Universidade do Porto
Faculdade de Engenharia
FEUP



Fernando Augusto Carneiro Sereno

Regressão com Redes Neuronais e Máquinas de Vectores de Suporte Aplicada à Estimação do Peso Fetal

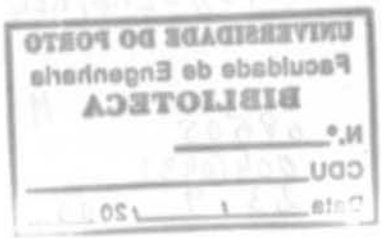
Porto, 2002

Faculdade de Engenharia da Universidade do Porto

Regressão com Redes Neuronais e Máquinas de Vectores de Suporte Aplicada à Estimação do Peso Fetal

Fernando Augusto Carneiro Sereno

Tese submetida ao Departamento de Engenharia Electrotécnica e de Computadores para obtenção do grau de Doutor em Engenharia Electrotécnica e de Computadores, realizada sob orientação científica do Doutor Joaquim P. Marques de Sá, Professor Associado com Agregação da Faculdade de Engenharia da Universidade do Porto


UNIVERSIDADE DO PORTO
Faculdade de Engenharia
BIBLIOTECA
N.º
CDU
Data / /20

Novembro 2002

004(043)/SERf/REG
UNIVERSIDADE DO PORTO
Faculdade de Engenharia
BIBLIOTECA M
N.º 68505
CDU 004(043)
Data 23, 9, 2003

Dedicatória
À minha família

Resumo

As redes neuronais são uma solução muito elegante quando se pretende encontrar respostas a muitos problemas relacionados com a previsão a partir de um conjunto de observações. Por exemplo, quais são as observações que se devem utilizar, quais as características que se devem extrair de cada observação, como se incorporaram informações suplementares sobre a verdadeira relação entre as entradas e as saídas no processo de aprendizagem ? Na investigação de respostas para estas e outras questões, o desenvolvimento das redes neuronais evoluiu consideravelmente distanciando-as bastante das suas raízes biológicas. Uma compreensão destas soluções constitui um dos objectivos principais desta dissertação, a qual também procura apresentar uma visão da recentemente introduzida teoria estatística da aprendizagem.

As máquinas de vectores de suporte constituem um paradigma recente para a aprendizagem automatizada, baseando-se em dois ingredientes principais. O primeiro consiste na utilização do conceito de transformação dos algoritmos lineares em algoritmos não-lineares, recorrendo a uma função real multivariável que aplica o espaço das entradas num espaço de características e o segundo consiste na combinação de uma mistura de técnicas provenientes da física, estatística, teoria da optimização e redes neuronais.

Nesta dissertação os aspectos teóricos e práticos das máquinas de vectores de suporte são tratados de forma equilibrada. Apresentamos um estudo cuidadoso dos desenvolvimentos mais recentes na teoria estatística da aprendizagem, e adaptamos e desenvolvemos experimentalmente novas ferramentas que permitem uma previsão precisa do erro de generalização de uma variedade de autómatos de aprendizagem.

Baseado neste estudo apresentamos sugestões de abordagens que permitem obter os melhores resultados com uma rede neuronal ajustada aos dados e ao problema, assim como garantir a melhor performance de generalização. Um estudo preliminar do problema da aprendizagem supervisada no quadro da teoria da informação, utilizando uma família de máquinas diferentes, sugere que uma abordagem da aprendizagem com várias máquinas, cada uma aprendendo à sua maneira, seguida de uma combinação cooperativa, torna possível melhorar a precisão da tarefa preditiva.

As redes neuronais comandadas pelos dados ("data-driven") têm sido aplicadas aos sistemas de apoio à decisão médica há vários anos e em várias especialidades. Contudo, a sua aplicação tem sido limitada por um conjunto de razões, uma das quais sendo a falta de provas suficientes de efectivamente conduzirem a estimativas mais precisas e portanto tornarem-se mais úteis que os métodos actualmente em uso clínico.

Utilizamos os dados de 414 exames ecográficos do feto, obtidos em quatro hospitais portugueses, cada exame compreendendo cinco medidas. Estes dados foram tratados com modelos de perceptrão de multicamada, redes de funções de base radiais e máquinas de vectores de suporte com o fim de obter modelos de regressão para a previsão do peso fetal antes do nascimento. A utilização destes modelos originou como melhor resultado um erro relativo de $\pm 6.71\%$. Este constitui um melhoramento considerável comparativamente com os erros relativos obtidos pelos métodos tradicionais e referidos na literatura, que são da ordem de $\pm 10\%$, ou mesmo $\pm 15\%$ segundo um estudo específico.

Abstract

Neural networks provide good solutions to several problems related to the prediction from a given set of observations. For example, which observations should be used, which features should be extracted from each observation and how additional information about the relationship between inputs and outputs can be incorporated in the learning process? In the search for answers to these and many other questions the development of neural networks went a long way from their biological roots. A theoretical understanding of some of these solutions is one of the main goals in this dissertation that attempts as well to present a vision within the modern statistical learning theory framework.

Support vector machines modeling is a very recent paradigm for machine learning that has two main ingredients. The first one is using the concept of transforming linear algorithms into nonlinear ones via a map into feature spaces, and the second one is combining a mixture of techniques taken from physics, statistics, optimization theory and neural networks.

In this dissertation the theory and practice of support vector machines receive similar attention. We present a careful study of up to date theoretical developments in statistical learning theory and experimentally adapt and create new tools that allow an accurate prediction of the generalization error of a variety of learning machines. Based on this study we present suggestions for how to get the most out of a neural network, appropriately tuned to the data and the problem, and for how to guarantee a maximum of the generalization performance. An exploratory study within the information-theory framework of supervised learning, using a family of different machines, suggests that using the available examples to train several machines, each one in a different way, and then combining them cooperatively, it will be possible to improve the accuracy of the predicting task.

Data-driven neural networks have been applied to medical decision support for many years in many clinical specialties. Nevertheless, their applicability has been limited for a variety of reasons, one of them is the lack of demonstration that they can effectively produce more accurate estimates and become more useful than the methods currently used in routine clinical practice.

We used a data set of 414 echographic examinations of the foetus collected in four Portuguese hospitals, each examination resulting in five distinct measurements. These data were processed using multilayer perceptrons, radial basis functions and support vector machines to get regression models predicting foetal weight before birth. Using these models results in a best prediction error rate of $\pm 6.71\%$. This is

a significant improvement over the reported errors using traditional methods which are of the order $\pm 10\%$, or even $\pm 15\%$ according to one specific source.

Résumé

Les réseaux de neurones donnent de bonnes solutions à plusieurs problèmes de prévision basée sur un ensemble donné d'observations. Par exemple, quelles sont les obsevations à utiliser, quelles les caractéristiques à extraire de chaque observation et comment incorporer informations suplémentaires sur la relation entre les entrées et les sorties dans le processus d'apprentissage ? Dans la recherche de réponses à ces et maintes autres questions, le développement des reseaux de neurones a évolué considérablement au delà de ses racines biologiques. Une compréhension théorique de ces solutions est un des objectifs principaux de cette thèse, qui présente aussi une vision dans la perspective de la théorie statistique de l'apprentissage.

Les machines à support vectoriel constituent un paradigme très récent pour l'apprentissage automatique, présentant deux ingrédients principaux. Le premier est l'utilisation du concept de transformation des algorithmes linéaires en algorithmes non-linéaires, ayant recours à une application sur les espaces de caractéristiques et le second est la combinaison d'un mélange de techniques empruntées à la physique, la statistique, la théorie de l'optimisation et, les réseaux de neurones.

Dans cette thèse les aspects théoriques et pratiques des machines à support vectoriels sont traités d'une manière équilibrée. Nous présentons une étude soignée des développements les plus récents en théorie statistique de l'apprentissage, et adaptons et développons de manière expérimentale de nouveaux outils qui permettent une prévision précise de l'erreur de généralisation d'une variété de automates d'apprentissage.

Basé sur cette étude nous présentons des suggestions pour les approches à suivre pour obtenir les meilleurs résultats avec un réseaux de neurones ajusté aux données et au problème, et aussi pour garantir la meilleure performance de généralisation. Une étude préliminaire dans le cadre de la théorie de l'information du problème de l'apprentissage supervisée, utilisant une famille de machines différentes, suggère que dans une approche d'apprentissage avec plusieurs machines, chacune apprenant à sa manière, suivi par combinaison coopérative, rend possible une précision améliorée de la tache prédictive.

Les réseaux de neurones pilotés aux données ("data driven") ont été appliqués aux systèmes d'aide à la décision médicale depuis plusiers années et dans plusieurs spécialités. Cependant, leur application a été limitée pour un certain nombre de raisons, une desquelles étant le manque de preuve qu'ils peuvent effectivement conduire à des estimations plus précises et donc peuvent devenir plus utiles que les méthodes actuellement en usage clinique.

Nous avons utilisées les données de 414 examens echographiques du fétus, obtenues en quatre hôpitaux portugais, chaque examen comportant cinq mesures. Ces données ont été traitées avec des modèles à perceptrons multicouches, réseaux aux fonctions de base radiales et machines à support vectoriels afin d'obtenir des modèles de régression pour la prédiction du poids fétal avant la naissance. L'utilisation de ces modèles donne comme meilleure résultat une erreur relative de prédiction de $\pm 6.71\%$. Ceci est une amélioration considérable comparée aux erreurs relatives obtenues avec les méthodes traditionelles et rapportées dans la littérature, qui sont de l'ordre de $\pm 10\%$, ou même $\pm 15\%$ selon une étude spécifique.

Agradecimentos

Ao meu orientador, Doutor Joaquim Pontes Marques de Sá, quero agradecer o seu empenho e qualidade na orientação científica desta dissertação, bem como a confiança que sempre me transmitiu.

Ao Doutor Luís António Pereira Leite, Doutor João Bernardes, Dr.ª Ana Maria Tavares de Matos e a toda a equipa de médicos do Serviço de Obstetrícia do Departamento de Ginecologia e Obstetrícia do Hospital de S. João, Serviço de Obstetrícia do Hospital Pedro Hispano, Serviço de Obstetrícia do Hospital Garcia de Orta e Maternidade Daniel de Matos, e a todas as pessoas que participaram na colheita de dados agradeço a ajuda e confiança que depositaram neste trabalho, assim como a disponibilidade demonstrada.

Ao Instituto de Engenharia Biomédica (INEB) da Universidade do Porto e, em particular, aos seus Directores, Doutor Mário Barbosa, Doutor Armando Padilha (2001-2002), Doutor Aurélio Campilho (1999-2001), apresento os meus sinceros agradecimentos pelas excelentes condições de trabalho que me foram proporcionadas, assim como a todos os seus membros pelo ambiente de cordialidade no qual fui inserido.

A todos os membros do painel de avaliação internacional do INEB nos anos 1999-2002 e, em particular, ao Doutor Bernard Buxton e Doutor Jacques Jossinet desejo exprimir os meus agradecimentos pelas relevantes sugestões recebidas.

Ao Doutor Willem van Meurs, Doutor Miguel Monteiro e Doutora Mariofanna Milanova expresso a minha gratidão pela disponibilidade que sempre tiveram quando alguma ajuda foi necessária.

Aos meus colegas de trabalho Carlos Felgueiras, Miguel Velhote Correia, Luís Alexandre, Jorge Barbosa e Pedro Marques desejo expressar a minha gratidão pela disponibilidade que sempre tiveram quando alguma ajuda foi necessária, assim como ao Paulo de Sousa pela ajuda e sua inteira disponibilidade.

Às Direcções do Centro de Informática Professor Correia de Araújo e da Biblioteca da Faculdade de Engenharia do Porto, assim como às Direcções da Biblioteca de Matemática Pura, Biblioteca de Matemática Aplicada da Faculdade de Ciências Porto, e aos funcionários destes serviços, agradeço a colaboração e o bom aten-

dimento que sempre me foi facultado, tendo em atenção que todos estes serviços sofreram mudanças de instalações em 2000 e 2001 sem se notar qualquer quebra significativa na boa qualidade dos serviços prestados.

Ao Presidente do Instituto Politécnico do Porto, Doutor Luís de Jesus Santos Soares, à Vice-Presidente do Instituto Politécnico do Porto, Doutora Maria de Fátima Morgado, às Presidentes do Conselho Directivo da Escola Superior de Educação do Porto, Dr.ª Maria de Jesus Lima (1999-2001), Doutora Maria do Carmo Castelo Branco (2001-2002), e à Presidente do Conselho Científico da Escola Superior de Educação do Porto, Doutora Fátima Lambert, agradeço as condições que me proporcionaram para levar a bom termo este doutoramento.

Agradeço ao programa PRODEP do Ministério da Educação pela concessão da bolsa de doutoramento, ao Instituto de Engenharia Biomédica da Universidade do Porto e seus Laboratório de Biomateriais e Laboratório de Sinal e Imagem, ao Departamento de Engenharia Electrotécnica e Computadores da Faculdade de Engenharia da Universidade do Porto, à Fundação Luso-Americana para o Desenvolvimento, à Fundação Calouste Gulbenkian e à Fundação para a Ciência do Ministério da Ciência e Tecnologia, toda a disponibilidade e ajuda recebida para o desenvolvimento deste trabalho, nomeadamente, o apoio para a participação em reuniões científicas e conferências internacionais de elevada qualidade que muito enriqueceram esta dissertação.

Agradeço à D. Gabriela Afonso e D. Teresa Teixeira a colaboração e atenção que sempre exprimiram em relação ao suporte do trabalho quotidiano de investigação no Laboratório de Sinal e Imagem do Instituto de Engenharia Biomédica, e ao trabalho descrito nesta dissertação em particular.

Conteúdo

1 Introdução 1

1.1 O problema: previsão do peso fetal 2
1.2 Modelos de regressão linear generalizados 3
1.3 Analogia estrutural entre rede neuronal e estatística 5
1.4 Origem histórica do modelo de rede neuronal 7
1.5 Aprendizagem e indução 9
1.6 Analogia estrutural entre a máquina de vectores de suporte e a estatística 11
1.7 Contribuições 12
1.8 Organização da tese 13

2 Aprendizagem Supervisada 15

2.1 A Dimensão de Vapnik-Chervonenkis 16
2.1.1 O problema da aprendizagem 16
2.1.2 Erro verdadeiro, amostral e de aproximação 20
2.1.3 Formalização de conceitos chave sobre aprendizagem 23
2.1.4 Aprendizagem de classes finitas de funções 25
2.1.5 Função de crescimento 27
2.1.6 Dimensão de Vapnik-Chervonenkis de uma classe de funções (d_{VC}) 29
2.1.7 Limites superiores da complexidade amostral 32
2.1.8 Classes de funções reais 33
2.1.9 Número de cobertura de subconjuntos do espaço Euclidiano . 33
2.1.10 Número de cobertura uniforme de uma classe de funções . . . 34
2.1.11 Pseudo-dimensão e dimensão fat-estilhaçamento 34
2.2 Aprendizagem de Funções Reais 39
2.2.1 Função de custo para funções reais (*regressão*) 43
2.2.2 Funcional de risco para funções reais *(regressão)* 44
2.2.3 Consistência dos processos de aprendizagem 48
2.2.4 Minimização do risco estrutural 52

2.3 Selecção de Modelos 53
2.3.1 O compromisso viés-variância 54
2.3.2 Critérios clássicos de selecção de modelos 56
2.4 Sumário 61
2.5 Referências bibliográficas e notas históricas 62

3 Máquina de Vectores de Suporte e Rede Neuronal 65
3.1 Máquina de Vectores de Suporte 66
3.1.1 Classificação 66
3.1.2 Hiperplano de classificação óptimo 67
3.1.3 Classificação de padrões não separáveis linearmente 77
3.1.4 Generalização num espaço de características de dimensionalidade elevada 78
3.1.5 Regressão linear com função de custo insensitiva 85
3.2 Elementos de Teoria da Regularização 90
3.2.1 Funcionais para suavização 91
3.2.2 *Kernel* da transformada de Fourier 92
3.2.3 Regularização da função aproximada 93
3.2.4 *Kernel* Reprodutivo no Espaço de Hilbert 94
3.2.5 Teorema de Mercer 96
3.3 Redes de Funções de Base Radiais (*FBR*) 96
3.3.1 Problema da aproximação de funções 97
3.3.2 Interpolação exacta com ruído 97
3.3.3 Aprendizagem 100
3.4 Perceptrão 100
3.4.1 Descrição segundo a teoria estatística da aprendizagem . . . 101
3.4.2 Máquinas de aprendizagem lineares 106
3.4.3 Regras clássicas de aprendizagem 111
3.4.4 Perceptrão de Multicamada (PMC) 116
3.4.5 Capacidade de representação 118
3.5 Sumário 119
3.6 Referências bibliográficas e notas históricas 119

4 Biometria Fetal, Análise Exploratória de Dados e Regressão 123
4.1 Biometria fetal pelo método ecográfico 124
4.1.1 Metodologia da colheita de dados 128
4.2 Análise exploratória de dados 130
4.2.1 Tipo de experiência 130
4.2.2 Análise multivariável 131

4.2.3 Um modelo probabilístico simples para a densidade dos pesos fetais . . . 135
4.2.4 Análise das componentes principais . . . 141
4.2.5 Regressão do PFE sobre as principais variáveis . . . 144
4.2.6 Regressão simples dos PFE baixos . . . 152
4.2.7 Regressão de PFE baixos sobre duas variáveis ecográficas . . 157
4.2.8 Regressão dos PFE baixos sobre múltiplas variáveis . . . 163
4.2.9 Efeito na regressão dos erros nas medidas . . . 167
4.3 Tópicos sobre o modelo de regressão linear . . . 171
4.3.1 Esperança condicional . . . 171
4.3.2 Variância condicional . . . 174
4.3.3 Modelo linear . . . 176
4.3.4 Modelos linear generalizado e não-linear . . . 179
4.4 Máxima verosimilhança e estimação paramétrica Bayesiana . . . 180
4.4.1 Aprendizagem Bayesiana . . . 183
4.4.2 Aprendizagem não paramétrica . . . 184
4.4.3 Fronteira de decisão . . . 185
4.5 Sumário . . . 189
4.6 Referências bibliográficas e notas históricas . . . 189

5 Regressão por Rede Neuronal e Máquina de Vectores de Suporte 191
5.1 Rede Neuronal de Perceptrão de Multicamada . . . 193
5.1.1 Resultados obtidos com um conjunto de treino de 140 casos . 194
5.1.2 Resultados obtidos com um conjunto de treino de 213 casos . 197
5.2 Rede Neuronal de Funções de Base Radiais . . . 199
5.2.1 Resultados obtidos com um conjunto de treino de 283 casos . 205
5.2.2 Resultados obtidos com um conjunto de treino de 220 casos . 207
5.3 Máquina de vectores de suporte . . . 209
5.3.1 Resultados obtidos com um conjunto de treino de 72 casos . . 210
5.3.2 Resultados obtidos com um conjunto de treino de 187 casos . 212
5.3.3 Resultados de MVS treinadas com um conjunto de treino obtido por edição . . . 213
5.3.4 Experiências com uma MVS que utiliza um algoritmo de decomposição . . . 217
5.4 Comparação de Resultados obtidos com PMC, FBR e MVS (algoritmo de decomposição) . . . 219
5.5 Análise numérica do algoritmo de aprendizagem de uma MVS . . . 221
5.5.1 Aproximação . . . 221
5.5.2 Estimação . . . 232

5.5.3 Problemas de álgebra linear numérica 235
5.6 Limites da capacidade de generalização 238
5.6.1 Complexidade amostral 238
5.6.2 Escolha óptima dos parâmetros C e ε 239
5.6.3 Selecção de Modelos 240
5.6.4 Métodos de kernel e métodos de dicionário 242
5.7 Sumário 243
5.8 Referências bibliográficas e notas históricas 243

6 Estimação Baseada na Teoria da Informação 245
6.1 Perspectiva teorético-informacional da aprendizagem 247
6.2 Entropia e informação 250
6.2.1 Informação de Fisher 251
6.2.2 Entropia de Vapnik 252
6.2.3 Entropia de Shannon 256
6.2.4 Entropia de Kullback-Leibler diferencial 261
6.2.5 Informação mútua 261
6.2.6 Aprendizagem teorético-informática 263
6.2.7 Entropia de Renyi 265
6.2.8 Entropia quadrática e o seu estimador não paramétrico . . . 267
6.3 Estimação de uma fdp pelo método das janelas de Parzen 269
6.4 Experiências com redes PMC estimadoras do peso fetal treinadas por um algoritmo informático-teorético 272
6.4.1 Estudo do erro de treino e generalização 272
6.4.2 Comparação de duas redes PMC 274
6.4.3 Combinação de modelos usando critérios de informação mútua 278
6.5 Sumário 279
6.6 Referências bibliográficas e notas históricas 281

7 Discussão e Conclusões 283
7.1 Discussão dos Resultados 283
7.2 Discussão epistemológica de quatro tópicos sobre a Teoria Estatística da Aprendizagem 286
7.2.1 Interpretação das demonstrações dos teoremas em TEA . . . 287
7.2.2 A lógica indutiva 289
7.2.3 Aproximação bayesiana e processos gaussianos 292
7.2.4 Medição indirecta e a matemática da inversão 294
7.3 Discussão da evidência de benefícios provenientes da utilização de modelos de rede neuronal 297

7.3.1 Tomada de decisão assistida por computador e por inferência estatística em medicina 297
7.3.2 Estimador por rede neuronal baseado no Windows 300
7.3.3 Perspectivas de desenvolvimento no futuro: pesquisa por conteúdo e *data mining* 301
7.4 Conclusões 304

Lista de Figuras

2.1 (A) Modelo de Aprendizagem Supervisada (B) Esquema de Algoritmo de Aprendizagem 18

2.2 Erro amostral e verdadeiro. 22

2.3 Representação gráfica dos limites superiores da Função Crescimento $\Pi_{\mathbb{H}}$: (a) recta, 2^m (b) curvas $\left(\frac{\exp(1)*m}{d_{VC}}\right)^{d_{VC}}$ para $d_{VC} = 2,3,4$ (4, curva superior) 29

2.4 Dimensão VC de hiperplanos 31

2.5 Os centros das "*bolas*"numa métrica d_∞ formam uma ε-cobertura da região a tracejado $\mathbb{F}_{|\mathbb{X}} \subset \mathbf{R}^2$ 35

2.6 Pseudo-estilhaçamento 36

2.7 Minimização do erro amostral $\widehat{er}_{\mathbf{z}}(h)$ equivale a aproximar o erro verdadeiro $er_P(h)$ (Anthony M. and Barlett P. L.,1999,p.20) 42

2.8 Grandes desvios 51

2.9 Minimização do Risco Estrutural 53

2.10 Compromisso viés-variância ilustrado no domínio da regressão (Duda et al., 2000, p.467) 57

2.11 Melhor ordem do modelo em conformidade com o critério de Akaike 61

3.1 Exemplo 46: a recta a contínuo separa os quadrados das cruzes, os círculos assinalam os vectores de suporte. 76

3.2 A idea das MVS é uma transformação Φ aos dados de treino mergulhando-os num *espaço de características* em que se pode determinar um hiperplano separador de margem máxima. 80

3.3 Regressão com MVS com uma precisão desejada $\pm\varepsilon$. O gráfico da direita representa a função de custo $\varepsilon - insensitiva$. 89

3.4 Arquitectura de uma rede de funções de base radial 99

3.5 Maximização da distância entre dois planos de suporte paralelos através da resolução do problema dual. 104

3.6 Um hiperplano de separação $(\mathbf{w}, b)$ para um conjunto bidimensional 107

3.7 Margem de um hiperplano de separação $(\mathbf{w}, b)$ para um conjunto bidimensional. 109
3.8 Descida do gradiente 114
3.9 PMC com três entradas e uma camada escondida 117
3.10 Capacidade de representação de três arqutecturas de redes se as unidades de processamento forem em degrau de Heaviside. 118

4.1 Diagrama de Caule-e-folhas de pesos à nascença (" fetais") 138
4.2 Teste de normalidade aos pesos fetais ($m = 414$) 140
4.3 Gráficos de caixa das variáveis 142
4.4 Gráfico dos dados segundos as duas primeiras novas variáveis (ou componentes principais) 143
4.5 Gráfico de Pareto ("*scree*"). Percentagem de variabilidade explicada. 143
4.6 Var. explicativa, regressora ou independente: perimetro abdominal, PAB. 144
4.7 Var. resposta, regressanda ou dependente: peso fetal, PFE. 145
4.8 Histograma do PFE com 10 intervalos 145
4.9 Histograma do peso fetal com 70 intervalos e a curva Gaussiana ajustada $\mathcal{N}(3225, 628)$ 146
4.10 Histograma do perimetro abdominal com 70 intervalos e a curva Gaussiana ajustada $\mathcal{N}(33.1, 3.1)$ 147
4.11 Diagrama de dispersão do PFE versus PAB, histogramas e recta de regressão, permite visualizar a correlação entre as variáveis. 148
4.12 Valores previstos versus valores residuais para a variável peso fetal (PFE) 149
4.13 Histograma dos resíduos 150
4.14 Valores previstos versus valores dos resíduos da variável PFE 151
4.15 Diagrama de normalidade dos resíduos 151
4.16 Diagrama de dispersão e histogramas das variáveis realativos a 47 casos de pesos baixos, $PAB < 30$cm e $PFE < 3210$ gr. 153
4.17 Valores previstos vs resíduos para a var PFE (pesos baixos, 47 casos) 154
4.18 Valores previstos vs. observados para a var. PFE ($n = 47$, pesos fetais baixos) 155
4.19 Valores observados vs. residuais para a var. PFE ($n = 47$, pesos baixos) 155
4.20 Gráfico de normalidad dos resíduos ($n = 47$) 156
4.21 Histograma da distribuição dos resíduos 156
4.22 Representação $3D$ dos casos ($n = 46$). Segundo o eixo vertical está representada a var. PFE 157

4.23 Correlações para pesos baixos (n=46) 158
4.24 Valores previstos vs. observados, duas var. PAB e CFE $(m = 46)$. 160
4.25 Valores dos previstos vs. resíduos, duas var. PAB e CFE, $(m = 46)$ 161
4.26 Valores observados vs. resíduos, duas var. PAB e CFE, $(m = 46)$. 161
4.27 Gráfico de normalidade, duas var. PAB e CFE $(m = 46)$ 162
4.28 Histograma dos resíduos, duas var. PAB e CFE $(m = 46)$ 162
4.29 Perímetros cefálicos medidos no recém-nascido (PCR) e por ecografia fetal (PCE). 170
4.30 Representação gráfica tridimensional de $P(a, p)$. 172
4.31 Representação das curvas de nível (P = *constante*) da densidade $P(a, p)$ projectadas no plano a, p, ... 173
4.32 Representação I da densidade de probabilidade condicional $P(p|a)$. 174
4.33 Representação II da densidade de probabilidade condicional $P(p|a)$. 175
4.34 Representação da regressão $E(p|a)$ 176
4.35 Evolução do perímetro abdominal em função do número de semanas de gestação 178
4.36 Histogramas do perímetro abdominal para 3 classes de PFE 185
4.37 Digramas de dispersão CFE vs. PAB para 2 classes de PFE 188

5.1 Gráficos dos erros amostrais de um *PMC-5:5:2* em função do número de épocas. 199
5.2 Gráfico (*PMC-5:5:2*), para um dos subconjuntos de teste, das previsões do: (a) Peso fetal, círculos na parte superior, (b) Comprimento fetal, na parte inferior. 200
5.3 Gráfico (PMC-5:5:2), para um dos subconjuntos de casos, das previsões do: (a) Peso fetal (esquerda) (b) Comprimento fetal (direita). 200
5.4 Gráfico das estimativas (círculos) obtidas com uma rede $FBR - 5 : 7 : 2$ (71 casos): (a) Peso fetal, parte superior, (b) Comprimento fetal, parte inferior. Os valores verdadeiros estão representados por pontos. 206
5.5 Estimação de pesos fetais **baixos** (PFE) por regressão com uma máquina de vectores de suporte (n *casos teste* $= 30$). 212
5.6 Estimação de pesos fetais **altos** (PFE) por regressão com uma máquina de vectores de suporte (n *casos teste* $= 38$) 213
5.7 Variabilidade das medidas de perímetro abdominal (PAB), diâmetro biparietal (DBP) e peso fetal (PFE). 214
5.8 Representação do sobre-ajustamento de uma MVS regressora aos dados. 215
5.9 Peso fetal (PFE) estimado por uma MVS em 26 casos teste. 217

5.10 Peso-à-nascença (gramas) em função do PAB, perímetro abdominal a (cm). A Expressão 5.4 está representada a traço interronpido. . . . 223
5.11 Peso-à-nascença (PFE,gramas). Pab (cm). Aproximação (linha contínua). Exacto (linha interrompida). $\varepsilon - tubo$ (linha a ponteado). $\varepsilon = 0.3$, $C = 20000$, kernel polinomial de grau 4. Vectores de suporte (círculos).231
5.12 Estimativa por uma MVS regressora $\varepsilon = 0.3$, $C = 20000$, kernel Polinomial de grau 4. O conjunto de treino éconstituido por 11 pontos sem ruído [A] e com ruído [B] e [C] 235
5.13 Aproximação por uma MVS-Regressora C, ε e grau do polinómio vriáveis (n^{o}casos=6). 241

6.1 Codificador e descodificador 257
6.2 Treino de uma máquina de aprendizagem pelo processo de adaptação dos parâmetros livres w. 264
6.3 Resultados de um PMC treinado por minimização do potencial informativo ($\alpha = 1.5$). 273
6.4 Risco esperado, PMC: "ent"significa que foi treinado com optimização entrópica e "std"com optimização MSE 276
6.5 Resultados preliminares obtidos com um perceptrão simples treinado por um critério informático-entrópico ($\alpha = 2$). 279

7.1 Estimador de rede neuronal em Windows 301

Lista de Tabelas

2.1 Table Caption 21
2.2 Classificações(exemplo) 22

3.1 Conjunto de treino no espaço de entradas 79
3.2 Conjunto de treino no espaço de entradas 80
3.3 Exemplos de funções kernel 84
3.4 Forma primal do algoritmo do perceptrao 108
3.5 Forma dual do algoritmo do perceptrao 111

4.1 Matriz dos Dados em Bruto (var- 1-7, extracto) 132
4.2 Matriz dos Dados em Bruto (var. 8-17, extracto) 133
4.3 Variáveis recolhidas 133
4.4 Métodos de medida 134
4.5 Matriz das correlações (I) 134
4.6 Matriz das correlações (II) 134
4.7 Estatísticas descritivas e Correlações com o PFE 139
4.8 As primeiras 3 componentes principais 142
4.9 Variabilidade explicada (I) 142
4.10 Variabilidade explicada (II) 143
4.11 Sumário dos resultados da regressãodo PFE sobre o PAB 146
4.12 Análise da variância da regressão 147
4.13 Resíduos superiores a +sigma e inferiores a -sigma (n=414) 148
4.14 Partição dos Pesos 152
4.15 Estatisticas descritivas de 47 casos de pesos baixos PAB¡30 e PFE¡3210 152
4.16 Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414) 153
4.17 Análise da variância da regressão 154
4.18 Tabela (n=47) dos valores máximos e mínimos dos resíduos 154

4.19 Estatisticas descritivas de 46 casos de pesos baixos PAB¡30 e PFE¡3210 157
4.20 Matriz das Correlações de pesos baixos PAB¡30 e PFE¡3210 158
4.21 Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414) 159
4.22 Análise da variância da regressão pesos baixos (n=46) 159
4.23 Correlações Parciais pesos baixos (n=46) 159
4.24 Resíduos superiores a +sigma e inferiores a -sigma (I) 160
4.25 Resíduos superiores a +sigma e inferiores a -sigma (II) 160
4.26 Estatisticas descritivas de 11 var. e 46 casos de pesos baixos PAB¡30 e PFE¡3210 163
4.27 Matriz (I) das Correlações de pesos baixos PAB¡30 e PFE¡3210 . . . 164
4.28 Matriz (II) das Correlações de pesos baixos PAB¡30 e PFE¡3210 . . 164
4.29 Sumário dos resultados da regressão lin. com todas as var. explic. . 165
4.30 Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414) 165
4.31 Análise da variância da regressão pesos baixos (n=46) 165
4.32 Correlações Parciais (I) pesos baixos (n=46) 165
4.33 Correlações Parciais (II) pesos baixos (n=46) 166

5.1 Resultados de experiências com PMC treinados com140 casos e testadas com 140 casos (disjuntos) 195
5.2 Melhor rede PMC-5-4-2-1 (140 casos treino, 140 casos teste) 196
5.3 Resultados PMC-5:5:1 (213 casos treino, 71 casos teste) 198
5.4 Resultados PMC-5:5:2 (213 casos treino, 71 casos teste) 198
5.5 Exemplo de uma estrutura de rede FBR-5:5:2 (Netlab) 201
5.6 Resultados FBR-5:7:2 (283 casos treino, 71 casos teste) 206
5.7 Resultados PMC-5:5:2 (213 casos treino, 71 casos teste) 207
5.8 Resultados FBR-5:10:2 (220 casos treino, 57 casos teste) 209
5.9 Matriz de confusão da rede 4 (Erro de teste=6.87 %.) 209
5.10 Resultados FBR-5:10:2 (220 casos treino, 55 casos teste) 210
5.11 Erros de Clasificação de MVS num conjunto de teste (37 casos) . . . 211
5.12 Resultados obtidos no teste com três MVS e para vários valores de C 216
5.13 Resultados de um algoritmo de decomposição com kernel polinomial 218
5.14 Resultados de um algoritmo de decomposição com kernel FBR gaussiano219
5.15 Resultados de um algoritmo de decomposição com kernelsigmoidal . 219
5.16 Valores dos parâmetros usados no treino das MVS (com Algoritmo Decmposição) 220

5.17 Estatísticas dos erros relativos obtidos com 2 modelos de rede neuronal, 3 modelos de MVS e 2 modelos clássicos de regressão (Con. treino 210 casos, Conj. Teste 70 casos) (. 221

6.1 Aprendizagem de uma rede neuronal vista como um processo de comunicação 248
6.2 Resultados de treino de um PMC 274
6.3 Riscos previstos com PMC 276

Notação

A^{-1}	matriz inversa
A^T	matriz transposta
A^*	matriz adjunta (ou, operador, ou vector)
A^Ψ	pseudo-inverso da matriz A
$\mathcal{A}$	$\sigma-$álgebra, ou tribo, sobre Ω
$\mathcal{B}$	$\sigma-$álgebra de Borel definida em $\mathbb{R}$
B	elemento do conjunto $\mathcal{B}$ (tal que se $X^{-1}(B) \in \mathcal{A}$, então X é var.aleatória)
c_f	forma compacta da função custo em (x, y) para a saída $f(x)$
$c(x, y, f(x))$	função de custo em (x, y) para a saída $f(x)$
C	constante de regularização ($C = 1/\lambda$)
C_i	classe i, ou partição de um conjunto
d_{VC}	dimensão VC (por vezes representada por h)
$d(x, y)$	distância entre x e y na métrica d (tb. se usa às vezes ρ)
e	base dos logaritmos naturais $lim\,(1 + (1(n))^n$, quando $n \to \infty$
e	*eigenvector*, vector próprio
er_P	erro verdadeiro supondo a lei de probabilidades P
$\widehat{er}_{\mathbb{Z}}$	erro amostral (ou empírico) para a matriz de exemplos $\mathbf{Z}$
$\mathbf{E}[X]$	esperança matemática, ou valor esperado, da variável aleatória X
$\mathbf{E}[X\|Y]$	esperança matemática, ou valor esperado *condicional* da var. aleatória X dado Y

f	função real de variável real (ou de variável vectorial em $\mathbf{R}^n$)
$\mathbb{F}$	conjunto de funções reais (classe de)
$fat_{\mathbb{F}}$	dimensão fat-estilhaçamento
h	dimensão VC (por vezes representado por d_{VC})
$\mathbb{H}$	conjunto de funções binárias (etiquetas $\{\pm 1\}$ou $\{0,1\}$)
iid	independentes e identicamente distribuidas (exemplos, amostras ...)
i,j	variáveis para índices de sucessões
$\mathcal{I}_A(w)$	função indicatriz do conjunto A (ou: $=1$ se $w \in A$,ou: $=0$ se $w \notin A$
$\mathbf{I}$	matriz identidade
$\inf_x f(x)$	infimo de $f(x)$,ou máximo dos minorantes, ou mínimo global de $f(x)$, $\forall x$
$\mathbf{J}$	matriz Jacobiana
k	kernel de Mercer
K	kernel (definido e positivo)
$\mathbb{H}$	espaço de características com um produto interno tal que $K(x,x') = \langle \Phi(x), \Phi(x') \rangle$
$\mathbf{H}$	matriz kernel, matriz de Gram ou hessiana
H_{ij}	entrada da matriz (Kernel, por exemplo) correspondente à linha i coluna
$\mathcal{L}$	verosimilhança
L_p^{μ}	espaços de funções que são μ^p integráveis
log	logaritmo na base 10

$\log_2$	logaritmo na base 2
ln	logaritmo natural
m	número de exemplos de treino
$[\mathbf{x}]_i$	componente ou entrada i do vector $\mathbf{x}$
n	número de variáveis de entrada, ou de caracteristicas
$\mathbf{N}$	conjunto dos números naturais
$N(\varepsilon, \mathbb{F}, d)$	número cobertura de um conjunto $\mathbb{F}$, na métrica d com precisão ε
$N(\mu, \sigma)$	função Gaussiana, ou distribuição normal
	de parametros: média μ, variância σ
$O(p(x))$	grande oh da ordem de $p(x)$
$opt_P(\mathbb{H})$	função da classe $\mathbb{H}$ que é optima em relação à distribuição P
$P(C)$	probabilidade de um conjunto (ou acontecimento) C
$P\{\cdot\}$	probabilidade de uma fórmula lógica
$R(f)$	risco esperado da função f (ou, o funcional do ...)
$R_{emp}(f)$	risco empírico da função f (ou, o funcional do ...)
$\mathbf{R}$	conjunto dos números reais (ou, corpo de ...)
$\mathbf{R}^n$	espaço euclideano n-dimensional
s	desvio padrão de uma variável numa amostra
$sign$	função não linear, sinal de, $sign(x) = 1$ se $x > 0$, ou $= -1$se $x \leq 0$
$\sup_x f(x)$	supremo de $f(x)$, minimo majorante,
	ou máximo global de $f(x)$, $\forall x$

$\mathbf{w}$	vector de pesos
$\mathbb{W}$	espaço dos parâmetros
X	variável aleatória real definida sobre $(\Omega, \mathcal{A})$
$\mathbb{X}$	espaço de entrada dos vectores exemplos (em geral $\mathbb{R}^n$)
$\mathbf{x}_i$	vector de entrada do i-ésimo exemplo
Y	variável aleatória real definida sobre $(\Omega, \mathcal{A})$
$\mathbb{Y}$	espaço das etiquetas (ou rótulos) dos exemplos
y	etiqueta do i-ésimo exemplo
$\mathbf{z}_i$	exemplo constituido pelo par ordenado $(\mathbf{x}_i, y_i)$, retirado independentemente dos restantes, em conformidade com uma distribuição de probabilidade P.
$\mathbb{Z}$	espaço de treino, produto cartesiano do conjunto das entradas pelo conjunto das etiquetas $\mathbb{X} \times \mathbb{Y}$
$\mathbb{Z}^m$	espaço de amostragem de onde são retiradas amostras de treino com m exemplos, independentes, de acordo com a distribuição de probabilidade conjunta (produto) P^m
$\mathbf{Z}$	matriz de de exemplos de treino $m \times n$ pertencente a $(\mathbb{X} \times \mathbb{R})^m$
α	multiplicador de Lagrange
α	vector de todos os multiplicador de Lagrange
β_i	coeficientes de expansão auxiliares
δ	grau ou nivel de confiança

δ_{ij}	delta de Kronecker que tem o valor 1 se os dois indices coincidem
$\delta_{x_i}(x)$	distribuição delta de Dirac centrada em x_i
∇	operador gradiente
$\epsilon_n(\mathbb{X})$	$n-$ésimo número de entropia do conjunto $\mathbb{X}$
ε	precisão, vizinhança ou parametro da função de custo $\varepsilon-$insensitiva
γ	parâmetro
θ	parâmetro de uma lei de distribuição de probabilidade standard
$\hat{\theta}$	estimativa de θ dada por verosimilhança
λ	constante de regularização ($\lambda = 1/C$), tb. valor próprio (*eigenvalue*)
$\mathbf{\Lambda}$	matriz diagonal de *eigenvalues*
μ	medida de probabilidade
$\mu(x)$	medida de x
η	taxa de aprendizagem
η_i	multiplicador de Lagrange correspondente a ξ_i
$\triangle$	operador de Laplace
Φ	aplicação linear no espaço de caracteristicas tal que $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$
ϕ_i	função de base para a aplicação linear Φ
ρ	métrica num espaço $\mathbb{X}$. $\rho : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. (tb. se usa às vezes d)
σ	desvio padrão, largura de kernel
σ^2	variância,

$\boldsymbol{\Sigma}$	matriz das covariancias
ξ_i	variáveis de forlga
$\Pi_{\mathbb{F}}$	função crescimento
ω	frequencia espacial
Ω	conjunto de acontecimentos elementares de um espaço probabilizável
$\Omega\left(p\left(x\right)\right)$	grande omega da ordem de $p\left(x\right)$
$\tilde{f}\left(\omega\right)$	transformada de Fourier de $f\left(x\right)$
$\langle \mathbf{u}, \mathbf{v} \rangle$	produto interno de $\mathbf{u}$ por $\mathbf{v}$
$\left\|\mathbb{X}\right\|$	cardinal do conjunto $\mathbb{X}$
$\left\|x\right\|$	módulo de x
$\left\|\cdot\right\|$	norma, em geral a distancia euclideana $\left\|\mathbf{u}\right\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$
$\left\|\cdot\right\|_p$	$p-$norma, por exemplo, $\left(\sum_i \left\|x_i\right\|^p\right)^{1/p}$
$\left\|\mathbf{A}\right\|$	operador norma de $\mathbf{A}$, i.é, $\sup_x \left\|\mathbf{A}x\right\| / \left\|x\right\|$
$\lceil x \rceil$	tecto de x, mínimo inteiro não inferior a x
$\lfloor x \rfloor$	soalho de x, máximo inteiro não superior a x
$\binom{n}{k}$	combinações de n, k a k, $\left(= n!/\left(k!\left(n-k\right)!\right)\right)$
$\mathbf{1}$	matriz unidade, também operador identidade
$\mathbf{1}_n$	vector de comprimento n consistindo apenas de 1s
$\mathbf{0}$	vector de zeros
$a \succ b$	os objectos a "têm preferencia em relação" ao objectos b

Capítulo 1

Introdução

A *aprendizagem supervisada* a que se faz referência na presente tese consiste numa estimação de uma dada função real de mais de uma variável real, conhecendo uma imagem da distribuição de probabilidades a que pertencem um conjunto finito de *exemplos* conhecidos, e que satisfazem a função a estimar. Pode resumidamente dizer-se que uma tal estimação requer a determinação de um conjunto finito de parâmetros de uma dada máquina de aprendizagem, mais precisamente de um dado *algoritmo de aprendizagem*, o qual deve ser capaz de determinar os valores óptimos desses parâmetros. Esta optimização deve ser entendida no sentido de maximizar a capacidade de generalização e a simplicidade da solução a ser implementada em qualquer *computador* moderno.

Existe um certo paralelismo entre a aprendizagem de uma máquina e a aprendizagem de um ser vivo. Uma máquina aprende sempre que ela altera a sua estrutura, programa ou dados memorizados, em resposta a informações recebidas do meio exterior, ou a dados de entrada provenientes do seu meio interior, e de forma a que se possa sempre esperar a verificação de um melhoramento da sua performance no futuro. O mesmo se pode aplicar a um ser vivo, ou a um animal, com a diferença de que, no caso da actividade de um ser humano, que é o mais complexo de todos, existe ainda uma procura de um *significado*, específico de cada indivíduo, ou grupo de indivíduos, uma capacidade de *representação* desse significado, por exemplo, através de um gesto, fala, escrita, equação, livro, filme, e, finalmente, existe um processo através do qual a partir desse, ou desses significados, se podem desenvolver outros, todos formando uma teia de significados, os quais podendo tornar-se operacionais constituem um processo que se designa por *intencionalidade*.

Nesta última geração a humanidade viu-se confrontada com a inevitável necessidade de partilhar o mundo com uma nova espécie de estranhas máquinas, em particular destacam-se as máquinas do tipo computador. Ao contrário das máquinas que manipulam a energia, as quais deram origem à revolução industrial do século XIX,

os computadores lidam com o controlo, a informação e os processos intelectuais. Poucas pessoas duvidam que estas máquinas se estão a desenvolver rapidamente, quer em capacidade, quer em complexidade, e que, portanto, estarão destinadas a desempenhar um papel cada vez mais importante nas sociedades actuais. Em 1956 foi criado o nome *inteligência artificial* para designar um novo domínio do conhecimento cujos objectivos eram a construção de máquinas "*inteligentes*" e, por outro lado, a compreensão do seu funcionamento, de forma a que tais máquinas fossem capazes de atingir um nível tão similar quanto possível ao da inteligência humana.

Uma das mais importantes preocupações da inteligência artificial desde a sua fundação tem sido o desenvolvimento de capacidades de raciocínio com vista ao apoio à *decisão médica*. Os problemas teóricos e práticos que um tal desenvolvimento tem levantado nas últimas décadas deram origem a um grande impulso quer nos métodos da teoria das probabilidades e estatística, quer nos métodos da ciência dos computadores. Verifica-se que na década de mil novecentos e noventa os *sistemas de apoio à decisão* passam a ser cada vez mais utilizados por rotina na prática clínica e em larga escala, tal como refererem, por exemplo, Fauci et al. (1998), Hudson and Cohern (2000) e Lisboa (2002).

As máquinas que irão ser referidas nesta tese têm igualmente o objectivo de apoiar a decisão médica em situações clínicas, particularmente na especialidade de obstetrícia, e caracterizam-se por envolverem modelos *artificiais* de *rede neuronal* e, outros modelos que se inspiram nestes, mas como têm princípios de funcionamento diversos têm uma designação própria, são chamados modelos de máquina de aprendizagem por *kernel*, ou *máquina de vectores de suporte*.

1.1 O problema: previsão do peso fetal

O trabalho descrito na presente tese foca-se num projecto cujo objectivo é o desenvolvimento de um método de previsão do peso do recém-nascido antes do nascimento, utilizando um modelo de rede neuronal que processa medidas bimétricas extraídas por ecografia fetal. Pretende-se saber em que medida se pode estimar o peso fetal de forma mais precisa do que através da utilização das actuais fórmulas baseadas em *modelos de regressão linear generalizados*, que apresentam um grau de erro considerável, sobretudo em dois casos especiais em que o feto ou apresenta um peso muito alto, ou um peso muito baixo.

Este projecto de investigação de estimativa ecográfica do peso fetal insere-se no âmbito de um estudo multicêntrico que compreendeu quatro Hospitais Portugueses, nos quais se procedeu à colheita de informações em conformidade com um protocolo, que foi cuidadosamente analisado e considerado como realista na prática clínica pelos especialistas de obstetrícia que participaram na colheita de dados. Este protocolo

contém especificações detalhadas dos critérios de inclusão de parturientes, dados maternos a recolher, e métodos de medida ecográfica do *perímetro cefálico*, *diâmetro biparietal*, *perímetro abdominal*, *comprimento do fémur*, *índice de resistência umbilical* (ou *índice de Pourcelot*) e *índice de pulsatilidade*.

Os estudos exploratórios das primeiras recolhas de dados começaram em 1999, a base de dados, compreendendo um total de 414 casos, e que foi utilizada nas experiências de modelação relatadas nos capítulos quarto, quinto e sexto, ficou estabelecida em finais de 2000 .

1.2 Modelos de regressão linear generalizados

Em qualquer ciência exacta sabe-se que uma representação quantitativa de um fenómeno contém uma parte que tem a ver com o acaso, ou com factores aleatórios, e outra parte que tem a ver com as incertezas que podem ocorrer nas medidas, e que ambas as partes podem ser fontes de flutuações, maiores ou menores, à volta de um certo valor médio que se pode teoricamente admitir que exista. Sempre que se procura um bom ajustamento de uma função a um conjunto de dados experimentais deve ter-se em linha de conta todas as fontes de incerteza, procurando traçar-se o melhor trajecto possível sempre no interior de uma núvem de pontos. Se é o acaso que, de facto, caracteriza uma grande maioria dos fenómenos da natureza, então, muitos aspectos desta deixam de estar determinados rigorosamente, pelo que é a partir das *correlações* melhores, ou piores, que se pode tentar encontrar explicações, ou direcções para a acção.

É possível estabelecerem-se correlações múltiplas sempre que se conhecem m exemplos, (X_i, Y_i), em que $X_i = (X_{i1}, \ldots, X_{in})$ é um vector aleatório conhecido com n componentes, cada uma sendo um valor medido de uma variável explicativa, e em que Y_i é um valor conhecido de uma variável aleatória explicada, no i-ésimo exemplo, em que $i = 1, 2 \ldots m$. É também possível determinarem-se n coeficientes w_i, e um termo constante w_0, que satisfaçam um sistema de m equações:

$$Y_i = w_0 + w_1 X_{i1} + w_2 X_{i2} + \cdots + w_n X_{in} + \varepsilon_i, \quad i = 1, 2 \ldots m$$

em que a parcela ε_i representa um erro aleatório de média $E[\varepsilon_i] = 0$, variância constante, $\sigma^2[\varepsilon_i] = \sigma^2$, e em que os erros ε_i e ε_j de dois exemplos diferentes não estejam correlacionados, isto é, a sua covariância é nula, $\sigma[\varepsilon_i, \varepsilon_j] = 0$, $i \neq j$, e em que $i = 1, \ldots, m$.

Este modelo de *regressão linear normal* emprega-se no estudo de um determinado fenómeno com o objectivo de explicar o seu comportamento, tendo-se como informação de base para este estudo um conjunto finito de exemplos. Este método

remonta aos finais do século XIX quando se procedia a estudos sobre a hereditariedade, levados a cabo por Mendel e principalmente por Galton (1869-1890).

O modelo de regressão pressupõe que as observações da variável explicada Y tenha uma distribuição normal gaussiana, $Y \sim N\left(X\mathbf{w}, \sigma^2\mathbf{I}\right)$, com valor esperado $\mathbf{E}[Y] = X\mathbf{w}$, e a matriz das covariâncias $\sigma^2\mathbf{I}$. Pode então obter-se a relação:

$$Y = w_0 + w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

que mais se aproxima da realidade dada pelas m observações (X_i, Y_i), no sentido em que minimiza um dado funcional dos erros $r(\mathbf{w})$, em que $\mathbf{w} = [w_0, w_1, \ldots w_n]$.

Pode empregar-se neste modelo o chamado estimador dos mínimos quadrados para determinar os coeficientes $\mathbf{w}$. Este estimador minimiza o funcional dos erros, $r(\mathbf{w})$, que se pode exprimir pelo somatório dos quadrados de m desvios $U_i = Y_i - \langle x, \mathbf{w}\rangle$:

$$r(w) = \sum U_i^2 = U^TU = (Y - \langle \mathbf{x}, \mathbf{w}\rangle)^T (Y - \langle \mathbf{x}, \mathbf{w}\rangle)$$

Pode mostrar-se que este estimador é o mais eficiente estimador de entre os estimadores possíveis não enviesados Murteira et al. (2001), no entanto, podem utilizar-se outros estimadores, como, por exemplo, o da máxima verosimilhança.

Designam-se por modelos de *regressão linear generalizado* (*GLM*, em inglês "*generalized linear model*") os modelos lineares, mas também outros modelos mais complexos, como, por exemplo, o modelo de regressão *polinomial,* numa ou mais variáveis explicativas, e linear nos parâmetros. Por exemplo, o seguinte modelo contém duas variáveis, inclui termos lineares, quadráticos e de interacção:

$$Y_i = w_0 + w_1X_{i1} + w_2X_{i1}^2 + w_3X_{i2} + w_4X_{i2}^2 + w_5X_{i1}X_{i2} + \varepsilon_i$$

Alternativamente, podem usar-se as variáveis transformadas mantendo-se a linearidade nos parâmetros, por exemplo, o seguinte modelo:

$$\log_{10} Y_i = w_0 + w_1\sqrt{X_{i1}} + w_2 \exp(X_{i2}) + \varepsilon_i.$$

Em geral pode-se exprimir este tipo de modelos na fórmula ($i = 1, 2 \ldots m$):

$$Y_i = f(\mathbf{X}_i, \mathbf{w}) + \varepsilon_i$$

em que $\mathbf{X}_i = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{i,n} \end{bmatrix}^T$ constitui a $i-$ésima linha da matriz do modelo $\mathbf{X}$, o vector de coeficientes de regressão $\mathbf{w}$ terá de ser estimado e $f(\mathbf{X}_i, \mathbf{w})$ representa o valor esperado de $\mathbf{E}[Y_i]$.

Subjacente aos modelos de regressão linear generalizados está a ideia de se obter

um modelo de regressão linear através da transformação das médias dos dados. A resolução, por exemplo, das equações de máxima verosimilhança do modelo equivale a um processo de cálculo repetido de uma regressão linear ponderada de uma variável dependente modificada sobre a matriz do modelo $\mathbf{X}$, usando-se uma função de peso que se modifica no algoritmo iterativo, e que em geral converge rapidamente quando o número de exemplos m for suficientemente grande.

O processo de construção de um modelo de regressão começa pelo *estabelecimento* de uma expressão para o referido modelo, seguindo-se uma fase de *ajustamento* dos parâmetros desta expressão a um conjunto de dados observacionais. Finalmente, entra-se numa fase de *inferência*, em se procede a uma verificação do grau de adequação do modelo como um todo, estudando-se detalhadamente em que medida as discrepâncias locais não só afectam significativamente o modelo, mas também poderão torná-lo inadequado para novos valores. Em geral, para este efeito faz-se uma avaliação dos erros das estimativas, determinam-se regiões de confiança, realizam-se testes sobre os parâmetros de interesse e analisam-se estatisticamente os resíduos.

Este processo conduziu ao estabelecimento de modelos de estimação do peso fetal, por exemplo, o modelo de Hadlock et al. (1985), em que se utilizam como variáveis "*explicativas*" as medidas obtidas por ecografia do *perímetro abdominal* (PAB, em cm) e do *comprimento do fémur* (CFE,em cm). Neste modelo o peso fetal P (em g) é dado pela fórmula:

$$\log P = 1.304 + 0.05281\,(PAB) + 0.1938\,(CFE) - 0.004\,(PAB)\,(CFE)$$

Este modelo ainda é um dos mais utilizados na prática clínica, uma vez que tem a vantagem da simplicidade e de ser rigoroso em muitas situações, tal como se refere mais detalhadamente no quarto capítulo, no entant, é considerado pelos obstetras inapropriado em muitas situações da prática clínica, podendo originar erros de previsão do peso fetal elevados, os quais podem ultrapassar $\pm 15\%$ do peso fetal Mandruzzato et al. (1994).

1.3 Analogia estrutural entre rede neuronal e estatística

Um modelo de rede neuronal emprega técnicas de *pré-acção*, por outras palavras, dispõe de um mecanismo de "*antero-alimentação*" ("*feedforward*", em inglês), que lhe permite resolver problemas de inferência em situações de classificação, regressão, estimação de densidades de probabilidade de conjuntos de dados, assim como nos chamados "*problemas de inversão*".

Nestas situações, tal como num raciocínio bayesiano, é possível ter-se uma apro-

ximação probabilística da inferência admitindo o pressuposto de que as quantidades que interessam são governadas por distribuições de probabilidade e que as decisões óptimas se podem obter através de raciocínios acerca dessas mesmas probabilidades juntamente com os dados observados. Seja, por exemplo, um conjunto de m vectores $\mathbf{x}_i$, $i = 1, \ldots m$, e seja p o número de classes C_k, $k = 1, \ldots p$, que se podem discriminar. Pode fazer-se a estimação das densidades condicionais correspondentes $p(\mathbf{x}_i \mid C_k)$ atribuindo a cada vector $\mathbf{x}_i$ uma etiqueta y_i, representativa da classe de que esse vector é membro. Se forem conhecidas estas densidades de probabilidade pode-se, usando o teorema de Bayes, determinar a probabilidade *a posteriori* de um novo vector e obter-se assim a etiqueta que melhor o classifica.

Um modelo de rede neuronal pode também ser interpretado como um modelo semi-paramétrico de estimação da densidade condicional. Existe uma relação forte, por vezes mal compreendida, entre alguns modelos de rede neuronal e os classificadores bayesianos de erro-mínimo. As saídas dos modelos de rede neuronal, nalguns casos, nem são verosimilhanças, nem são valores lógicos binários próximos de zero ou da unidade. Pelo contrário, os valores das saídas são estimativas de probabilidades *a posteriori* (bayesianas). Para um problema com k classes, as probabilidades a posteriori bayesianas estimam-se, em geral, no sentido do mínimo erro quadrático, isto é, optimiza-se um certo funcional de custo quadrático, sempre que o modelo tiver uma saída para cada uma das classes, as saídas desejadas sendo 1 de entre k possíveis. Por outras palavras, uma unidade de saída corresponde à classe correcta, todas as restantes têm valores nulos.

Estas condições verificam-se frequentemente em modelos de rede neuronal com unidades não-lineares sigmóidais, treinadas por rectropropagação, em modelos de rede de funções de base radiais, e em modelos de rede com polinómios de grau elevado treinadas com base numa função de custo erro quadrático. Quando as probabilidades bayesianas se estimam rigorosamente, se minimiza a percentagem de erros de classificação e a soma das saídas é unitária, então, as saídas da rede podem ser encaradas como probabilidades. Adicionalmente, esta forma de interpretar as saídas de uma rede neuronal como probabilidades *a posteriori* permite uma compensação das diferenças entre as probabilidades a posteriori de classe entre conjuntos de treino e de teste, fazer a combinação das saídas de vários classificadores para um nível de decisão mais elevado, usar funções de risco diferentes do risco de erro mínimo, implementar regras convencionais óptimas de rejeição de exemplos e calcular medidas alternativas da performance da rede Richard and Lippmann (1991).

1.4 Origem histórica do modelo de rede neuronal

Um modelo de rede neuronal pode ser visto como um neurocomputador, isto é, um *processador distribuído* dotado de um paralelismo massivo de muitas unidades de processamento simples, as quais lhe conferem uma propensão natural para o armazenamento de conhecimento empírico e a sua disponibilização para ser utilizado sempre que ele for necessário Haykin (1999). Pode encontrar-se algumas semelhanças entre um modelo de rede neuronal e um modelo da mente humana, particularmente em dois aspectos primordiais. O primeiro tem a ver com o facto de em ambos a fonte onde é extraída a informação ser o ambiente que os rodeia, e a aquisição dessa informação se processar por aprendizagem. O segundo aspecto tem a ver com o facto de ambos os modelos possuírem conexões interneurónicas, conhecidas pela a designação de *pesos sinápticos*, os quais têm valores variáveis que realizam o armazenamento da informação que vai sendo adquirida.

A principal fonte de inspiração do estudo dos modelos de rede neuronal no passado, como já se referiu, foi a observação dos sistemas de aprendizagem biológicos constituídos por redes complexas de neurónios interligados. Muito grosseiramente pode estabelecer-se uma analogia entre os sistemas biológicos e os sistemas de redes neuronais na medida em que estes podem conter um conjunto de unidades simples densamente interligadas. Para cada uma destas unidades convergem várias entradas, por onde podem passar sinais com valores numéricos determinados, provenientes das saídas de outras unidades da vizinhança. Por sua vez cada unidade tem uma saída que apresenta um dado valor numérico que poderá ser a entrada de uma ou mais unidades.

O desenvolvimento do modelo de rede neuronal remonta à fase da história da informática, por volta dos anos mil novecentos e quarenta, em que estava em discussão o modelo da máquina de John von Neumann (1903-57) . Já naquele tempo se colocava a questão de saber se seria possível uma aprendizagem automatizada que se auto-aperfeiçoasse a partir da sua própria experiência. A possibilidade de um computador modelar com elevada aproximação a mente humana era objecto de atenção entre matemáticos, engenheiros, físicos e biólogos que trabalhavam na criação dos sistemas artificiais de cálculo electrónico, os quais eram objecto de uma necessidade premente naqueles anos. Desta pesquisa intensiva surgiu nos E.U.A. o "*conexionismo*", o qual se constituiu numa ciência a partir do momento em que Frank Rosenblatt iniciou a publicação dos seus primeiros trabalhos em 1957. Existiram muitas outras contribuições, em particular as de Warren McCulloch e Walter Pitts que haviam proposto, já em 1943, um modelo de rede de autómatos para uma máquina calculadora. Na Europa foi principalmente Allan Turing (1912-54), destacado co-inventor do computador moderno, que em 1948 escreveu vários ensaios

sobre uma máquina inteligente ("*inteligent machinery*"). Esta máquina desconhecida, que foi por ele designada "*B-type unorganized machine*", consistia num elevado número de neurónios artificiais, com múltiplas conexões entre si, a que se associavam dispositivos com a capacidade de fazer a adaptação destas conexões aos dados provenientes do meio envolvente e que se passou a designar-se por "*rede neuronal*". O principal documento que Turing escreveu sobre este tópico só foi conhecido postumamente vinte anos após ter sido escrito Copland and Proudfoot (1999). Assim, é a Rosenblatt que se atribuíem as bases teóricas da *rede neuronal* e do *conexionismo*, podendo-se sintetizar-se o seu pensamento na seguinte frase:

> "A informação armazenada pode assumir as formas de *conexões novas*, *canais de transmissão* no sistema nervoso (ou, então, uma *criação de condições* que sejam funcionalmente equivalentes a conexões novas) Rosenblatt (1959)."

Os primeiros algoritmos neuronais para a aprendizagem supervisada, que foram propostos nos anos 1960, são essencialmente da lavra de Rosenblatt, tal como a designação do dispositivo que concretizava o princípio de aprendizagem: "*o perceptrão.*" Estes algoritmos de aprendizagem baseada em exemplos conferiam ao perceptrão uma capacidade de aprender classes de funções muito simples. Widrow introduziu aperfeiçoamentos nestes algoritmos introduzindo a famosa regra da aprendizagem dos mínimos quadrados, conhecida por regra do delta de Widrow-Hoff para a minimização do erro empírico Widrow and Hoff (1960).

Até finais da década de setenta o impulso industrial e comercial parece ter sido dirigido principalmente para o desenvolvimento de um computador baseado no modelo determinístico das máquinas propostas por Turing e von Neumann, tendo os modelos estocásticos, inspirados na física e biologia, sido relegados para um segundo plano. Tal facto poderá ter criado alguma desilusão nalguns investigadores que se interessavam pelas áreas da cibernética e neurociência, porém, o desenvolvimento sem paralelo de uma das mais sofisticadas máquinas jamais concebidas, *o microprocessador*, veio finalmente dar a possibilidade, em finais dos anos mil novecentos e oitenta, de se implementar sem qualquer tipo de restrições, seja de natureza económica ou outras, o ideal conexionista, no qual alguns investigadores acreditam estar o cerne de um possível sistema capaz de simular funções relevantes de aprendizagem já existentes no sistema neural dos primatas superiores.

Uma das mais importantes características de qualquer sistema de aprendizagem é a sua plasticidade, ou adaptabilidade, a qual constitui também um dos aspectos mais característicos dos modelos de rede neuronal. Em lugar destes modelos serem construidos em conformidade com uma lista de especificações *a priori* muito precisas e exaustivas, eles estão concebidos para se irem automaticamente ajustando em

conformidade com os dados recebidos do exterior, modificando os seus parâmetros internos de acordo com as necessidades. Isto significa que eles podem ser encarados como sistemas parametrizados, que têm "*consciência*" das suas respostas através de um anel de retroalimentação, o qual por sua vez compreende uma função de custo com a missão de avaliar o desempenho (*performance*) e que, por consequência, o sistema modifica os seus próprios parâmetros até conseguir o melhor desempenho possível, tendo em vista determinados objectivos.

As aplicações comerciais significativas de sistemas de aprendizagem automática baseados em modelos de rede neuronal começaram a surgir em força nos inícios dos anos 1990. Exemplos relevantes de aplicações civis são o reconhecimento da fala, o reconhecimento de caracteres manuscritos, a extracção de conhecimento em bases de dados massivas em áreas financeiras e industriais ("*data mining*"), e a descodificação do DNA de genomas nomeadamente o do ser humano.

1.5 Aprendizagem e indução

Em certo sentido "*aprendizagem*" e "*indução*" podem considerar-se sinónimos, diferindo apenas em que o primeiro conceito pode ser considerado mais abrangente que o segundo. Indução, literalmente, significa um raciocínio que permite estabelecer uma proposição geral a partir de alguns casos particulares, enquanto que a aprendizagem pode também processar-se por *dedução*. No entanto, para alguns autores, por exemplo Jeffreys (1939), a dedução pode ser vista como um caso particular da indução e, portanto, os dois conceitos tendem a confundir-se em muitas aplicações práticas. A demarcação entre estes dois conceitos é um problema às vezes complexo. Vai proceder-se, seguidamente, a um delineamento de alguns aspectos relevantes da indução, que se julga ajudarem a compreender outros conceitos, tais como, contingência, plausibilidade, generalização e complexidade, os quais aparecerão referidos avulsamente em capítulos que se seguem.

São bem conhecidas as *lógicas "clássicas"* que assentam nos processos de formalização típicos da lógica matemática e da teoria da demonstração. Também são reconhecidas, apesar de serem menos comuns que as primeiras, *as lógicas "não clássicas"* que assentam num conjunto de sistemas, que sem deixarem de ser também um cálculo formal, fogem a uma ou várias das características da lógica formal clássica. É o caso das lógicas *multivalentes*, a lógica *difusa* (fuzzy), assim como as lógicas opostas destas, as chamadas lógicas *de verdade não verificável*, tais como a lógica *deôntica*, a lógica *modal* e a lógica *epistémica*, por exemplo.

Entre estes dois extremos, existem outros sistemas que se encarregam da análise dos *graus de confirmação*, como é o caso da *lógica indutiva*, ou que recorrem a *critérios de plausibilidade*, caso das lógicas *não monotónicas*, ou ainda as lógicas

que empregam critérios de implicação "*enfraquecidos*" através da introdução de restrições nas regras da dedução.

Um meio de definir a *lógica indutiva* consiste em opô-la à lógica dedutiva, supostamente conhecida. Considerem-se, por exemplo, os seguintes enunciados que se enquadram no contexto desta tese, a estimação do peso fetal:

$\mathbf{A}_1$: "Todos os fetos que têm *peso baixo* têm *perímetro abdominal inferior a* 30cm ."

$\mathbf{A}_2$: "X é um feto que tem *perímetro abdominal inferior a* 30cm ."

$\mathbf{B}$: "X é um feto com *peso baixo.*"

Designando por $\mathbf{A}$ a conjunção de $\mathbf{A}_1$ e $\mathbf{A}_2$ é evidente que $\mathbf{A}$ *implica* $\mathbf{B}$. A lógica dedutiva tem por objecto relações deste tipo.

Considerem-se, agora, os enunciados do seguinte tipo:

$\mathbf{A}'_1$: "A base de dados tem 50 casos de *perímetro abdominal inferior a* 30cm."

$\mathbf{A}'_2$ "A base de dados tem 40 casos de *perímetro abdominal inferior a* 30cm *e* que têm um *peso fetal baixo.*"

$\mathbf{B}'$ "X é um feto com *peso fetal baixo.*"

Seja $\mathbf{A}'$ a conjunção das proposições $\mathbf{A}'_1$, $\mathbf{A}'_2$ e $\mathbf{A}_2$ (do enunciado anterior).

É claro que $\mathbf{A}'$ não implica dedutivamente $\mathbf{B}'$, uma vez que existem casos que têm um *perímetro abdominal inferior a* 30cm, mas podem não ter um *peso fetal baixo.* Mas, sabendo-se que se verifica a condição $\mathbf{A}'$ existe uma maior tendência para se acreditar na possível veracidade de $\mathbf{B}'$, isto é, a hipótese $\mathbf{B}'$ fica melhor apoiada pela evidência $\mathbf{A}'$ do que pela ignorância de $\mathbf{A}'$.

A lógica indutiva propõe-se dar uma formulação rigorosa a esta ideia intuitiva utilizando a teoria das probabilidades e a matemática estatística. Apesar de eminentes matemáticos do século dezassete, tais como, Pascal, Leibniz e Bernoulli, terem feito as primeiras tentativas de descrever a relação entre duas proposições, das quais uma sustenta a outra sem que a verdade da primeira seja logicamente incompatível com a falsidade da segunda, só a partir dos anos 1920 é que se deu o desenvolvimento da ideia seminal da *lógica indutiva* relacionando-a com os fundamentos matemáticos do cálculo das probabilidades. Este tópico voltará a ser objecto de mais desenvolvimentos noutros capítulos.

1.6 Analogia estrutural entre a máquina de vectores de suporte e a estatística

Os detalhes da *máquina de vectores de suporte* serão apresentados nos capítulos seguintes, aqui, vai-se apresentar apenas uma descrição sumária. A ideia de base consiste em encontrar um *hiperplano* capaz de estabelecer a separação entre um conjunto finito de pontos no espaço, cada ponto sendo classificado com uma dada *etiqueta,* de forma a que seja máxima a *margem de separação* entre os pontos cujas etiquetas são opostas. O termo *margem* quer dizer a distância mínima entre o hiperplano e um qualquer ponto do conjunto de pontos. Os modelos de rede neuronal aspiram a conseguir isto, porém, nem é fácil saber teoricamente se o poderão conseguir, nem como tal será possível, em particular no caso de os dados não serem linearmente separáveis. Os modelos baseados na *máquina de vectores de suporte* garantem sempre uma maximização da margem. A convergência é garantida, a solução é única, e o tempo de computação é aceitável (graças às elevadas frequências de trabalho, da ordem dos Megahertz, dos processadores que equipam os computadores actuais).

Uma vez que na maioria dos casos os conjuntos de pontos a classificar não são linearmente separáveis recorre-se a uma "*mudança de coordenadas*", quer dizer, acrescentam-se novas dimensões, ou atributos suplementares aos que já existem no espaço de entradas, em grande número, sendo cada novo atributo definido a partir de uma função dos já existentes. A realizabilidade dos cálculos num espaço com uma tão "*amaldiçoada*" dimensionalidade é sempre possível graças ao "*truque do kernel*". Uma boa escolha deste *kernel* pode ser crucial, sendo este um dos pontos fracos deste tipo de modelação, uma vez que ele torna a amostra linearmente separável no novo espaço. A designação "*vectores de suporte*" significa que existe um conjunto de vectores que de facto definem o hiperplano separador, os quais deverão ser uma percentagem pequena do conjunto de pontos de treino, e pertencer a uma dada vizinhança definida por uma distância mínima ao hiperplano separador .

Os modelos de *máquina de vectores de suporte* não só parecem mostrar ter uma mais sólida fundamentação estatística que os de rede neuronal, mas também são capazes de os substituir em quase todas as aplicações podendo até ter um funcionamento melhor numa vasta gama de aplicações. Estes modelos baseados nos métodos *kernel* também podem ser vistos na perspectiva da teoria da regularização e das redes de funções de base radiais, no entanto, é muito difícil estabelecer-se um relacionamento entre as ideias originais de Rosenblatt e o conceito matemático de perceptrão dos anos 1960, e as ideias surgidas a partir dos anos 1990 subjacentes aos modelos de *máquina de vectores de suporte.*

1.7 Contribuições

Os capítulos seguintes desta dissertação procuram descrever os caminhos que se utilizaram para tentar obter uma solução do problema dos pesos fetais com modelos de rede neuronal e máquina de vector de suposte. Os principais resultados, que se ordenaram segundo o seu grau de importância, são os seguintes:

- Redução do erro relativo de estimação do peso fetal para cerca de $\pm 7\%$ quando os valores admissivéis na prática clínica para este erro são actualmente de $\pm 15\%$.
- Estabelecimento de um modelo de rede neuronal de perceptrão de multicamada, PMC, que apresentou um erro relativo médio de estimação do peso fetal de $\pm 7.17\%$.
- Estabelecimento de um modelo de rede neuronal de funções de base radiais, FBR, que apresentou um erro relativo médio de estimação do peso fetal de $\pm 6.95\%$.
- Estabelecimento de um modelo de máquina de vectores de suporte, MVS, que apresentou um erro relativo médio de estimação do peso fetal de $\pm 6.71\%$.
- Concepção e realização de um calculador do peso fetal em $Visual - Basic$ que permite testar na prática clínica os modelos estabelecidos.
- Projecto de um modelo de estimação treinado por critério entrópico, baseando-se na teoria da informação tem em vista a obtenção de um único estimador combinatório e cooperativo de três modelos estimadores PMC, FBR e MVS.
- Concepção e implementação de um método de edição e pré-processamento de dados por sobre-ajustamento de uma máquina de kernel aos exemplos e detecção dos casos desviantes quando estes mesmos exemplos são reprocessados pela máquina.
- Interpretação e tradução original para língua portuguesa dos principais conceitos da teoria estatística da aprendizagem aplicados à aprendizagem supervisada de funções reais.
- Interpretação, na perspectiva da teoria da aprendizagem, e tradução original para língua portuguesa dos princípios de funcionamento, propriedades e algoritmos dos modelos baseados na máquina de vectores de suporte e de rede neuronal.

1.8 Organização da tese

O capitulo *segundo* aborda de uma forma estruturada aspectos relevantes e comuns às teorias combinatória e estatística da aprendizagem e faz uma primeira abordagem dos principais aspectos teóricos relativos à selecção dos modelos. Procede-se a uma revisão e interpretação, incluindo uma tradução original para língua portuguesa com propostas de alguns neologismos, dos principais conceitos da teoria estatística da aprendizagem aplicados à aprendizagem supervisada de funções reais.

O capítulo *terceiro* foca os modelos que se considera terem mais potencialidades de constituirem uma solução com sucesso, as diferentes soluções baseadas na máquina de vectores de suporte que permitem maximizar a margem de separação, a teoria da regularização que permite satisfazer a necessidade de fazer uma restrição ao conjunto de funções que podem ser solução de um problema de aprendizagem, a aproximação de funções com os modelos de rede neuronal dos tipos *rede de funções de base radiais* e *perceptrão de multicamada*. Este capítulo contém ainda uma narrativa sobre o funcionamento, propriedades e algoritmos da máquina de vectores de suporte e uma interpretação, incluindo tradução original para língua portuguesa e proposta de neologismos, sobre aspectos teóricos do modelo de rede neuronal, na perspectiva da teoria da aprendizagem.

O capítulo *quarto* tem por objectivo estabelecer algum conhecimento *a priori* sobre a biometria fetal. Faz-se uma revisão de literatura de alguns trabalhos de estimação do peso fetal clássicos e uma análise exploratória dos dados recolhidos, uma análise exploratória dos dados colhidos em quatro serviços de obstetrícia de hospitais portugueses numa perspectiva clássica e com base numa modelação com regressores multilineares. Esboça-se o potencial interesse da chamada aprendizagem bayesiana da densidade de probabilidade e da determinação da fronteira de decisão segundo o princípio da máxima verosimilhança.

O capítulo *cinco* refere experiências conducentes à concepção e optimização da performance de modelos estimadores de rede neuronal PMC ou FBR e *máquina kernel*, MVS. Procede-se à descrição, concepção e implementação do método de pré-processamento e edição de dados, e apresentam-se os resultados práticos obtidos.

O capítulo *seis* faz uma revisão dos aspectos mais relevantes para a aprendizagem da teoria da informação, a entropia, a entropia relativa e a informação mútua, segundo Shannon e Reyni. Esboça-se o método não-paramétrico de determinação de funções de distribuição, designado método das janelas de Parzen (gaussianas). Finalmente, descrevem-se três experiências de estimação do peso fetal com modelos PMC treinados por um critério informático-teorético, e apresenta-se em esboço teórico e alguns resultados preliminares de um estimador combinatório e cooperativo de três modelos estimadores.

O capítulo *sétimo,* e último, contém uma discussão de três temas e a conclusão. O primeiro tema que se discute são os resultados das experiências com modelos de rede neuronal e de máquina de vectores de suporte apresentados no quinto e sexto capítulo. O segundo tema de discussão é constituído por quatro aspectos da *teoria estatística da aprendizagem* vistos segundo uma perspectiva epistemológica. Finalmente, no terceiro tema de discussão, aborda-se a evidência de benefícios na tomada de decisão em obstetrícia provenientes, primeiro, da utilização de modelos de rede neuronal e inferência estatística, segundo, de um projecto de um estimador de peso fetal por rede neuronal implementado em *Visual-Basics* para Windows e, finalmente, de um projecto de desenvolvimento no futuro de uma técnica de pesquisa por conteúdo e *data mining* de uma base de dados clínicos e e imagens.

Capítulo 2

Aprendizagem Supervisada

Introdução

Este capítulo contém uma revisão de conceitos teóricos relacionados com o princípio de funcionamento e as técnicas para controlar a capacidade de generalização de algoritmos de aprendizagem supervisada. Estes conceitos aplicam-se na optimização da performance de modelos de *rede neuronal* e a *máquina kernel.*

Esboçam-se alguns procedimentos propostos recentemente, tendo por base a também recente *teoria estatística da aprendizagem,* para o estabelecimento da relação entre um algoritmo de aprendizagem e a sua capacidade de generalização.

Na secção 2.1 faz-se uma abordagem de métodos de análise da capacidade de famílias de funções, focando-se a importância do conceito de *dimensão de Vapnik Chervonenkis.* Referem-se, a seguir, os conceitos de algoritmo de aprendizagem minimizador do erro amostral, função de crescimento, número de cobertura, pseudo-dimensão e FAT-estilhaçamento.

Na secção 2.2 abordam-se alguns conceitos ligados à *aprendizagem de funções reais* na perspectiva da teoria estatística da aprendizagem, definem-se, no contexto de problemas de regressão, os conceitos de *função de custo*, de *funcional de risco* e de *consistência* de um processo de aprendizagem e, finalmente, enunciam-se os princípios de *minimização do risco empírico* e de *minimização do risco estrutural.*

Na secção 2.3 abordam-se alguns tópicos sobre a capacidade de generalização de um algoritmo que se podem designar genricamente por *selecção de modelos.* Refere-se o compromisso *viés-variância* no caso da regressão e discute-se os critérios clássicos de Akaike, Craven-Wahba, Shibata e Rissanen para a selecção de modelos.

2.1 A Dimensão de Vapnik-Chervonenkis

2.1.1 O problema da aprendizagem

O processo de procura de um modelo que se ajuste a dados empíricos e possua uma boa capacidade de generalização é, por vezes, um problema que não é fácil de resolver, uma vez que pertence ao tipode problemas que se costumam designar "incompletamente formulados". Seja, por exemplo, a estimativa da regressão em que se considera o problema de determinar uma função real de várias variáveis reais, a partir de medições indirectas Vapnik (1998) (p.28). Supondo que se pretende estimar uma função $f(t)$ que pode ser medida no ponto t e, ao mesmo tempo, tem-se uma outra função $F(x)$, em que se fizeram medidas com um certo erro, tal que $y_i = F(x) + \varepsilon_i$. Admitindo que esta função se relaciona com $f(t)$ através de um operador A tal que, $Af(t) = F(x)$, pode acontecer que, tendo-se uma boa aproximação da função de regressão $F(x)$, não se consiga obter uma correspondente aproximação da função desejada $f(t)$, porque operador inverso A^{-1} não possui certas propriedades que garantem a existência de uma solução estável.

Uma solução possível para problemas deste tipo consiste em utilizar-se um processo de *indução,* que permita construir modelos locais com uma boa expectativa de generalização, isto é, que originem respostas correctas, ou com erros muito reduzidos, quando se introduzem exemplos novos que não tenham sido processados anteriormente. A construção de um modelo por um processo de indução depende em grande medida da quantidade e da qualidade dos exemplos do conjunto de treino. As soluções podem não ter a capacidade de generalização desejada, devido a que a cardinalidade do conjunto de treino é finita, e pode acontecer que a amostragem dos exemplos padeça de não-uniformidade e escassez de exemplos nalgumas regiões do espaço dos vectores de entrada da função de distribuição conjunta, à qual os exemplos pertencem.

Uma das formas de realizar um processo de indução consiste em usarem-se redes neuronais artificiais dotadas de algoritmos de optimização de uma dada medida da performance. Estas redes uma vez treinadas podem ficar com uma capacidade de suavização dos resultados insuficiente devido a terem utilizado uma medida estatística dos erros inadequada. Constroem-se, assim, modelos que se sobreajustam demasiado aos dados de aprendizagem e, por consequência, perdem uma grande parte da sua capacidade de generalização.

Em geral considera-se existir uma equivalência entre os *processos de indução* e de *aprendizagem*, uma vez que ambos os procesos visam encontrar uma relação de dependência usando um número limitado de observações, casos, ou exemplos. Um processo de indução implica uma aprendizagem da relação de dependência existente entre dois conjuntos, um composto por *vectores de entrada*, e o outro por "*etiquetas*",

quando se conhece um número finito de exemplos, isto é, de instâncias da correspondência entre um vector de entrada e a respectiva etiqueta. Estas instâncias, que globalmente constituem o chamado *conjunto de treino*, admitem várias designações, tais como, padrões, exemplos, casos ou observações.

A aprendizagem baseada em exemplos processa-se segundo um modelo de aprendizagem (ver figura 2.1), constituído pelo seguinte terno de constituintes principais Vapnik (1995):

1. Um *gerador* de exemplos aleatórios, $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)...,(\mathbf{x}_m, y_m)$, isto é, pares constituídos por um vector de números reais $\mathbf{x}_i$ e uma etiqueta y_i, $(\mathbf{x}_i, y_i) \in \mathbf{R}^n \times \mathbf{R}$, obtidos de forma independente numa população em que se supõe existir uma determinada função de distribuição de probabilidade P que se desconhece.

2. Um *supervisor* S,tal que para cada vector de entrada $\mathbf{x}_i$ fornece um valor de saída, a etiqueta y_i, em conformidade com uma função que se supõe existir para a distribuição de probabilidade condicional $P(y|\mathbf{x})$ mas que se desconhece.

3. Um *algoritmo de aprendizagem* $L(\mathbf{z})$capaz de implementar uma dada classe de funções, ou hipóteses, $h(\mathbf{x}, \mathbf{w})$, em que $\mathbf{w}$ é um vector de parâmetros, $\mathbf{w} \in \mathbb{W}$, e $\mathbb{W}$ é um espaço ao qual esses parâmetros pertencem

Pode enunciar-se o problema da aprendizagem em termos de uma busca de uma função particular numa classe de funções, $h(\mathbf{x}, \mathbf{w})$, $\mathbf{w} \in \mathbb{W}$, que se apresenta como sendo a melhor aproximação às respostas fornecidas pelo supervisor S, e cuja selecção tem por base unicamente o conjunto de m exemplos, $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, independentes e identicamente distribuídos($i.i.d.$).

Supondo que se tem um universo U com um número de exemplos $(\mathbf{x}_i, y_i)$, $i = 1, 2, ..., m_U$, em que m_U é muito superior a m. As sucessivas ocorrências do acontecimento "*extrair um exemplo aleatoriamente do universo U*" dizem-se que são *independentes* se a probabilidade de sair um dado exemplo num ensaio é independente do exemplo que saiu nos ensaios anteriores, ou dos exemplos que vierem a sair nos ensaios seguintes. Esta expressão subentende que as probabilidades são as mesmas para todos os ensaios. Numa amostragem esta série de ensaios é por vezes referida como "*ensaios de Bernoulli*", quando se faz referência ao lançamento de um dado ou uma moeda. Recorde-se que neste caso se verifica um caso particular da lei do produto das probabilidades compostas, $P(z_1, z_2) = P(z_1) \,.P(z_2)$, uma vez que $P(z_1|z_2) = P(z_1)$ e $P(z_2|z_1) = P(z_2)$. Por outro lado, dizer-se que os exemplos são identicamente distribuídos quer dizer que os ensaios foram sempre realizados no universo U e que todos os exemplos deste universo têm igual probabilidade de serem extraídos.

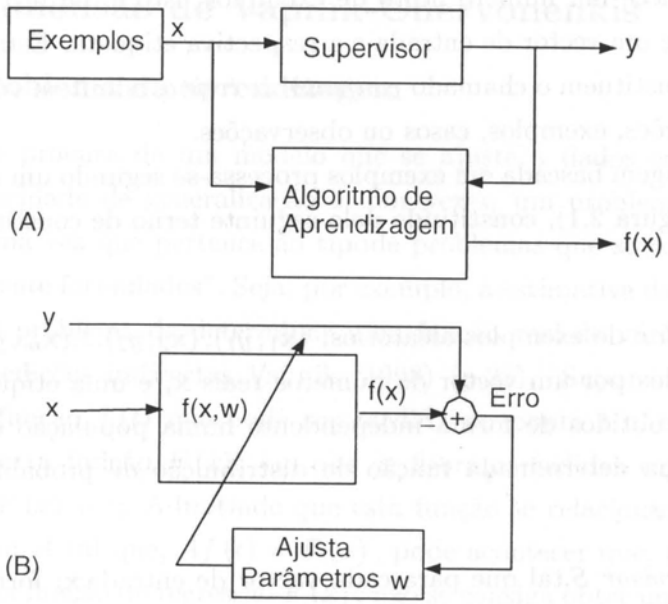


Figura 2.1: (A) Modelo de Aprendizagem Supervisada (B) Esquema de Algoritmo de Aprendizagem

Vai, agora, considerar-se os exemplos definidos num espaço de amostragem, que se designará por $\mathbb{Z}$, tal que $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, em que $\mathbb{X} \subset \mathbf{R}^n$ é o espaço das entradas, em geral sendo um subconjunto do espaço n-dimensional euclidiano usual, e $\mathbb{Y}$ é o espaço das respostas, em geral sendo um conjunto de números reais, o conjunto das "*etiquetas*". O espaço $\mathbb{Y}$ pode ser uma qualquer a partição dicotómica de um dado conjunto, um conjunto finito de números inteiros positivos ou um intervalo $[a, b] \in \mathbf{R}$ definido no conjunto dos números reais.

O conjunto de *exemplos*, $\mathbf{Z}$, é um conjunto constituído por m pares ordenados de vectores $\mathbf{z}_i=(\mathbf{x}_i, y_i)$ e $i = 1, 2, ..., m$. Nalguns textos recentes, por exemplo Scholkopf and Smola (2002), um vector de entrada $\mathbf{x}_i$ é também denotado $[\mathbf{x}]_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$, em que x_i. é um qualquer dos n componentes. Sem perda de generalização possível a um vector admite-se que a etiqueta y_i é um número real. Pode escrever-se que:

$$\mathbf{Z} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)..., (\mathbf{x}_m, y_m)\} = \\ = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m) \in \mathbb{Z}^m.$$

Habitualmente usa-se para conjunto de *exemplos* uma representação matricial e

denota-se com uma letra maiúscula:

$$\mathbf{Z} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} & y_1 \\ x_{21} & x_{22} & \dots & x_{2n} & y_2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} & y_3 \end{bmatrix}$$

O termo *amostra* a adjectivar um conjunto de exemplos pretende realçar o facto de cada exemplo etiquetado $z_i = (\mathbf{x}_i, y_i)$ ser extraído independentemente num espaço de probabilidade P. Aqui P designa uma medida de probabilidade numa dada σ-álgebra apropriada, ao conjunto $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$. Usa-se P^m para designar a medida de probabilidade conjunta. As funções de probabilidadedefinidas no conjunto $\mathbb{Z}$, ou nos seus subconjuntos, e que se vão mencionar seguidamente, pressupõem que são medíveis sem se referir explicitamente.

Alternativamente, também se podem ver os espaços $\mathbb{X}$ e $\mathbb{Y}$ como variáveis aleatórias, $\mathbf{X}$ e Y, respectivamente, num qualquer, espaço de probabilidades, não especificado, nomeadamente, quando se trata de vectores reais, cujas componentes são medidas de grandezas. Neste caso P é visto como uma distribuição conjunta em X e Y. A probabilidade de um conjunto $T \subset \mathbb{Z}$, sendo $\mathbf{z} = (\mathbf{x}, y)$, $\mathbf{x} \in \mathbf{X}$, e $y \in Y$, é definida por :

$$P(T) = \int_T dP(\mathbf{z})$$

e o valor esperado (ou esperança matemática) da função$f(\mathbf{z})$ em $\mathbb{Z}$ denota-se:

$$\mathbf{E}(f) = \int_{\mathbb{Z}} f(\mathbf{z})\, dP(\mathbf{z}).$$

Quando $\mathbb{Z}$ é um conjunto discreto, com um certo abuso de notação, usa-se P para a função massa da probabilidade, isto é, para $z \in \mathbb{Z}$, pode usar-se $P(z)$, para designar $P(\{z\})$. Por consequência pode usar-se neste caso as expressões:

$$P(T) = \sum_{z \in T} P(z)$$

$$\mathbf{E}(f) = \sum_{z \in \mathbb{Z}} f(z)\, P(z).$$

Quando $\mathbb{Z}$ é um conjunto contínuo, a densidade associada com P (se existir) designa-se por p.

As etiquetas, valores alvo ou desejados que caracteriza a *aprendizagem supervisada,* podem ser elementos: (i) De um conjunto com apenas dois elementos, por exemplo, $\{0, 1\}$ ou $\{-1, 1\}$, estando-se assim no caso de uma *classificação* em duas classes (também designada por classificação *binária*, partição *dicotómica* ou *dicoto-*

mia); (*ii*) De um conjunto finito com k classes que poderão estar ordenadas *estando-se assim neste último caso numa situação de aprendizagem preferencial.* (*iii*) De um intervalo $[a, b] \in \mathbf{R}$, do conjunto dos números reais (caso da *regressão).* (*iv*) De conjuntos sem estrutura, tais como, caracteres, sucessões de caracteres, imagens (esta situação está na transição para a aprendizagem *supervisada* e *não supervisada,* também conhecida por "*clustering*").

2.1.2 Erro verdadeiro, amostral e de aproximação

Seja um conjunto $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, no qual se define uma classificação binária, $\mathbf{x} \in \mathbb{X}$, $y \in \mathbb{Y} = \{-1, 1\}$, e uma dada uma função $h(\cdot) \in \mathbb{H}$, sendo $\mathbb{H}$ um conjunto de funções.

Definição 1 (Erro verdadeiro) *O erro verdadeiro dado pela função h com respeito à função de distribuição P do universo de onde os exemplos provêm, define-se como sendo a probabilidade da ocorrência em $\mathbb{Z}$ de um par $(\mathbf{x},y)$ em que a imagem de $\mathbf{x}$ dada por h é diferente da etiqueta y:*

$$er_P(h) = \Pr\{(\mathbf{x},y) \in \mathbb{Z} \mid h(\mathbf{x}) \neq y\}$$

Para um dado exemplo $(\mathbf{x}, y)$ extraído aleatoriamente de acordo com a distribuição P,esta expressão dá a probabilidade da função h classificar esse exemplo com uma etiqueta errada, $h(\mathbf{x}) \neq y$. Por outras palavras, pode dizer-se que o número $1 - er_P(h)$ dá uma medida da precisão com que a função h é capaz de estabelecer uma correspondência correcta entre um exemplo $\mathbf{x}$ e a sua etiqueta y, desde que $(\mathbf{x},y)$ seja gerado em conformidade com a distribuição P.

Uma outro tipo de erro, relacionado com anterior, é o *erro amostral* que se pode sempre determinar para uma dada amostra de treino $\mathbf{Z} \in \mathbb{X} \times \mathbb{Y}$.

Definição 2 (Erro amostral) *O erro amostral de uma hipótese h que se verica num conjunto de m exemplos, $\mathbf{Z}$, pertencentes a $\mathbb{X} \times \mathbb{Y}$, define-se como sendo o número total de exemplos para os quais a função h origina uma resposta, ou classificação errada.*

No caso de $(\mathbf{x}_i,y_i) \in \mathbb{X} \times \{-1, 1\}$ pode determinar-se o erro amostral, também conhecido como *erro de observação,* ou *erro empírico,* pela expressão:

$$\widehat{er}_{\mathbf{z}}(h) = \frac{1}{2m} \sum_{i=1}^{m} (1 - h(\mathbf{x}_i)\, y_i) \tag{2.1}$$

Esta expressão admite uma interpretação trivial em termos da proporção de exemplos etiquetados $(\mathbf{x}_i, y_i)$, existentes na amostra de treino $\mathbf{Z}$, para os quais a função h origina uma estimativa errada para a etiqueta (ou classe). O erro amostral

$x\,(cm)$	26	28	30	32	34	36	38
$y=-1$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	0	0
$y=1$	0	0	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

Tabela 2.1: Table Caption

apresenta uma grande utilidade uma vez que permite determinar facilmente a partir dos exemplos de treino uma estimativa do erro verdadeiro (ver Figura 2.2 do exemplo seguinte).

Exemplo 3 (classificação de pesos fetais) *Seja um conjunto* $\mathbb{Z}=\mathbb{X}\times\{-1,1\}$, *em que* $\mathbb{X}$ *é um conjunto de valores de perímetros abdominais* x *de um feto e a lei de distribuição* P *para* $\mathbb{Z}$ *é a seguinte:*

Para o problema em vista não é necessário saber se as etiquetas obedecem ao critério de ser $y=1$ *se o peso fetal* $\geq 2,5\,kg$, *e* $y=-1$ *no caso contrário. Apenas se incluí para dar um contexto próximo do real. Suponha-se agora que se tem uma classe de funções* h *estimadoras do peso fetal:*

$$h(x,w.b)=signum\left[(w\,x+b)\right]$$

Optimizaram-se (mal para uma melhor ilustração) os parâmetros: $w=0.15$ *e* $b=-2.4$. *Nesta hipótese, o erro verdadeiro calcula-se:*

(*i*) *determinando as imagens dadas por* $h\,(x,0.15,-2.4)$

$$h([26,28,30,32,34,36,38])=[-1,-1,-1,-1,1,1,1]$$

(*ii*) *Atendendo à função de distribuição e aos casos em que* h *origina erros :*

$$er_P(h)=P\{\,(x,y)\in\mathbb{Z}\mid h(x)\neq y\,)\}=P\left\{(30,1)\cup(32,1)\cup(34,-1)\right\}=\frac{3}{10}$$

Suponha-se agora que se extraíram aleatoriamente cinco casos z_i, *de um conjunto que simula a distribuição* P, *constituídos por pares de perímetros abdominais* x_i *e as correspondentes etiquetas* $y_i\in\{-1,1\}$:

$$z_1=\left[\begin{array}{cc}26 & -1\end{array}\right];\ z_2=\left[\begin{array}{cc}28 & -1\end{array}\right];\ z_3=\left[\begin{array}{cc}30 & -1\end{array}\right];$$
$$z_4=\left[\begin{array}{cc}30 & 1\end{array}\right];\ z_5=\left[\begin{array}{cc}34 & 1\end{array}\right]$$

O erro amostral determina-se:

$$\widehat{er}_{\mathbf{z}}\,(h)=\frac{1}{2m}\sum_{i=1}^{m}\left(1-h\left(\mathbf{x}_i\right)\,y_i\right)=\frac{1}{2\cdot 5}\left(0+0+0+2+0\right)=\frac{2}{10}$$

1	−1	−1	−1	−1	−1	1	1	11	−1	−1	−1	1	1	1	1
2	−1	−1	−1	−1	−1	1	1	12	−1	−1	−1	1	1	1	1
3	−1	−1	−1	−1	−1	1	1	13	−1	−1	−1	1	1	1	1
4	−1	−1	−1	−1	1	1	1	14	−1	−1	−1	1	1	1	1
5	−1	−1	−1	−1	1	1	1	15	−1	−1	−1	1	1	1	1
6	−1	−1	−1	−1	1	1	1	16	−1	−1	1	1	1	1	1
7	−1	−1	−1	−1	1	1	1								
8	−1	−1	−1	1	1	1	1								
9	−1	−1	−1	1	1	1	1								
10	−1	−1	−1	1	1	1	1								

Tabela 2.2: Classificações(exemplo)

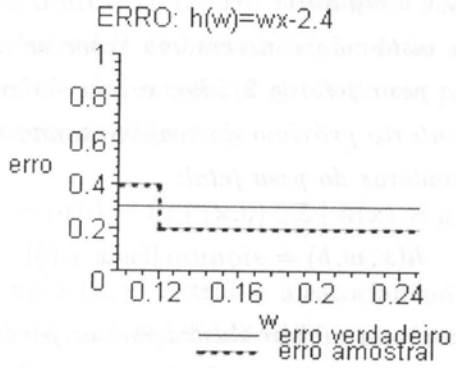


Figura 2.2: Erro amostral e verdadeiro.

Agora, determinam-se as classificações dadas por 16 *funções* h *para* $w = 0.10$ *até* 0.25 *com acréscimos de* 0.01, *estão representadas na Tabela 2.2 e figura 2.2* .

Aborda-se agora um tipo de erro que, tendo em consideração a função de distribuição de probabilidades P, se pode interpretar como sendo uma medida do grau de aproximação que a melhor função da classe de funções $\mathbb{H}$ apresenta com a verdadeira relação funcional realmente existente entre os vectores de entrada $\mathbf{x}$ e as respectivas etiquetas y dos exemplos de treino.

Definição 4 (Erro de aproximação) *Seja uma classe de funções* $\mathbb{H}$, *suponha-se conhecida a função de distribuição de probabilidades* P *do gerador de exemplos e seja* h *uma hipótese fornecida por uma dada máquina de aprendizagem ao fim de uma sucessão finita de passos do respectivo algoritmo de aprendizagem ou* treino. *Admitindo que a máquina de aprendizagem minimiza o erro verdadeiro* $\widehat{er}_P(h)$, *i.é,*

determina a função h tal que:

$$opt_P(\mathbb{H}) = \inf \widehat{er}_P(h),$$

então, o erro de aproximação *cometido pela classe de funções* $\mathbb{H}$ *é o segundo membro desta expressão,* $\inf \widehat{er}_P(h)$,

Dado o número real positivo $\varepsilon \in [0,1]$, que se designa *parâmetro de precisão*, diz-se que uma dada hipótese h é *ε-satisfactória* para P se a máquina de aprendizagem produzir $h \in \mathbb{H}$, tal que o erro verdadeiro:

$$er_P(h) < opt_P(\mathbb{H}) + \varepsilon$$

Em virtude da forma probabilística de geração das amostras de treino pode acontecer que uma dada amostra com muitos exemplos que não sejam representativos da distribuição P possa enganar o algoritmo de aprendizagem. À partida nunca se pode garantir que uma dada hipótese óptima h seja sempre *ε-satisfactória.* Contudo pode-se assegurar que ela seja *ε-satisfactória* com uma elevada probabilidade, especificamente, com uma probabilidade de pelo menos $1-\delta$, no qual δ é um *parâmetro de confiança* que pode tomar valores no intervalo $[0,1]$.

2.1.3 Formalização de conceitos chave sobre aprendizagem

As definições seguintes permitem apresentar um significado para o conceito seminal de *algoritmo de aprendizagem.* Um algoritmo de aprendizagem aplica o conjunto de exemplos de treino, através um funcional apropriado, ao conjunto dos números reais $\mathbf{R}$, de forma a determinar uma hipótese $h \in H$ para a qual, com um número de exemplos suficientemente grande, se pode verificar com uma probabilidade de pelo menos $1-\delta$ ser *ε-satisfactória* para P.

Definição 5 (Algoritmo de aprendizagem) *Seja* $\mathbb{H}$ *uma classe de funções que aplicam um conjunto* $\mathbb{X}$ *em* $\{-1,1\}$, *um algoritmo de aprendizagem para* $\mathbb{H}$ *é* uma função

$$L : \bigcup_{m=1}^{\infty} \mathbb{Z}^m \to \mathbb{H},$$

definida no conjunto dos exemplos de aprendizagem e cujo contradomínio é a classe de funções $\mathbb{H}$, *se existir uma função* $\varepsilon_0(m,\delta)$, *tal que, para um número inteiro* m, *que denota o tamanho do conjunto de treino, um número real* $\delta \in [0,1]$, *que denota o grau de confiança, e para uma dada função de distribuição* P, *para todo o exemplo* $\mathbf{Z} \in Z^m$ *escolhido em conformidade com* P^m *se verifica com uma probabilidade de*

pelo menos $1-\delta$ *que o erro verdadeiro:*

$$er_P(L(\mathbf{Z})) < opt_P(\mathbb{H}) + \varepsilon_0(m,\delta),$$

sendo $\varepsilon_0(m,\delta)$ *convergente para zero quando o número de exemplos cresce indefinidamente, i.é,* $m \to \infty$.

A função $\varepsilon_0(m,\delta)$ pode interpretar-se como sendo um *extremo superior (bound)* do *erro de estimação* dado pelo algoritmo L, podendo transformar-se permitindo obter-se um extremo superior da complexidade amostral $m_0(\varepsilon,\delta)$. O algoritmo de aprendizagem L deve originar sempre uma solução, para qualquer escolha dos parâmetros de precisão e confiança, ε e δ, respectivamente.

O valor da complexidade amostral $m_0(\varepsilon,\delta)$ para a qual L é capaz de realizar uma ε,δ-*aprendizagem* da classe de funções $\mathbb{H}$, é variável consoante os valores que de ε,δ, os quais, de certo modo, exprimem o grau de dificuldade da aprendizagem. Por definição, o *tamanho suficiente da amostra,* $m_0(\varepsilon,\delta)$, não depende da distribuição de probabilidade P. Poderia não ser assim, mas tal é vantajoso uma vez que P é desconhecida e, portanto, fixado uma amostra de treino deixa de ser relevante saber-se qual é a verdadeira distribuição de onde provém essa amostra. É claro que se esta suposição não for ser verdadeira, então, a aprendizagem pode não ser independente da distribuição, mas, se for reduzida a complexidade da classe de funções geradas pelo algoritmo de aprendizagem, como é o caso das redes neuronais, admite-se que a suposição é aceitável.

A aprendizagem supervisada de redes neuronais pode ter os três problemas:

1. *Aproximativo*, no qual se procura obter a melhor função h que minimiza o erro amostral $\widehat{er}_{\mathbf{z}}(h)$, dado por um funcional que pode ser, por exemplo, a o erro médio quadrático
$$c_{\mathbf{z}}(h) = \mathbf{E}\left[(h(\mathbf{x}) - y)^2\right].$$

2. *Estimativo*, no qual se determina o resíduo que se espera obter devido ao facto de o algoritmo de aprendizagem $L(\mathbf{Z})$ ser treinado apenas com um número finito m de exemplos:
$$\text{resíduo} = er_P(L(\mathbf{Z})) - opt_P(\mathbb{H})$$

3. *Computacional*, no qual se traduz a ignorância à partida de qual será o processo mais eficiente de se escolher a função h mais adequada.

Define-se, agora, um outro conceito que desempenha uma importância fundamental para se determinar se é possível a aprendizagem de um dado conjunto de

funções $\mathbb{H}$. Uma medida da eficiência de um algoritmo de aprendizagem L é o tamanho mínimo da amostra $m_0(\varepsilon, \delta)$ suficiente para uma aprendizagem em conformidade com os níveis prescritos para a precisão ε, e para um grau de confiança δ.

Definição 6 (Complexidade amostral) *Seja um algoritmo de aprendizagem L, a* complexidade amostral *é uma função das variáveis ε e δ, $m_L(\varepsilon, \delta)$. Sempre que exista um número inteiro mínimo, $m_0(\varepsilon, \delta)$, referido na definição 5, tal que o tamanho mínimo para um conjunto de treino suficiente para uma $(\varepsilon, \delta)-$*aprendizagem *de $\mathbb{H}$ é $m_L(\varepsilon, \delta)$. Se este tamanho mínimo for calculado para todos os algoritmos de aprendizagem de uma dada classe de funções $\mathbb{H}$, chama-se* complexidade amostral inerente ao problema de aprendizagem de $\mathbb{H}$. *e representa-se por $m_{\mathbb{H}}(\varepsilon, \delta)$.*

Definição 7 (Erro de estimação) *Seja um dado algoritmo de aprendizagem L, o* erro de estimação, *$\varepsilon_L(m, \delta)$, é o valor mínimo que a função $\varepsilon(m, \delta)$ pode ter. Igualmente, pode dizer-se que $\varepsilon_{\mathbb{H}}(m, \delta)$ é* o erro de estimação inerente ao problema de aprendizagem de $\mathbb{H}$.

A complexidade amostral tem a ver com o problema de se saber o número mínimo de exemplos necessários e suficientes para que o treino possa conduzir a um valor mínimo do erro amostral. Uma vez que o erro de aproximação é dado pelo valor de $opt_P(\mathbb{H})$, como se disse atrás, uma solução quantitativa que satisfaz o problema da estimação pode ser obtida determinando $\varepsilon_{\mathbb{H}}(m, \delta)$, ou $m_{\mathbb{H}}(\varepsilon, \delta)$.

2.1.4 Aprendizagem de classes finitas de funções

O objectivo da aprendizagem consiste em determinar-se uma função h, em $\mathbb{H}$, que tenha o mais pequeno erro verdadeiro possível, $er_P(h)$. Uma vez que os erros verdadeiros das funções de $\mathbb{H}$ são desconhecidos, parece natural que se usem os erros amostrais como estimativas daqueles. Pode caitar-se que se uma função h tiver um erro amostral pequeno, então, o seu erro verdadeiro também será pequeno. Para um algoritmo de aprendizagem L verifica-se que:

$$er_P(L(\mathbf{Z})) = \min_{h \in \mathbb{H}} \widehat{er}_{\mathbf{Z}}(h)$$

Teorema 8 (Convergência uniforme I) *Seja h uma função do conjunto $\mathbb{X}$ e contradomínio $\{0, 1\}$. Então, verifica-se:*

$$P^m\{|\widehat{er}_{\mathbf{Z}}(h) - er_P(h)| \geq \varepsilon\} \leq 2\exp\left(-2\varepsilon^2 m\right)$$

para qualquer distribuição P, qualquer ε, e qualquer inteiro positivo m.

Este resultado mostra que dada uma função $h \in \mathbb{H}$ e dada um conjunto de exemplos suficientemente grande, o *erro amostral* de h está muito próximo do *erro verdadeiro*. Isto não deve ser surpreendente uma vez que é precisamente uma consequência da *lei dos grandes números* da teoria da probabilidade, que exprime a rapidez com que as extremidades de uma distribuição binomial tendem para zero.

Enuncia-se, seguidamente, um primeiro teorema sobre um resultado da aprendibilidade.

Teorema 9 (Complexidade amostral I) *Seja $\mathbb{H}$ um conjunto* finito *de funções que classificam um conjunto $\mathbb{X}$ num conjunto das etiquetas $\{0,1\}$. Seja L um algoritmo de aprendizagem:*

$$L : \bigcup_{m=1}^{\infty} \mathbb{Z}^m \to \mathbb{H}$$

tal que, para qualquer m e qualquer $\mathbf{Z} \in \mathbb{Z}^m$

$$\widehat{er}_{\mathbf{Z}}(L(\mathbf{Z})) = \min_{h \in \mathbb{H}} \widehat{er}_{\mathbf{Z}}(h).$$

Então, L é um algoritmo de aprendizagem para $\mathbb{H}$ com erro de estimação

$$\varepsilon_L(m,\delta) \leq \left(\frac{2}{m} \ln\left(\frac{2\,|\mathbb{H}|}{\delta}\right)\right)^{1/2}$$

e complexidade amostral

$$m_L(\varepsilon,\delta) \leq \frac{2}{\varepsilon^2} \ln\left(\frac{2\,|\mathbb{H}|}{\delta}\right)$$

Anthony and Bartlett (1999), (p.21).

A suposição de que $\mathbb{H}$ é um conjunto *finito* de funções nem sempre é válida, no entanto, um algoritmo de aprendizagem L poder gerar uma hipótese que minimiza o erro amostral. É útil a seguinte definição de *algoritmo minimizador do erro amostral*, ou simplesmente algoritmo MEA, para um conjunto de funções $\mathbb{H}$ (que se supõe terem valores alvo em $\{0,1\}$).

Definição 10 (Algoritmo de aprendizagem MEA) *Um algoritmo de aprendizagem L para uma classe de funções $\mathbb{H}$ diz-se* minimizador do erro amostral *(MEA) sempre que dado um inteiro m e uma amostra $\mathbf{Z} \in \mathbb{Z}^m$, a função $L : \bigcup_m^{\infty} \mathbb{Z}^m \to \mathbb{H}$ pela qual esse algoritmo é definido origine um erro amostral igual ao mínimo dos erros amostrais possíveis na classe $\mathbb{H}$:*

$$\widehat{er}_{\mathbf{Z}}(L(\mathbf{Z})) = \min_{h \in \mathbb{H}} \widehat{er}_{\mathbf{Z}}(h).$$

Portanto, um algoritmo MEA é, de entre todas as hipóteses h possíveis em $\mathbb{H}$, aquele que produz a hipótese que apresenta o menor número possível de desacordos em relação aos exemplos rotulados que esse algoritmo processou no treino. Uma das consequências do teorema da complexidade amostral I (9) é poder-se demonstrar que, sendo $\mathbb{H}$ um conjunto finito de funções com valores em $\{-1, 1\}$, qualquer algoritmo MEA para $\mathbb{H}$ é também um algoritmo de aprendizagem para $\mathbb{H}$.

2.1.5 Função de crescimento

Introduz-se, agora, um conjunto de conceitos que permitem generalizar o teorema da complexidade amostral I (9) a classes de funções $\mathbb{H}$ que não sejam finitas. Por exemplo, as classes de funções geradas numa simples máquina como um perceptrão nem sempre são conjuntos *finitos*. Outras redes neuronais mais complicadas podem também gerar representações funcionais cujo número de parâmetros pode pertencer a conjuntos que também não são finitos. Para a quase totalidade destas classes de funções a aprendizagem continua a ser possível, como se verá, desde que a sua complexidade não seja muito exagerada.

Vai começar-se por definir duas medidas que permitem exprimir o grau de complexidade de uma classe de funções $\mathbb{H}$, que estão relacionadas entre si. A primeira baseia-se na *função de crescimento*, e a outra é a *dimensão de Vapnik-Chervonenkis*, .

Seja um subconjunto finito $\mathbb{S}$ do espaço de entradas $\mathbb{X}$, e considere-se na classe $\mathbb{H}$ apenas as funções com domínio no subconjunto $\mathbb{S}$, denotada $\mathbb{H}_{|\mathbf{S}}$ e designada restrições de $\mathbb{H}$ ao conjunto $\mathbb{S}$. Se $\mathbb{H}_{|\mathbf{S}}$ for o conjunto de todas as funções que aplicam $\mathbb{S}$ em $\{-1, 1\}$, então, facilmente se vê que $\mathbb{H}_{|\mathbf{S}}$ tem uma capacidade suficiente para fazer a classificação de todos pontos de $\mathbb{S}$. Pode comparar-se o cardinal de $\mathbb{H}_{|\mathbb{S}}$ a $2^{|\mathbb{S}|}$, e interpretar-se o cardinal $|\mathbb{H}_{|\mathbb{S}}|$como uma medida da complexidade da classe de funções $\mathbb{H}$ com respeito ao conjunto $\mathbb{S}$.

Definição 11 (Função de crescimento) *A função de crescimento de uma classe de funções $\mathbb{H}$ é uma função de valor inteiro positivo, $\Pi_{\mathbb{H}} : \mathbb{N} \to \mathbb{N}$, dado pelo máximo cardinal do conjunto das funções $\mathbb{H}_{|\mathbb{S}}$, em que $|\mathbb{S}| = m$ elementos :*

$$\Pi_{\mathbb{H}}(m) = \max\left\{ |\mathbb{H}_{|\mathbb{S}}| \ \mid \ \mathbb{S} \subseteq \mathbb{X} \, e \, |\mathbb{S}| = m \right\}$$

Notar que $\Pi_{\mathbb{H}}(m) \leq 2^m$, para todo o m, e se o conjunto $\mathbb{H}$ for finito, então, vê-se facilmente que $\Pi_{\mathbb{H}}(m) \leq |\mathbb{H}|$ para todo o m, verificando-se a igualdade para valores de m suficientemente grandes. Portanto, pode considerar-se que o conceito de função de crescimento enriquece a noção trivial de cardinal de um conjunto, ou é uma extensão a conjuntos infinitos de funções, .

Pode, por exemplo, determinar-se a função de crescimento de um perceptrão. Pode dizer-se que uma dicotomia dada por uma classe de funções $\mathbb{H}$, e definida num subconjunto finito $\mathbb{S}$ do espaço de entradas, é uma das $\Pi_{\mathbb{H}}(|\mathbb{S}|)$ maneiras pelas quais o conjunto $\mathbb{S}$ pode ser classificado pelas funções de $\mathbb{H}$ em exemplos positivos e negativos. Uma dicotomia pode expremir-se por uma função $f : \mathbb{S} \rightarrow \{-1, 1\}$ pertencente à classe de funções $\mathbb{H}_{|\mathbb{S}}$. Pode contar-se o número de dicotomias de um subconjunto do espaço de entradas considerando-se no espaço de parâmetros uma partição num número de regiões, de tal modo que em cada região, todos os parâmetros correspondam à mesma classe. O número dessas regiões permite obter um limite superior do número de dicotomias, o qual foi estabelecido por Cover Cover (1965), e se exprime no teorema seguinte.

Teorema 12 (Cover, 1965) *Seja um perceptrão simples com n entradas reais, e cujos pesos nestas entradas são números reais. Seja $\mathbb{H}$ um conjunto de funções que o perceptrão é capaz de modelar, então, a função crescimento $\Pi_{\mathbb{H}}(m)$, para m exemplos (em posição geral) tem a expressão seguinte:*

$$\Pi_{\mathbb{H}}(m) = 2\sum_{k=0}^{n}\left(\begin{array}{c} m-1 \\ k \end{array}\right) = \\ = 2\sum_{k=0}^{n}\frac{(m-1)...(m-k)}{k!}$$

Uma interpretação deste número, que se pode obter em Bishop (1995) (p.86), designando por $F(m, n)$ a percentagem de dicotomias com m exemplos que são linearmente separáveis por um perceptrão com n entradas, então, tem-se que:

$$\Pi_{\mathbb{H}}(m) = 2^{m-1}F(m, n).$$

Exemplo 13 *Os limites superiores da função crescimento $\Pi_{\mathbb{H}}$ admitem a representação gráfica que mostra a figura 2.3, em que d_{VC} é um inteiro (possivelmente infinito) designado VC-dimensão de um conjunto de funções (dicotómicas) $\mathbb{H}$, e em que cada função, elemento de $\mathbb{H}$, tem como variáveis independentes os exemplos $\mathbf{z}_i$, $i = 1, ..., m$, e os parametros w.*

$$\Pi_{\mathbb{H}}(m) \leq \min\left(2^m, \left(\frac{(\exp(1) * m)}{d_{VC}}\right)^{d_{VC}}\right)$$

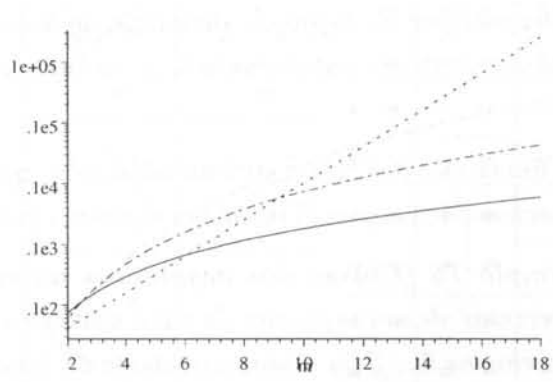


Figura 2.3: Representação gráfica dos limites superiores da Função Crescimento $\Pi_{\mathbb{H}}$: (a) recta, 2^m (b) curvas $\left(\frac{\exp(1)*m}{d_{VC}}\right)^{d_{VC}}$ para $d_{VC} = 2, 3, 4$ (4, curva superior)

2.1.6 Dimensão de Vapnik-Chervonenkis de uma classe de funções (d_{VC})

O conceito mais popular na teoria da aprendizagem estatística é sem dúvida o conceito de dimensão de Vapnik-Chervonenkis (d_{VC}). Cada uma das funções da classe $\mathbb{H}$separa os exemplos de uma certa maneira e, portanto, induz-lhes uma certa etiquetagem. Uma vez que as etiquetas ou são $+1$, ou -1, existem, no máximo, 2^m maneiras diferentes de colocar etiquetas em m exemplos. Uma classe $\mathbb{H}$ muito rica deveria ser capaz de realizar todas estas 2^m partições dizendo-se, neste caso, que essa classe de funções *estilhaça* os m exemplos. Contudo, certas classes de funções não são suficientemente ricas para realizar um *estilhaçamento* dos m pontos. Então, usa-se a dimensão VC de $\mathbb{H}$, $d_{VC}(\mathbb{H})$, que se define pelo máximo tamanho m do conjunto de exemplos que a classe de funções $\mathbb{H}$ *é capaz de estilhaçar*, e, no caso de m não existir, diz-se que $d_{VC}(\mathbb{H}) = \infty$. Por outras palavras, $d_{VC}(\mathbb{H})$ pode interpretar-se como uma espécie de resumo, expresso num único número, da capacidade de aprendizagem de uma máquina. Como se acaba de ver, a introdução deste conceito recorre à noção de "*estilhaçamento*" de um conjunto (*shattering*, em inglês; *pulverization*, em francês).

Diz-se que uma determinada classe de funções $\mathbb{H}$ é capaz de "*estilhaçar*" um conjunto $\mathbb{S}$ constituído por m exemplos, sempre que as funções de $\mathbb{H}$ permitam estabelecer todas as dicotomias possíveis de $\mathbb{S}$, isto é:

$$\left|\mathbb{H}_{|S}\right| = 2^m, \quad em\ que \quad |\cdot| \equiv card(\cdot).$$

Definição 14 (dimensão de Vapnik-Chervonenkis, d_{VC}) *Seja uma classe de funções $\mathbb{H}$, e um espaço de entradas $\mathbb{X}$ em que existe pelo menos um subconjunto*

$\mathbb{S}$ *que pode ser estilhaçado por* $\mathbb{H}$, *define-se* dimensão de Vapnik-Chervonenkis,d_{VC} , *como sendo o cardinal do maior subconjunto* $\mathbb{S}$,*e no caso de* $\mathbb{S}$ *não existir, diz-se que* $d_{VC} = \infty$.

Denota-se esta dimensão por d_{VC}, subentendendo-se que que é determinada relativamente a uma classe de funções $\mathbb{H}$ (também se chama por vezes *VC-dimensão).*

Definição 15 ***Exemplo 16 (Classe dos segmentos na recta numérica)*** *Considere-se o conceito de intervalos de um segmento da recta numérica* $[0,1]$, *o qual constitui o nosso espaço de exemplos* $\mathbb{Z}$. *Seja a seguinte classe de hipóteses definida por:*

$$\mathbb{H} = \{h_1, h_2, h_3, h_4\} = \{[0, r_1], [r_1, r_2], [r_2, r_3], [0,1]\}.$$

Considere-se $\mathbb{H}_{|S}$ *as restrições ao espaço de hipóteses, determinadas por um conjunto de dois pontos* $\mathbb{S} = \{z_1, z_2\}$. *Dado que:*

$$h_1 \cap \mathbb{S} = \emptyset, \; h_2 \cap \mathbb{S} = \{z_1\}, \quad h_3 \cap \mathbb{S} = \{z_2\}, \quad h_4 \cap \mathbb{S} = \mathbb{S},$$

isto é, para todos os sub-conjunto possíveis de $\mathbb{S}$ *existe uma hipótese que contém esses e só esses sub-conjuntos. Se* $S = \{z_1, z_2, z_3\}$, *com 3 pontos, nunca seria possível definir-se uma hipótese* h_i, *quer dizer um intervalo, que contivesse, por exemplo, os pontos* z_1 *e* z_3, *e não incluísse necessariamente* z_2. *Portanto, diz-se que existe um conjunto* $\mathbb{S}$ *que pode ser pulverizado pela classe de hipóteses* $\mathbb{H}_{|S}$. *Uma vez que o cardinal máximo de* $\mathbb{S}$ *é 2 conclui-se que o valor da* dimensão VC*:*

$$d_{VC} = 2.$$

Exemplo 17 (Funções afins) *A figura 2.4 mostra intuitivamente como se pode determinar que a dimensão VC é igual a 3 para a classe* $\mathbb{H}$ *de rectas (hiperplanos):*

$$h(z, w) = h(w.z + b)$$

(*i*) *Se um conjunto* $\mathbb{S}$ *de exemplos for constituído por apenas 3 pontos* $(m = 3)$,*em posição genérica, as rectas são capazes de o estilhaçar, uma vez que as etiquetagens são possíveis de obter usando-se uma recta a separar os pontos, logo* $\varphi_H(z_1, z_2, z_3, z) = 2^3$.

Exemplo 18 (*ii*) *Se um conjunto* $\mathbb{S}$ *de exemplos for constituído por 4 pontos* $(m = 4)$ *as duas etiquetagens que correspondem aos casos em que dois pontos estão numa qualquer das diagonais de um quadrilátero e pertencem à mesma classe, não são possíveis de obter com apenas uma recta, logo* $\varphi_H(z_1, z_2, z_3, z_4) = 14 < 2^4$.*Logo,*

$$d_{VC}(\mathbb{H}) = 3.$$

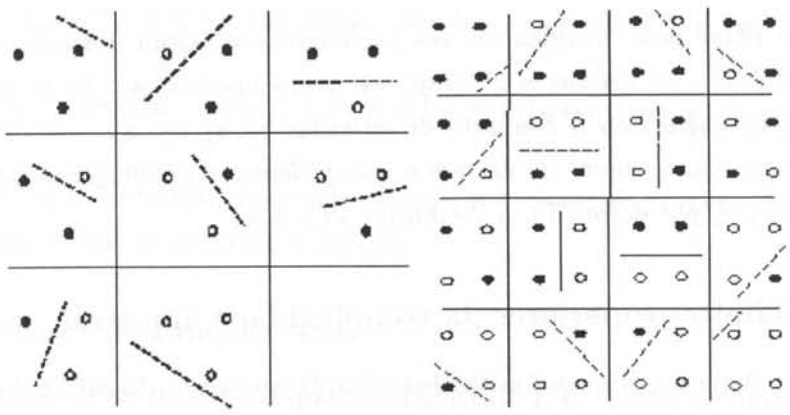


Figura 2.4: Dimensão VC de hiperplanos

Exemplo 19 (Perceptrão) *No caso de um perceptrão com n entradas para um conjunto de m exemplos sabe-se pelo teorema de Cover (12) que a função crescimento tem a seguinte expressão:*

$$\Pi_{\mathbb{H}}(m) = 2\sum_{k=0}^{n}\begin{pmatrix} m-1 \\ k \end{pmatrix} = \\ = \begin{cases} 2^m & se\ n \geq m-1 \\ 2^m - 2\sum_{k=n+1}^{m-1}\binom{m-1}{k} & caso\ contrario \end{cases}$$

Como esta é sempre menor que 2^m exactamente quando $m \geq n+2$, então,

$$d_{VC}(\mathbb{H}) = n+1.$$

Exemplo 20 *Seja uma família $h \in \mathbb{H}$ de funções, tal que $h : R^p \rightarrow \{\pm 1\}$, e seja uma restrição $\mathbb{H}_{|S}$ constituída pelo conjunto das funções de transferência duma rede neuronal no espaço dos parâmetros $\mathbb{W}$ realizável por uma certa máquina. Se esta for, por exemplo, um perceptrão de multicamada (PMC) com n unidades de entrada, p unidades escondidas e uma unidade de saída, a classe de funções tem $(n+1)p + p + 1 = c$ conexões, então, $\mathbb{H}_{|S}$ corresponde a uma escolha particular de pesos $\mathbf{w}$ no espaço dos parâmetros c-dimensional.*

Para cada conjunto de m exemplos, $\mathbf{z}_i$, toda a escolha de parâmetros $\mathbf{w} = \mathbf{w}_0$ especifica um vector de saída $h(\mathbf{z}, \mathbf{w})$ de tamanho m. Valores diferentes de $\mathbf{w}$ produzirão vectores resposta $h(\mathbf{z}, \mathbf{w})$ diferentes para um mesmo conjunto de exemplos. O número máximo de respostas possíveis de $h(\mathbf{z}, \mathbf{w})$ para m vectores de entrada é $2m$. Para $m = 3$ é evidente que se podem ter $2^3 = 8$ tipos de vectores resposta diferentes : $[-1,-1,-1]$, $[-1,-1,1]$ ... $[1,1,1]$. Se m aumentar, isto é, se crecer o número de exemplos, o número de variações das respostas possíveis de $h(\mathbf{z}, \mathbf{w})$ que

se podem obter pela modificação dos parâmetros $\bar{\mathbf{w}}$ *estão limitadas pela estrutura da rede PMC. Se existirem* 2^{μ} *respostas possíveis* $h(\mathbf{z}, \mathbf{w})$ *diz-se que a classe de funções* $\mathbb{H}_{|S}$ *pulverizou o conjunto de entradas* $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3 ... \mathbf{z}_{\mu}$ *. O máximo valor* μ*, isto é, o máximo número de exemplos que podem ser estilhaçados pelo conjunto de funções* $\mathbb{H}_{|S}$ *é, por definição, a dimensão VC,* d_{VC}*.*

2.1.7 Limites superiores da complexidade amostral

O teorema 9 que exprime os limites de $\varepsilon_L(m, \delta)$ em função do cardinal do do conjunto de funções $\mathbb{H}$ não é porém aplicável quando $\mathbb{H}$ tiver um número infinito de funções, mas, o teorema seguinte pode ser aplicado nesta situação.

Teorema 21 (Complexidade amostral II) *Seja uma classe de funções* $\mathbb{H}$ *que aplicam um conjunto* $\mathbb{X}$ *num contradomínio binário* $\{0, 1\}$*, e seja* d_{VC} *a dimensão de Vapnik Chernovenkis, finita, da classe de funções* $\mathbb{H}$*. Seja L um qualquer algoritmo MEA para* $\mathbb{H}$*. Então, L é um algoritmo de aprendizagem para* $\mathbb{H}$*, tal que se* $m \geq \frac{d_{VC}}{2}$*, verifica-se que o* erro de estimação *de L satisfaz :*

$$\varepsilon_L(m, \delta) \leq \varepsilon_0(m, \delta) = \left(\frac{32}{m}\left(d_{VC} \ln\left(\frac{2em}{d_{VC}}\right) + \ln\left(\frac{4}{\delta}\right)\right)\right)^{\frac{1}{2}}$$

e a complexidade amostral *de L satisfaz :*

$$m_L(\epsilon, \delta) \leq m_0(\epsilon, \delta) = \frac{64}{\varepsilon^2}\left(2.d_{VC} \ln\left(\frac{12}{\varepsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right)$$

O teorema da complexidade amostral *II* (21) exprime um resultado muito geral em que se estabelece que os extremos superiores do *erro de estimação* se podem determinar para todas as classes de funções $\mathbb{H}$ para as quais a *dimensão VC* é finita.

Os teoremas 9 e 21 podem demonstrar-se, Anthony and Bartlett (1999) (p.43), através da verificação de que os erros amostrais convergem uniformemente para os erros verdadeiros. A motivação para se utilizar um *algoritmo MEA* para a aprendizagem deve-se ao facto de que os erros amostrais sendo bons estimadores dos erros verdadeiros, então, ao escolher-se a hipótese com um erro mínimo está-se a escolher uma boa estratégia. O teorema seguinte, que é uma generalização do teorema 8, exprime que para um dado conjunto de treino aleatório suficientemente grande e toda a função $h \in \mathbb{H}$, se verifica com grande probabilidade que o *erro amostral* de h tem um valor bastante próximo do *erro verdadeiro* de h.

Teorema 22 (Convergência uniforme II) *Seja* $\mathbb{H}$ *um conjunto de funções com valores* $\{0, 1\}$ *definidas num conjunto* $\mathbb{X}$*, e seja P a função de distribuição sobre*$\mathbb{Z} =$

$\mathbb{X} \times \{0,1\}$. *Para* $0 < \varepsilon < 1$ *e um inteiro positivo* m, *verifica-se a que:*

$$P^m \left\{ \left| er_P(h) - \widehat{er}\,(h) \right| \geq \varepsilon\,,\ h \in \mathbb{H} \right\} \leq 4\Pi_{\mathbb{H}}\,(2m) \exp\left(-\frac{\varepsilon^2 m}{8} \right)$$

em que $\Pi_{\mathbb{H}}\,(2m)$ *é a* função de crescimento *para a classe de funções* $\mathbb{H}$ *e o conjunto de* m *exemplos (que se descreve a seguir).*

2.1.8 Classes de funções reais

Recorde-se que se afirmou atrás que um algoritmo L se diz ser de aprendizagem de uma classe de funções $\mathbb{H}$ sempre que, com um nível de confiança arbitrariamente grande, o erro amostral $\widehat{er}\,(L\,(\mathbf{Z}))$ for arbitrariamente pequeno.

Quando a classe de funções $\mathbb{H}$ tem um contradomínio binário $\{0,1\}$, e é finito o subconjunto $\mathbb{S} \in \mathbb{X}$, sendo $|\mathbb{S}| = m$, então, exprime-se a complexidade de $\mathbb{H}$ através da função de crescimento:

$$\Pi_{\mathbb{H}}\,(m) = \max \left| \left(\mathbb{H}_{|\mathbb{S}} \right) \right|,$$

em que $\mathbb{H}_{|\mathbb{S}}$ designa a restrição do domínio de $\mathbb{H}$ ao conjunto $\mathbb{S}$.

Quando a classe de funções $\mathbb{F}$ tem como contradomínio um conjunto de números reais, por exemplo, o intervalo $[0,1]$, a função de crescimento deixa de ter significado. Mesmo que o conjunto $\mathbb{S}$ seja finito, verifica-se em geral que qualquer sua restrição $\mathbb{F}_{|\mathbb{S}}$ é constituída por um número infinito de funções. Portanto, quando a classe $\mathbb{F}$ é constituída por funções de valores reais, o cardinal de $\mathbb{F}_{|\mathbb{S}}$ não permite saber se existe convergência uniforme entre os erros amostral e verdadeiro. Se o conjunto $\mathbb{F}_{|\mathbb{S}}$ tiver apenas alguns elementos que diferem muito entre si, então, em vez do cardinal do conjunto,pode passe a utilizar-se a noção de *cobertura* para medir a "*extensão*" (*extent*) de $\mathbb{F}_{|\mathbb{S}}$.

2.1.9 Número de cobertura de subconjuntos do espaço Euclidiano

Sejam um conjunto $\mathbb{W} \subseteq \mathbf{R}^k$ e um número real positivo ε, diz-se que $\mathbb{C} \subseteq \mathbf{R}^k$ é uma $d_\infty\varepsilon$-*cobertura* para $\mathbb{W}$, se $\mathbb{C} \subseteq \mathbb{W}$e, para todo $w \in \mathbb{W}$, existir um $\nu \in \mathbb{C}$, tal que:

$$\max_{i=1\ldots k} |w_i - \nu_i| < \varepsilon$$

Por outras palavras, dado um qualquer vector em $\mathbb{W}$ deve existir um vector em $\mathbb{C}$ cuja distância entre ambos seja menor que ε, sendo esta distância medida com a métrica d_∞ em $\mathbf{R}^k$, definida por:

$$d_\infty\left((x_1, x_2\,, \ldots x_k)\,, (y_1, y_2\,, \ldots y_k) \right) = \max_{i=1\ldots k} |w_i - v_i|$$

Para se verificar que $v \in \mathbf{R}^k$ e $\varepsilon > 0$ define-se em d_∞ uma esfera aberta no ponto ν e raio ε definida pelo conjunto de pontos:

$$\left\{u \in \mathbf{R}^k \mid d_\infty(u, v) < \epsilon\right\}.$$

Então, uma ε-cobertura para $\mathbb{W} \subseteq \mathbf{R}^k$ define-se como sendo um subconjunto $\mathbb{C}$ de $\mathbb{W}$ para o qual $\mathbb{W}$ está contido na união do conjunto das esferas abertas $d_\infty(u, v)$, de raio ε, centradas em todos os pontos de $\mathbb{C}$ (Figura 2.5).

Diz-se que $\mathbb{W}$ está totalmente limitado se, para cada $\varepsilon > 0$, existir uma *ε-cobertura* finita para $\mathbb{W}$ verificando-se, neste caso, que *o número da $d_\infty\varepsilon$-cobertura* de $\mathbb{W}$, $N(\varepsilon, \mathbb{W}, d_\infty)$, se pode definir pelo valor mínimo do cardinal da *$d_\infty\varepsilon$-cobertura* de $\mathbb{W}$.

2.1.10 Número de cobertura uniforme de uma classe de funções

Procede-se, agora, a uma generalização da função de crescimento introduzindo os números de cobertura, que se aplicam no estabelecimento de limites superiores do erro amostral. Suponha-se que $\mathbb{F}$ é uma classe de funções de $\mathbb{X}$, $(x_1, x_2, ..., x_k) \in \mathbb{X}^k$ é uma dada sucessão, e que $\mathbb{F}_{|\mathbb{X}}$ é um subconjunto de $\mathbf{R}^k$ dado por:

$$\mathbb{F}_{|\mathbb{X}} = \{f(x_1), f(x_2), ..., f(x_k) \mid f \in \mathbb{F}\}$$

Se forem distintos os termos da sucessão x_i, existe claramente uma bijecção entre $\mathbb{F}_{|\mathbb{X}}$ e a restrição $\mathbb{F}_{|\mathbb{S}}$ a $\mathbb{F}$, definida pelo conjunto $\mathbb{S} = (x_1, x_2, ..., x_k) \subseteq \mathbb{X}$. Para um número positivo ε define-se o *número de cobertura uniforme*, $N_\infty(\varepsilon, \mathbb{F}, k)$, como sendo, para todos as sucessões $(x_1, x_2, ..., x_k) \in \mathbb{X}^k$, o número cobertura máximo, $N(\varepsilon, \mathbb{F}_{|\mathbb{X}}, d_\infty)$, e se este máximo não existir considera-se que é infinito, isto é:

$$N_\infty(\varepsilon, \mathbb{F}, k) = \max\left\{N(\varepsilon, \mathbb{F}_{|\mathbb{X}}, d_\infty) \mid x \in \mathbb{X}^k\right\}$$

Um número de cobertura $N_\infty(\varepsilon, \mathbb{F}, k)$ pode interpretar-se como uma medida da riqueza da classe $\mathbb{F}$ à escala ε. Para um dado valor de ε este número indica em que medida o conjunto $\mathbb{F}_{|\mathbb{X}}$ "*preenche*" o espaço $\mathbf{R}^k$ quando este é examinado à escala k e, em particular, como esse número cresce com k.

2.1.11 Pseudo-dimensão e dimensão fat-estilhaçamento

A *VC-dimensão* sendo uma medida com um papel importante na avaliação da viabilidade da aprendizagem em redes com saídas binárias $\{-1, 1\}$ pode também ser usada como um limite da função de crescimento. Esta, por sua vez, pode ser generalizada usando os números de cobertura em redes de classificação de saídas reais.

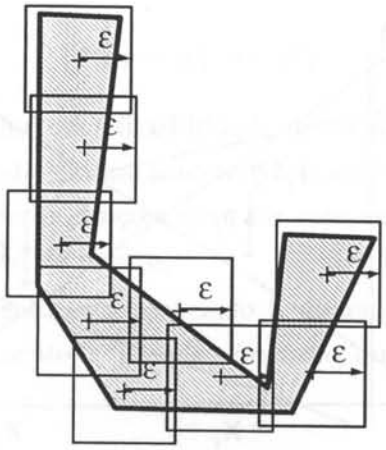


Figura 2.5: Os centros das "*bolas*"numa métrica d_∞ formam uma ε-cobertura da região a tracejado $\mathbb{F}_{|\mathbb{X}} \subset \mathbf{R}^2$

No caso que se foca a seguir importa encontrar uma outra medida '*combinatória*' análoga à *VC-dimensão* que sirva para limitar os números de cobertura de uma classe de funções reais e, por consequência, quantificar a complexidade amostral de um algoritmo de classificação. Tal dimensão designa-se por *pseudo-estilhaçamento* ('*pseudo-shattering*').

Primeiramente, vai-se proceder a uma reformulação da definição de $VC-$dimensão. Recorde-se que para um conjunto de funções:

$$f \in \mathbb{F}, \quad f : \mathbb{X} \to \{0,1\}$$

diz-se que um subconjunto de pontos:

$$\mathbb{S} = \{x_1, x_2, ..., x_m\} \subset \mathbb{X}$$

é estilhaçado por $\mathbb{F}$, se o cardinal do conjunto $\mathbb{F}_{|\mathbb{S}}$ for igual a 2^m. Isto significa que a um qualquer vector binário:

$$b = \{b_1.b_2, ..., b_m\} \in \{0,1\}^m$$

corresponderá uma função $f_b \in \mathbb{F}$ tal que:

$$(f_b(x_1), f_b(x_2), ..., f_b(x_m)) = \{b_1.b_2, ..., b_m\} = b$$

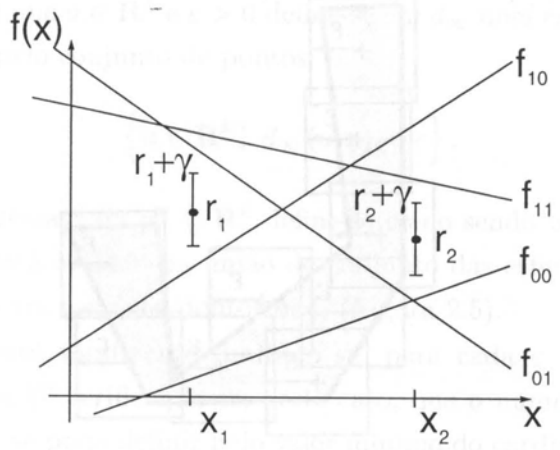


Figura 2.6: Pseudo-estilhaçamento

ou, por outras palavras, tal que:

$$f_b(x_i) = b_i, \text{ qualquer que seja } i.$$

Para a classe $\mathbb{F}$ de funções reais pode obter-se uma nova definição de estilhaçamento recordando que a função (de sinalização) $sgn(a)$ toma o valor 1 se $a \geq 0$, e o valor 0 nos outros casos alternativos.

Definição 23 (*pseudo-estilhaçamento*) *Seja $\mathbb{F}$ um conjunto de funções reais, $f : \mathbb{X} \to \mathbf{R}$, e seja $\mathbb{S} = \{x_1, x_2, ..., x_m\}$ um subconjunto de $\mathbb{X}$, então, diz-se que $\mathbb{S}$ é* pseudo-estilhaçado *por $\mathbb{F}$ se existirem m números reais $r_1, r_2, ..., r_m$, tal que para cada $b \in \{1, 0\}^m$ existe a função $f_b \in \mathbb{F}$, tal que:*

$$sgn(f(x_i) - r_i) = b_i \quad para \quad 1 \leq i \leq m.$$

*Diz-se que $r = r_1, r_2, ..., r_m$ é uma "*testemunha*" do estilhaçamento.*

Portanto, um conjunto $\mathbb{S}$ é *pseudo-estilhaçado* se existirem os números reais $r_1, r_2, ..., r_m$, e as 2^m funções f_b que realizam todas as possíveis combinações, "*para cima e para baixo*", relativamente aos números r_i, os quais constituem as suas testemunhas (ver figura 2.6).

Notar que o conjunto $\{x_1, x_2\} \subseteq \mathbf{R}$ é pseudo-estilhaçado pela classe de funções afins sobre $\mathbf{R}$:

$$\{f(x) \to ax + b \mid \ a, b \in R\}$$

e os pontos r_1 e r_2 são as '*testemunhas*' desse estilhaçamento, mas não é estilhaçado

pela classe de funções

$$\{f(x) \to a \mid \ a \in R\}.$$

Esta classe de funções não tem a flexibilidade suficiente para aproximar os pontos $\{x_1, x_2\}$.O conceito de pseudo-estilhaçamento relaciona-se com a expressividade, ou flexibilidade, que uma classe de funções deve ter com vista a poder constituir uma aproximação de um conjunto de pontos.

Tal como o estilhaçamento está associado à dimensãoVC, também, agora, se associa o pseudo-estilhaçamento à '*pseudo-dimensão,*'que se denota $Pdim(\mathbb{F})$, e se define a seguir.

Definição 24 (*pseudo-dimensão*) *Seja $\mathbb{F}$ um conjunto de funções $f : \mathbb{X} \to \mathbf{R}$, então, diz-se que d é a pseudo-dimensão de $\mathbb{F}$ se d for o* máximo *cardinal de um subconjunto $\mathbb{S}$ de $\mathbb{X}$ que seja pseudo-estilhaçado por $\mathbb{F}$. Se esse máximo não existir diz-se que $\mathbb{F}$ tem pseudo-dimensão infinita. .*

Exemplo 25 (Funções indicatrizes: $P\dim = d_{VC}$) *Seja o caso duma função indicatriz da região por cima, ou por baixo, B_f , em que $f \in \mathbb{F}$, sendo dada por:*

$$B_f(x, y) = sgn(f(x) - y).$$

Verifica-se, então, que a pseudo-dimensão de $\mathbb{F}$ é igual à $d_{VC}(\mathbb{B}_F)$ em que a classe do sub-gráfico é dada por:

$$\mathbb{B}_F = \{B_f \mid f \in \mathbb{F}\}.$$

Relembre-se que uma classe $\mathbb{F}$ de funções reais é um espaço vectorial se para todas as funções $f, g \in \mathbb{F}$, e escalares $\lambda, \mu \in \mathbf{R}$, se verifica que a combinação linear $\lambda f + \mu g$ também pertence a $\mathbb{F}$. Se $\mathbb{F}$ for um espaço vectorial de funções reais, então,

$$Pdim(\mathbb{F}) = \dim(\mathbb{F}).$$

Uma vez que para a classe $\mathbb{B}_F$ se sabe que:

$$Pdim(\mathbb{F}) = d_{VC}(\mathbb{B}_F),$$

logo, por definição de classe de funções $\mathbb{B}_F$, acima indicada, tem-se que as funções desta classe são da forma:

$$sgn(f + g)$$

em que f é uma função do espaço vectorial e g é a função constante $g(x, y) = -y$. Dado o facto, que se pode demonstrar, de que se $\mathbb{F}$ for um espaço vectorial de funções,

e g uma função real, então:

$$\mathbb{H} = \{sgn\,(f+g) \mid f \in \mathbb{F}\}$$

verifica-se que $d_{VC}\,(\mathbb{H}) = \dim(\mathbb{F})$ *Anthony and Bartlett (1999) (p.37).*

Logo, fica demonstrado que:

$$Pdim\,(\mathbb{F}) = \dim(\mathbb{F})\,.$$

Recorde-se que um conjunto $S = \{x_1, x_2, ..., x_m\}$ é estilhaçado pela classe de funções $\mathbb{F}$ se existirem os números $r_1, r_2, ..., r_m$ tal que todas as 2^m combinações, por cima e por baixo, com respeito aos r_i sejam possíveis. Pode exprimir-se esta condição mais rigorosamente requerendo que os valores da função em x_i não só estejam por cima e por baixo dos r_i, mas também dentro de uma certa margem γ. Formaliza-se esta ideia com a seguinte definição de conjunto γ- *estilhaçado.*

Definição 26 (**γ- *estilhaçamento***) *Seja* $\mathbb{F}$ *um conjunto de funções reais,* $f : \mathbb{X} \to \mathbf{R}$. *Seja um conjunto* $\mathbb{S}$ *tal que* $\mathbb{S} = \{x_1, x_2, ..., x_m\} \subset \mathbb{X}$ *e um número real positivo* γ. *Então, diz-se que* $\mathbb{S}$ *é* γ- estilhaçado *por* $\mathbb{F}$ *se para qualquer* $b \in \{0,1\}^m$ *e* $1 \leq i \leq m$ *existirem os números reais* $r_1, r_2, ..., r_m$, *e uma função* f_b *em* $\mathbb{F}$, *que satisfaz as desigualdades:*

$$f_b\,(x_i) \geq r_i + \gamma \quad se\ b_i = 1 \quad e \quad f_b\,(x_i) \leq r_i - \gamma \quad se\ b_i = 0\,.$$

Diz-se que $r = r_1, r_2, ..., r_m$ *é uma* 'testemunha' *do estilhaçamento.*

Portanto, $\mathbb{S}$ é $\gamma-$ *estilhaçado* se for estilhaçado com uma largura de estilhaçamento igual ou superior a γ. Esta noção conduz naturalmente à seguinte definição de γ-*dimensão.*

Definição 27 (**γ-*dimensão ou dimensão de fat-estilhaçamento***) *Seja* $\mathbb{F}$ *um conjunto de funções* $f : \mathbb{X} \to \mathbf{R}$, *e* $\gamma > 0$, *então, diz-se que* d *é a* γ-*dimensão de* $\mathbb{F}$ *se* d *for o máximo cardinal do subconjunto* $\mathbb{S}$ *de* $\mathbb{X}$ *que é* $\gamma-$*estilhaçado por* $\mathbb{F}$. *Se este máximo não existir diz-se, então, que* $\mathbb{F}$ *tem uma* $\gamma-$*dimensão infinita. Representa-se a* γ-*dimensão de* F *por* $fat_F(\gamma)$. *Define-se assim uma função* $fat_F : R^+ \to \mathbb{N} \cup \{0, \infty\}$, *designada dimensão de* $fat-$*estilhaçamento. Diz-se que* $\mathbb{F}$ *tem uma dimensão de* $fat-$*estilhaçamento finita sempre que for ocaso de* $\gamma > 0$ *e, por consequência,* $fat_F(\gamma)$ *é finita.*

Uma vez que esta dimensão depende da escala γ pode descrever-se como uma dimensão sensível à escala. Como ilustração do conceito de *dimensão fat-estilhaçamento* vamos considerar seguidamente classes de funções de variação limitada.

Exemplo 28 (Função de de variação limitada) *Diz-se que a função* $f : [0,1] \rightarrow \mathbf{R}$ *é de variação limitada se existir um número real* V *tal que para todo o inteiro* n, *e toda a sucessão crescente de números positivos* $y_1, y_2, ..., y_n$ *se tem*

$$\sum_{i=1}^{n-1} |f(y_{i+1}) - f(y_i)| \leq V$$

Neste caso diz-se que f *tem uma variação total limitada a menos de* V. *Pode mostrar-se que se tivermos uma classe de funções* $\mathbb{F}$ *que aplicam o intervalo* $[0,1]$ *no intervalo* $[0,1]$, *e tenham uma variação total limitada de pelo menos* V *que se verifica a seguinte relação* $fat_F(\gamma) = 1 + \left\lfloor \frac{V}{2\gamma} \right\rfloor$ *em que* $\lfloor \cdot \rfloor$ *é a função menor inteiro (*floor*).*

Existe um teorema devido a Simon (1997), e referido em Anthony and Bartlett (1999) (p.161) e Marques_de_Sá (2001a) (p.198), que permite determinar uma estimativa da *dimensão fat-estilhaçamento* de funções com domínio e contradomínio $[0,1]$:

$$fat_F(\gamma) = 1 + trunc\left(\frac{V}{2\gamma}\right)$$

em que a função $trunc(x)$ dá o maior inteiro menor ou igual a x e V é a variação total da função:

$$V = \sum_{i=1}^{m-1} |f(x_{i+1}) - f(x_i)| .$$

2.2 Aprendizagem de Funções Reais

A aprendizagem de funções reais é um dos problemas a resolver para permitir a estimação do peso à nascença de uma criança. Envolve um estudo de famílias de funções reais de várias variáveis reais, $\mathbb{F}$, e que se descreve nos capítulos quarto e seguintes. As variáveis independentes destas funções são vectores cujas componentes são algumas medidas biométricas do feto, obtidas por ecografia, assim como o peso à nascença.

Vai, seguidamente, introduzir-se uma formalização do conceito de variável aleatória, inspirada em Murteira et al. (2001), Billingsley (1995) ou Spanos (1999), para permitir utilizar alguns resultados, a serem referidos ao longo dos próximos capítulos, quando se requer um rigor mais apertado na especificação dos espaços probabilizados. Considere-se um vector exemplo em que se admite que as componentes x_i, $i = 1, ..., n$, de cada vector-exemplo (podem ser medidas ecográficas e x_n pode ser o peso fetal) se interpretam como sendo resultantes de experiências aleatórias em espaços probabilizáveis do tipo $(\Omega, \mathcal{A})$. Um espaço probabilizável do tipo $(\Omega, \mathcal{A})$, quer dizer que é constituído por um dado conjunto de acontecimentos elementares

Ω, e uma dada *σ-álgebra*, ou *tribo*, sobre Ω, designada por $\mathcal{A}$. Seja P uma lei de probabilidade definida sobre $(\Omega, \mathcal{A})$ e suponha-se que se define para cada componente de um vector-exemplo uma variável aleatória real X_i, sobre $(\mathbf{R}, \mathcal{A})$. Portanto, X_i é uma função mensurável de $(\Omega, \mathcal{A})$ sobre um espaço probabilizável $(\mathbf{R}, \mathcal{B})$, constituído pelo conjunto dos números reais, $\mathbf{R}$, e por uma $\sigma-$álgebra de Borel $\mathcal{B}$ definida em $\mathbf{R}$. Ao $n-$uplo de variáveis aleatórias $\mathcal{X} = (X_1, X_2, ...X_n)$, cada uma definida sobre um espaço de probabilidade $(\mathbf{R}, \mathcal{A}, P)$,dá-se a designação de *vector aleatório real.* Uma *lei de probabilidade* para o vector aleatório real $\mathcal{X}$, pode ser definida como sendo a medida de probabilidade sobre $(\mathbf{R}^n, \mathcal{B}^n)$, e expressa pela função:

$$P_{\mathcal{X}} : \mathcal{B}^n \to [0, 1],$$

definida para todo o boreliano B de $\mathcal{B}^n$ pela relação:

$$P_{\mathcal{X}}(B) = P\left(X^{-1}(B)\right).$$

Vai-se referir, agora, a alguns aspectos do problema da regressão comparativamente com o problema da classificação que foi revisto na subsecção 2.1.2 e seguintes. O problema da classificação de padrões foi estudado supondo que os exemplos $(\mathbf{x}, y)$ eram vectores $n-$dimensionais constituídos por um vector de entradas $\mathbf{x}$ e outro de etiquetas y unidimensionais, ambos gerados por uma lei de probabilidade definida num espaço $\mathbb{Z}_0 = \mathbf{R}^n \times \{-1, 1\}$. Presentemente procura-se generalizar a um espaço $\mathbb{Z} = \mathbf{R}^n \times \mathbf{R}$. Uma outra modificação, que importa realçar desde já, tem a ver com o facto de na classificação interessar escolher uma função, pertencente a uma dada classe de funções, tal que seja mínimo o número de classificações erradas. Isto é, tal que a $\Pr\{f(\mathbf{x}) \neq y\}$ seja tão pequena quanto possível. Na aprendizagem de funções reais interessa estimar até que ponto uma dada função $f(\mathbf{x})$ se afasta do valor verdadeiro y da respectiva etiqueta, afastamento este que se pode medir por uma função de erro apropriada, a mais utilizada sendo a função *quadrado das diferenças,* $(f(\mathbf{x})-y)^2$. A melhor função f_0 que aproxima um dado conjunto de exemplos, $\{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, m\}$, é a que minimiza o erro quadrático médio, $E\left\{(f(\mathbf{x}) - y)^2\right\}$.

Finalmente, vai-se formalizar os conceitos de *algoritmo de aprendizagem* para funções reais e *complexidade amostral.* Considera-se um funcional de custo, $c_f = (f(\mathbf{x}) - y)^2$, dependente da função f, e admite-se que um conjunto de pares $(\mathbf{x}, y)$, em que $\mathbf{x} \in \mathbb{X} \subset \mathbf{R}^n$, e $y \in \mathbf{R}$, são gerados por uma lei de probabilidade P. Pode escrever-se a seguinte expressão funcional:

$$er_P(f) = E\left\{(f(\mathbf{x}) - y)^2\right\} = E\{c_f\}$$

Um algoritmo de aprendizagem, neste contexto, recebe uma matriz $m \times (n + 1)$

de exemplos:

$$\mathbf{Z} \in (\mathbb{X} \times \mathbf{R})^m$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{x}_{11} & \cdots & \mathbf{x}_{1n} & y_1 \\ \mathbf{x}_{21} & \cdots & \mathbf{x}_{2n} & y_2 \\ \vdots & \cdots & \vdots & \vdots \\ \mathbf{x}_{m1} & \cdots & \mathbf{x}_{m,n} & y_m \end{bmatrix}$$

e determina uma função $f \in \mathbb{F}$.

Definição 29 (Algoritmo de aprendizagem II) *Seja $\mathbb{F}$ um conjunto de funções de $\mathbb{X}$ num intervalo real $[0,1]$. Uma função L tal que:*

$$L = \bigcup_{m=1}^{\infty} (\mathbb{X} \times \mathbf{R})^m \to \mathbb{F}$$

diz-se que é um algoritmo de aprendizagem *para $\mathbb{F}$, se tiver as propriedades seguintes.*

Para quaisquer $\varepsilon \in [0,1]$, $\delta \in [0,1]$, então, existe um inteiro $m_0(\varepsilon, \delta, B)$ tal que, se $m \geq m_0(\varepsilon, \delta)$, e para a distribuição de probabilidade P sobre $X \times [0,1]$, se $\mathbf{Z}$ é um conjunto de treino com m exemplos retirados aleatoriamente de acordo com a lei de probabilidade P, então, com uma probabilidade no mínimo de $1 - \delta$ pode dizer-se que a função f dada por $L(\mathbf{Z})$ satisfaz a desigualdade:

$$er_p(f) < opt_P(\mathbb{F}) + \varepsilon$$

em que:

$$opt_P(f) = \inf_{f \in \mathbb{F}} er_P(f) .$$

Por outras palavras, para $m \geq m_0(\varepsilon, \delta, B)$ a função $f(\cdot) = L(\mathbf{Z})$ satisfaz a desigualdade:

$$P^m \{er_P[f(x_i)] < opt_P[f(x_i)] + \varepsilon\} \geq 1 - \delta$$

Sempre que existir um algoritmo de aprendizagem para $\mathbb{F}$ diz-se que a classe de funções $\mathbb{F}$ pode ser aprendida.

Definição 30 (Complexidade amostral) *A complexidade amostral, $m_L(\varepsilon, \delta)$ de um algoritmo de aprendizagem para $\mathbb{F}$ define-se como sendo o mínimo valor que pode ter o número inteiro $m_0(\varepsilon, \delta)$, quaisquer que sejam ε, δ.*

Seja L um algoritmo de aprendizagem, então, é necessário e suficiente que exista uma função $\varepsilon_L(m, \delta)$ tal que para todo $\delta \in [0,1]$, se verifica que $\varepsilon_L(m, \delta)$ converge

para zero quando $m \to \infty$, e tal que para todo m, δ e P com probabilidade no mínimo de $1 - \delta$, para qualquer f dado por $L(\mathbf{Z})$ se verifica:

$$er_P[f] < opt_P[\mathbb{F}] + \varepsilon_L(m, \delta) .$$

Neste caso ao ínfimo de $\varepsilon_L(m, \delta)$ chama-se *erro de estimação.*

O erro empírico ou amostral duma função para uma dada amostra de treino é um indicador útil do seu erro. Dada a função $f \in \mathbb{F}$ e a amostra $\mathbf{Z}$, o erro amostral é dado por:

$$\widehat{er}_{\mathbf{Z}}[f] = \frac{1}{m} \sum_{i=1}^{m} (f(\mathbf{x}_i) - y_i)^2$$

Um algoritmo minimizador do erro amostral (MEA) para uma classe de funções finita $\mathbb{F}$ é uma função L com a propriedade de que

$$\widehat{er}_{\mathbf{Z}}[L(\mathbf{Z})] = \min_{f \in \mathbb{F}} \widehat{er}_{\mathbf{Z}}[f]$$

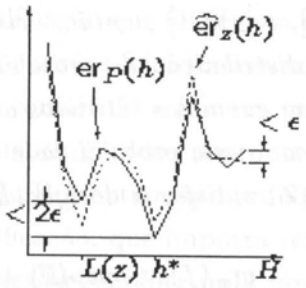


Figura 2.7: Minimização do erro amostral $\widehat{er}_{\mathbf{z}}(h)$ equivale a aproximar o erro verdadeiro $er_P(h)$ (Anthony M. and Barlett P. L.,1999,p.20)

Teorema 31 *Seja $\mathbb{F}$ uma classe finita de funções de um conjunto $\mathbb{X}$ em$[0,1]$ e suponhamos que L é um algoritmo MEA para $\mathbb{F}$. Então L é um algoritmo de aprendizagem para f cujo erro de estimação admite como majorante:*

$$\varepsilon_L(m, \delta) \leq \varepsilon_0(m, \delta) = \left(\frac{2}{m} \ln \left(\frac{2|\mathbb{F}|}{\delta} \right) \right)^{1/2}$$

Os dois teoremas seguintes que são enunciados e demonstrados em Cucker and Smale (2001), pressupõem a existência de $\mathbb{C}(\mathbb{X})$, um espaço de Banach de funções contínuas em $\mathbb{X}$, com a norma $\|f\|_\infty = \sup_{x \in \mathbb{X}} |f(x)|$. Seja $\mathbb{F}$ um subconjunto de $\mathbb{C}(\mathbb{X})$ o *espaço de hipóteses* no qual os algoritmos de aprendizagem vão tentar encontrar as funções f que melhor aproximam a função óptima. Seja $\sigma^2 = \sigma^2(\mathbb{F}) = \sup_{f \in \mathbb{F}} \sigma^2(f_{\mathbb{Y}}^2)$ em que $\sigma^2(f_{\mathbb{Y}}^2)$ é a variância de $f_{\mathbb{Y}}^2$, f é uma função desconhecida

$f : \mathbb{X} \to \mathbb{Y}$ e existe uma medida de probabilidade que permite retirar exemplos aleatoriamente e i.i.d., em $\mathbb{X} \times \mathbb{Y}$.

Teorema 32 (Convergência uniforme III) *Seja um conjunto compacto de* $\mathbb{C}(\mathbb{X})$. *Suponha-se que , para todo* $f \in \mathbb{F}$ *tem-se que* $|f(x) - y| \leq M$ *quase sempre. Então, para todo o* $\varepsilon > 0$ *verifica-se:*

$$P^m \left\{ \sup_{f \in \mathbb{F}} |er_P(f) - \widehat{er_Z}(f)| \leq \varepsilon \right\} \geq 1 - N\left(\mathbb{F}, \frac{\varepsilon}{8M}\right) 2\exp\left(-\frac{m\varepsilon^2}{4\left(2\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}\right)$$

Teorema 33 (Estimativa do erro amostral) *Seja* $\mathbb{F}$ *um conjunto compacto de* $\mathbb{C}(\mathbb{X})$, *suponha-se que para toda* $f \in \mathbb{F}$ *tem-se que* $|f(x) - y| \leq M$ *quase sempre, então para todo o* $\varepsilon > 0$ *verifica-se:*

$$\Pr\left(\widehat{er_Z}(f) \leq \varepsilon\right) \geq 1 - N\left(\mathbb{F}, \frac{\varepsilon}{16M}\right) 2\exp\left(-\frac{m\varepsilon^2}{8\left(4\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}\right)$$

Este teorema ajuda a responder à questão de se saber quantos exemplos vão ser necessários para que, com uma confiança maior que , se verifique que $\widehat{er_Z}(f) < \varepsilon$, isto é, dados $\varepsilon, \delta > 0$, para assegurar que:

$$\Pr\left(\widehat{er_Z}(f) \leq \varepsilon\right) \geq 1 - \delta$$

é suficiente que o número de exemplos m satisfaça a condição:

$$m \geq \frac{8\left(4\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}{\varepsilon^2}\left[\ln\left(2N\left(\mathbb{F}, \frac{\varepsilon}{16M}\right)\right) + \ln\left(\frac{1}{\delta}\right)\right].$$

2.2.1 Função de custo para funções reais (*regressão*)

Um dos requisitos mais imediatos em qualquer problema de aprendizagem consiste em especificar convenientemente o que se pretende de facto conseguir, seja minimizar, limitar ou aproximar uma dada quantidade. Para este efeito é imperativo que se defina um critério em conformidade com o qual seja sempre possível avaliar-se a qualidade das estimativas dadas por uma dada função estimadora $f : \mathbb{X} \to \mathbb{Y}$, obtida com base em exemplos.

Vamos começar por definir o que se entende por um *custo* incorrido por uma função f num ponto $\mathbf{x}$ ao qual está associada uma etiqueta y.

Definição 34 (Função de Custo) *Se designarmos por* $(\mathbf{x}, y, f(\mathbf{x})) \in \mathbb{X} \times \mathbb{Y} \times \mathbb{Y}$ *o terno composto por um exemplo* $\mathbf{x}$, *uma etiqueta* y *e uma estimativa* $f(\mathbf{x})$, *então, chama-se* função de custo *à aplicação* $c : \mathbb{X} \times \mathbb{Y} \times \mathbb{Y} \to [0, \infty)$ *com a propriedade de que* $c(\mathbf{x}, y, y) = 0$ *para todo o* $\mathbf{x} \in \mathbb{X}$ *e* $y \in \mathbb{Y}$.

Note-se que se requer que c seja uma função não-negativa e que todas as estimativas para as quais $f(\mathbf{x}) = y$ não dêem origem a qualquer custo, pelo que se sabe que o mínimo do custo é sempre 0.

Quando se fazem estimativas de funções de valores reais é usual avaliar-se a diferença $f(\mathbf{x}) - y$, isto é o grau de previsão errada, pelo que em muitos casos se usa uma função de custo do tipo:

$$c(\mathbf{x}, y, f(\mathbf{x})) = \tilde{c}(f(\mathbf{x}) - y) \tag{2.2}$$

em que a função $\tilde{c}(\mathbf{x})$ dá conta da dependência das entradas. A escolha mais popular, chamada *função quadrática dos custos* e que consiste em minimizar-se a soma dos quadrados dos resíduos $f(\mathbf{x}) - y$, corresponde à suposição de que existe um ruído gaussiano aditivo a corromper as observações y_i. Consequentemente minimiza-se

$$c(\mathbf{x}, y, f(\mathbf{x})) = (f(\mathbf{x}) - y)^2 \quad \text{ou, o que é equivalente} \quad \tilde{c}(\xi) = \xi^2 \tag{2.3}$$

e, por conveniência de notação no que vai seguir-se usa-se frequentemente $\frac{1}{2}\xi^2$ em vez de ξ^2.

Utiliza-se em geral na aprendizagem *kernel* em problemas de regressão a chamada função de custo ε-insensitiva:

$$\tilde{c}(\xi) = \max\left[(|\xi| - \varepsilon), 0\right] \tag{2.4}$$

para a qual os desvios inferiores a uma tolerância ε não são penalizados, e os restantes sofrem uma penalização linear. Para que a implementação de um procedimento de aprendizagem seja eficiente é necessário que a função de custo satisfaça algumas propriedades, em particular, não ser computacionalmente muito difícil de calcular e que tenha um número de descontinuidades pequenas ou mesmo nenhum na primeira derivada.

2.2.2 Funcional de risco para funções reais *(regressão)*

Uma vez que se definiu como se calcula o erro pode agora abordar-se a questão de saber como se deve penalilizá-lo em casos específicos $(\mathbf{x}, y, f(\mathbf{x}))$, isto é, definir um método que permita combinar várias penalizações localizadas.

Quando se está a tentar aprender algo acerca da regularidade ou dos parâmetros de uma dada distribuição de probabilidade de onde os exemplos são extraídos, de facto, aquilo que se faz é realizar m experiências m que se admite (para simplificar) serem aleatórias e em condições idênticas. Chama-se a isto extrair exemplos (ou amostras) de P independente e identicamente distribuídos *(iid)*. Formalmente pode exprimir-se a situação em que se fazem extracções de exemplos *iid* através do espaço

de probabilidades $(\mathbb{X}^m, C^m, P^m)$, no qual $\mathbb{X}^m$ designa um produto cartesiano de m factores todos iguais a $\mathbb{X}$, C^mdesigna a mais pequena σ-álgebra que contém os elementos do produto cartesiano de m factores todos iguais a C. A medida produto P^m é determinada unicamente pela seguinte expressão que codifica o conceito de independência e de identidade da medida:

$$P^m(C_1, \ldots, C_m) = \prod_{i=1}^{m} P(C_i)$$

Como já foi dito repetidas vezes, uma das principais preocupações da teoria estatística da aprendizagem consiste em considerar sempre os factores responsáveis por uma boa generalização. Nesta perspectiva um dos problemas a resolver, perante um dado um conjunto de treino:

$$\mathbb{Z} = \{(x_i, y_i) \in \mathbb{X} \times \mathbb{Y},\ i = 1 \ldots m\},$$

consiste em obter-se uma função estimadora $f : \mathbb{X} \to \mathbb{Y}$ que permita a melhor generalização possível. Isto é, determinar-se uma função que possa ser usada para prever o valor de y que corresponde a qualquer novo caso x, suposto que este foi extraído em conformidade com a mesma lei de distribuição de probabilidades P dos exemplos de treino.

Suponha-se que existe uma distribuição de probabilidade $P(x, y)$ em $\mathbb{X} \times \mathbb{Y}$ que governa a geração dos dados e a dependência funcional dos dados subjacente. Designe-se a distribuição condicional de y dado x por $P(y|x)$. Suponha-se que existem as diferenciais de cada uma destas duas funções no domínio que interessa, que se representam por $dP(x, y)$ e $dP(y|x)$, respectivamente. Suponha-se que não existe qualquer conhecimento sobre o conjunto de teste, então, tem-se a seguinte definição:

Definição 35 (Risco Esperado) *É o mínimo erro esperado* $\mathbf{E}[c(x, y, f(x))]$ *para todos os m exemplos de treino, e que pode ser obtido pela minimização da seguinte funcional com respeito a P e a c:*

$$R[f] = \int_{\mathbb{X} \times \mathbb{Y}} c(x, y, f(x))\, dP(x, y) \tag{2.5}$$

sempre que $P(x, y)$ *for conhecido, o que não é obviamente aplicável a uma aprendizagem baseada em exemplos. Se existir (não é necessário) a função densidade de probabilidade* $p(x, y)$, *então, pode fazer-se a substituição* $dP(x, y) = p(x, y)\, dxdy$.

Infelizmente $p(x, y)$ também não é conhecida, a única coisa que se tem são os

dados pelo que habitualmente substitui-se esta função pela *densidade empírica*:

$$p_{emp} = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}(x)\, \delta_{y_i}(y)$$

em que $\delta_{x'}(x)$ designa a distribuição δ de Dirac que satisfaz $\int \delta_{x'}(x) f(x)\, dx = f(x')$. Admite-se que substituindo p por p_{emp} se obtém um valor suficientemente aproximado para o erro esperado, pelo que substituindo esta expressão se obtém a seguinte definição.

Definição 36 (Risco Empírico) *Define-se risco empírico*

$$R_{emp}[f] = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i)) \tag{2.6}$$

o qual tem a vantagem de poder ser calculado e minimizado a partir dos conjunto de exemplos dado.

A determinação da função f de uma dada classe $\mathbb{F}$ que minimiza o risco $R_{emp}[f]$ com respeito a $\mathbb{F}$ é um problema do tipo mal-formulado, como se mostra no seguinte exemplo.

Exemplo 37 *Vamos supor que se pretende resolver um problema de* regressão *usando a função de custo quadrática, e em que se supõe que a classe de funções $f \in F$ é linear, $f(x) = \sum_{i=1}^{p} \alpha_i g_i(x)$, em que α_i são números reais e $g_i(x)$ são funções que aplicam $\mathbb{X}$ em $\mathbf{R}$. Pretende-se minimizar o risco empírico:*

$$\min_{f \in F} R_{emp}[f] = \min_{\alpha \in R} \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{i=1}^{p} \alpha_i g_i(x) \right)^2 \tag{2.7}$$

Calculando a derivada do $R_{emp}[f]$ com respeito a α, e definindo G_{ij} como sendo $g_i(x_j)$, pode ver-se que o mínimo ocorre em:

$$G^T \mathbf{y} = G^T G\, \alpha.$$

Uma condição suficiente para se verificar esta igualdade é que seja $\alpha = \left(G^T G\right)^{-1} G^T \mathbf{y}$, em que $\left(G^T G\right)^{-1}$ designa a matriz pseudo-inversa. Se esta tiver uma característica de má condição, isto é, se a razão entre o maior e o menor dos valores próprios de $G^T G$ for elevada, é difícil resolver o problema para α. Se adicionalmente $p > m$, quer dizer se o número de funções de base g_i for maior que o número de exemplos, existe um subespaço de soluções com dimensão de pelo menos $p - m$ que satisfaz (2.7) o qual é indesejável quer teórica quer computacionalmente.

Pode também esperar-se que se a classe $\mathbb{F}$ for muito rica a discrepância entre os riscos $R_{emp}[f]$ e $R[f]$ pode tornar-se bastante grande, pelo que este problema resolve-se pela adição a $R_{emp}[f]$ de um termo chamado de regularização, que se descreve mais detalhadamente no Capítulo três.

Em certas situações pode estar-se interessado em determinar $P(y|x)$,a distribuição condicional de y dado um exemplo x, isto é , desejar-se conhecer $\mathbf{E}[c(x, y, f(x))]$ para determinar qual é o risco que se pode esperar se ocorrer o exemplo x. Em estatística Bayesiana é frequente procurar-se obter estimativas de densidade correspondentes a variáveis aleatórias (x, y) , como por exemplo em casos de diagnóstico médico em que pode mesmo ser necessária ter-se informação sobre $p(x, y)$ para se poder tirar alguma conclusão de um dado conjunto de dados de treino. Se se começar por definir um modelo aproximado para a densidade p e, a seguir, com base nesta aproximação se calcular um minimizador do risco esperado tem que se fazer duas aproximações, o que pode conduzir a resultados de qualidade inferior ou erros imprevisíveis. Portanto, sempre que possível deve evitar-se situações deste tipo.

No caso de se pretender de facto determinar a função probabilidade, ou densidade condicionada $p(y|x)$, é necessário obter um modelo para o processo de geração de y baseado nalguma dependência subjacente, $f(x)$, para a qual se especifica uma forma funcional $p(y|x, f(x))$ e se maximiza a expressão relativamente a $f(x)$, obtendo-se, assim, a função $f(x)$ *que é mais provável* de ter gerado os dados.

Definição 38 (Verosimilhança) *A verosimilhança da amostra* $(x_1, y_1), \ldots (x_m, y_m)$ *dada uma dependência funcional subjacente a* $\{f(x_1), \ldots, f(x_m)\}$ *é dada por*

$$p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\} \mid \{f(x_1), \ldots, f(x_m)\}) =$$

$$= \prod_{i=1}^{m} p(x_i, y_i \mid \{f(x_1), \ldots, f(x_m)\}) =$$

$$= \prod_{i=1}^{m} p(y_i | x_i, \{f(x_1), \ldots, f(x_m)\})\ p(x_i)$$

Como se vê a verosimilhança depende apenas dos valores de $f(x_1), \ldots, f(x_m)$, não sendo pois em si uma funcional de $f(x)$. Por razões práticas pode retirar-se o factor $p(x_i)$, uma vez que ele não é dependente de $f(x)$, e converter-se os produtos em somas aplicando o negativo do logaritmo de:

$$p(\{x_1, \ldots, x_m\}, \{y_1, \ldots, y_m\} \mid \{f(x_1), \ldots, f(x_m)\}),$$

obtendo-se assim uma expressão que se designa *log-verosimilhança,* e que pode ser

convenientemente minimizada:

$$\mathcal{L}\left[\{f(x_1),\ldots,f(x_m)\}\right] = \sum_{i=1}^{n} -\ln p\left(y_i|x_i,\{f(x_1),\ldots,f(x_m)\}\right)$$

Dado que a função de custo com base na máxima verosimilhança pode não ser convexa e criar problemas de minimização procura-se em geral não utilizar esta função.

Note-se que no caso das funções de custo para a *regressão* obtém-se o mesmo resultado quer minimizando a *log-verosimilhança,* $\mathcal{L}\left[\{f(x_1),\ldots,f(x_m)\}\right]$, quer o *risco empírico,* $R_{emp}\left[\{f(x_1),\ldots,f(x_m)\}\right]$, desde que a função de custo escolhida seja $c(x,y,f(x)) = -\ln p(y|x,f)$. Supondo que os valores alvo foram gerados por uma dependência funcional subjacente $f_{verdad.}$com ruído aditivo ξ,cuja densidade é $p(\xi)$:

$$y_i = f_{verdad.}(x_i) + \xi_i$$

então, obtém-se $c(x,y,f(x)) = -\ln p_\xi(y - f(x))$

Assim, para a função de custo polinomial, por exemplo, $\frac{1}{d}|\xi|^d$ o modelo de densidade $p(\xi)$ correspondente é:

$$p(\xi) = \frac{d}{2\Gamma(1/d)} \exp\left(-|\xi|^d\right)$$

No caso de uma das funções de custo mais usadas em MVS regressoras, a função ε-insensitiva $|\xi|_\varepsilon$, corresponde-lhe o modelo de densidade:

$$p(\xi) = \frac{d}{2(1+\varepsilon)} \exp\left(-|\xi|_\varepsilon\right)$$

2.2.3 Consistência dos processos de aprendizagem

Um dos aspectos relevantes da teoria de Vapnik Vapnik (1998) da aprendizagem consiste em determinar-se as condições necessárias e suficientes para que um processo de aprendizagem baseado no princípio de indução da minimização do risco empírico seja consistente segundo a definição 42 que se vai dar mais à frente.

Recorde-se as seguintes três desigualdades probabilísticas clássicas que se utilizam na especificação e demonstração de teoremas sobre extremos superiores e inferiores das inferências das probabilidades associadas a variáveis aleatórias com base em processos empíricos.

Teorema 39 (Chernoff) *A seguinte desigualdade caracteriza a convergência da média empírica* $\frac{1}{m}\sum_{i=1}^{m}\xi_i$ *para o valor esperado de* $\mathbf{E}[\xi]$. *Seja* P *a medida de probabilidade (composta) da amostra* $\xi_1,\ldots,\xi_m$, *verifica-se que* P *goza da propriedade*

seguinte (indicada dentro das chavetas):

$$P\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi_i - \mathbf{E}\left[\xi\right]\right| \geq \varepsilon\right\} \leq 2\exp\left(-2m\varepsilon^2\right) \tag{2.8}$$

Teorema 40 (Markov) *A seguinte desigualdade caracteriza os extremos superiores no caso de uma variável aleatória não negativa ξ com distribuição P. Então, para todo o $\lambda > 0$, verifica-se a seguinte desigualdade:*

$$P\left\{\xi \geq \lambda\,\mathbf{E}\left[\xi\right]\right\} \leq \frac{1}{\lambda}$$

Teorema 41 (Hoeffding) *Caracteriza o limite do erro esperado a partir do erro empírico e da dimensão duma amostra. Seja a amostra $\xi_1, \ldots, \xi_m$ constituída por m exemplos independentes de uma variável aleatória ξ, com valores em $[a, b]$, e seja a sua média $Q_m = \frac{1}{m}\sum_{i=1}^{m}\xi_i$, então, para qualquer $\varepsilon > 0$, verifica-se:*

$$\left.\begin{array}{l} P\left\{(Q_m - \mathbf{E}\left[\xi\right]) \geq \varepsilon\right\} \\ P\left\{(\mathbf{E}\left[\xi\right] - Q_m) \geq \varepsilon\right\} \end{array}\right\} \leq \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right)$$

A desigualdade de Hoeffding uma vez que contém um extremo mais geral do que a de Chernoff tem mais aplicações do que esta. Por sua vez a desigualdade de Markov, que decorre da conhecida desigualdade de Chebyshev, é necessária para a demonstração da desigualdade de Hoeffding, a qual a qual se omite por falta de espaço, mas pode encontrar-se, por exemplo no apêndice de Anthony and Bartlett (1999).

Primeiramente, começa-se por considerar o caso simples da classificação binária, em que o conjunto das etiquetas $\mathbb{Y} = \{-1, +1\}$, usando-se uma função de custo que penaliza as classificações erradas. Assim, dada a função f, então, para cada exemplo (x_i, y_i) que se supõe extraído independentemente, o custo dado pela expressão

$$\xi_i = \frac{1}{2}\left|f\left(x_i\right) - y_i\right|$$

ou é nulo ou é unitário. Em termos de teoria das probabilidades esta situação corresponde a um ensaio de Bernoulli. Supondo, então, que $\xi_1, \ldots, \xi_m$ são amostras *iid* da variável aleatória

$$\xi = \frac{1}{2}\left|f\left(x\right) - y\right| \tag{2.9}$$

Usando a desigualdade de Chernoff (2.8), fazendo a substituição da variável aleatória ξ pela expressão (2.9) obtém-se uma expressão sobre o extremo superior que, de certa forma, exprime para uma dada função f o desvio do seu risco empírico

em relação ao seu risco verdadeiro:

$$P\left\{\left|R_{emp}\left[f\right]-R\left[f\right]\right|\geq\varepsilon\right\}\leq 2\exp\left(-2m\varepsilon^{2}\right)$$

No caso de se tratar de uma estimativa de regressão pode usar-se a desigualdade de Hoeffding e obter-se um extremos superiores semelhante desde que a função de custo $c\left(x,y,f\left(x\right)\right)$ seja limitada.

O erro de treino para uma dada função converge em probabilidade para zero:

$$\left|R_{emp}\left[f\right]-R\left[f\right]\right|\xrightarrow{P}0 \text{ quando } m\rightarrow\infty \tag{2.10}$$

o que quer dizer que, para todo o $\varepsilon>0$:

$$P\left\{\left|R_{emp}\left[f\right]-R\left[f\right]\right|>\varepsilon\right\}=0$$

A interpretação deste resultado requer alguns cuidados. Tendo-se em atenção a natureza probabilística da expressão (2.10), ela estabelece apenas que é pequena a probabilidade de se verificar um grande desvio entre os erros de teste e de treino, e tanto mais pequena quanto maior for o tamanho m do conjunto de treino. A expressão (2.10) não permite concluir que o desvio num exemplo particular nunca possa ser grande, nem que a função a ser aprendida pela máquina seja única e não possam existir muitas outras. É também preciso ter-se em consideração que quando se está a escolher uma função f tal que a média de ξ_i seja tão pequena quanto possível está-se a procurar activamente *no pior dos casos* uma função muito especial no que diz respeito ao custo médio e ao risco empírico.

Para se exprimir objectivamente o que é preciso para a minimização do risco empírico funcionar tem que se definir o conceito estatístico de consistência do *princípio de minimização do risco empírico* (MRE).

Definição 42 (Consistência. Vapnik, 1998) *O princípio é consistente para o conjunto de funções $c\left(x,y,f\left(x,w\right)\right)$ em que os parâmetros w pertencem a um dado conjunto finito $\mathbb{W}$, e para a distribuição de probabilidades $P\left(x,y\right)$, se o método de minimização do risco empírico, aplicado a cada amostra de dimensão m, produz uma sucessão de funções $c\left(x,y,f\left(x,w_{m}\right)\right)$ tal que:*

$$R\left[f\left(x,w_{m}\right)\right]\xrightarrow[m\rightarrow\infty]{P}\inf_{f\in\mathbb{F}}R\left[f\right]$$

$$R_{emp}\left[f\left(x,w_{m}\right)\right]\xrightarrow[m\rightarrow\infty]{P}\inf_{f\in\mathbb{F}}R\left[f\right]$$

Por outras palavras, o método MRE é consistente se ele gerar uma sucessão de funções $c\left(x,y,f\left(x,w_{m}\right)\right)$ para as quais o risco esperado e o risco empírico convergem

em probabilidade para o mesmo limite, o ínfimo do risco no conjunto $\mathbb{F}$.

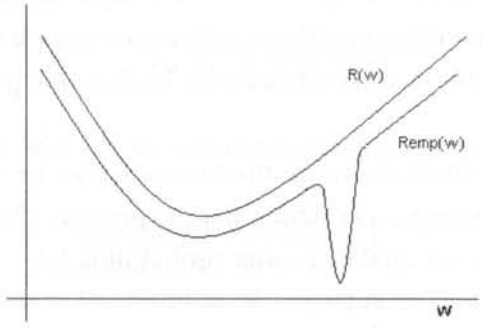


Figura 2.8: Grandes desvios

A minimização da função de risco pode colocar problemas sérios, como se mostra na figura 2.8 a existência de um intervalo de confiança entre os valores de duas funções, por exemplo duas funções aproximam-se em 95% dos casos (do seu intervalo de definição) não garante que os seus mínimos também estejam próximos. Esta garantia é dada através de teoremas, referidos em Vapnik (1998), que fornecem intervalos de confiança e que relacionam $R_{emp}[f]$ e $R[f]$ para uma dada função resultante de um processo de aprendizagem.

Um dos resultados chave da teoria de aprendizagem de Vapnik é o teorema seguinte.

Teorema 43 (Chave da T.E.A.) *Seja o conjunto de funções $c(x, y, f(x, w))$ em que os parâmetros w pertencem a um dado conjunto finito $\mathbb{W}$. Suponha-se que as funções de custo c são limitadas inferior e superiormente:*

$$A \leq \int c(x, y, f(x, w))\, dP(x, y) \leq B.$$

Então, uma condição necessária e suficiente de consistência não trivial do princípio de minimização do risco empírico consiste em o risco empírico $R_{emp}[f(x, w)]$ ser uniformemente convergente para o risco verdadeiro $R[f(x, w)]$ no conjunto de funções $c(x, y, f(x, w))$ e no sentido em que

$$\lim_{m \to \infty} P\left\{ \sup_{w \in \mathbb{W}} \left(R[f(x, w)] - R_{emp}[f(x, w)] \right) > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0 \tag{2.11}$$

A presença de um limite superior na expressão (2.11) implica que a consistência do principio MRE é necessariamente determinada pela pior das funções do conjunto de funções $c(x, y, f(x, w))$, $w \in \mathbb{W}$. Este teorema é um resultado teórico não sendo directamente aplicável, uma vez que nem o risco $R[f(x, w)]$ se pode determinar,

nem a distribuição de probabilidades $P(x, y)$ é conhecida. Utilizando várias noções intermediárias, tais como a entropia VC ou a função de crescimento, que se referiu atrás, pode deduzir-se outras condições suficientes para a consistência do princípio MRE, assim como obter-se uma sucessão de limites que permitem obter resultados úteis.

Por exemplo, a aprendizagem de funções binárias de contradomínio $\{-1, +1\}$, na qual as funções de custo são dadas pela expressão (2.9), pode obter-se um resultado muito simples, no qual com uma probabilidade $1 - \delta$ se verifica as seguintes desigualdades:

$$R_{emp}[f] - \varepsilon(m, d_{vc}, \delta) \leq R[f] \leq R_{emp}[f] + \varepsilon(m, d_{vc}, \delta) \tag{2.12}$$

em que:

$$\varepsilon(m, d_{vc}, \delta) = \sqrt{\frac{d_{vc}}{m}\left(1 + \log_2 \frac{2m}{d_{vc}}\right) - \frac{1}{m}\log_2 \frac{\delta}{4}}$$

A largura $\varepsilon(m, d_{vc}, \delta)$ deste intervalo de confiança depende do tamanho m do conjunto de exemplos, da dimensão VC e da família de funções a aprender e do nível de confiança $1 - \delta$.

2.2.4 Minimização do risco estrutural

O principio da minimização do risco empírico é mais eficaz quando a razão $\frac{m}{d_{VC}}$ tem valores elevados (superiores a 20), porque neste caso a largura $\varepsilon(m, d_{vc}, \delta)$ do intervalo de confiança torna-se mais estreita e o mínimo do risco empírico $R_{emp}[f]$ aproxima-se do mínimo do risco real $R[f]$. Por outras palavras o *erro de aprendizagem* (ou amostral, num conjunto de treino) converge para o *erro de generalização* (ou verdadeiro, que se pode estimar num conjunto de teste disjunto do conjunto de treino). Reciprocamente, quando a razão $\frac{m}{d_{VC}}$ toma valores pequenos a largura do intervalo de confiança começa a ser muito significativa.

Desde que o número de exemplos m seja constante convém então minimizar o *risco garantido*, dado pela soma do risco empírico $R_{emp}[f]$ e do intervalo de confiança $\varepsilon(m, d_{vc}, \delta)$. Define-se no seio do conjunto das funções $\mathbb{F}$ em que o algoritmo de aprendizagem trabalha uma sucessão de subconjuntos imbricados

$$\mathbb{S}_1 \subset \mathbb{S}_2 \subset \ldots \subset \mathbb{F},$$

com *capacidades*, que se podem exprimir no caso de funções discretas pela dimensão VC de cada $\mathbb{S}_i$, crescentes,

$$d_{vc1} < d_{vc2} < \ldots < d_{vc}.$$

Realiza-se, seguidamente, uma optimização em duas etapas. Primeiramente, calcula-se o mínimo dos riscos empíricos em cada subconjunto:

$$R_i[f_*] = \min_{f \in \mathbb{S}_i} R[f]$$

e, em seguida, faz-se a selecção do sub-conjunto que apresenta o melhor risco garantido (figura 2.9):

$$\min_i R_i[f_*] + \varepsilon\left(m, (d_{vc})_i, \delta\right).$$

Quando se escolhem subconjuntos de capacidade crescente, o valor óptimo do risco empírico diminui por se dispor de uma família de funções mais rica. Ao mesmo tempo a largura do intervalo de confiança aumenta porque a capaciade d_{vc} também aumenta. A sucessão de subconjuntos define uma "*estrutura*" que representa uma solução de compromisso para o problema da aprendizagem, desde que a selecção de um subconjunto de fraca capacidade não penalize em demasiado o risco empírico, mas, antes pelo contrário, reduza o desvio entre o risco empírico e o risco esperado.

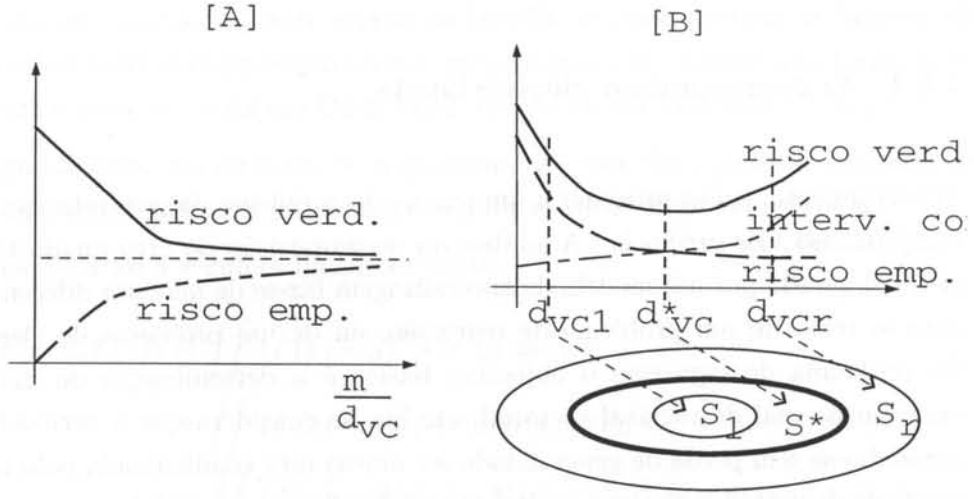


Figura 2.9: Minimização do Risco Estrutural

2.3 Selecção de Modelos

A aplicabilidade dos resultados inspirados na teoria desenvolvida por Vapnik para o problema da generalização é nalgumas situações de aprendizagem ainda um problema em aberto. A garantia de uma boa generalização passa, em geral, por uma problema de determinação de um compromisso entre a complexidade e a robustez da aprendizagem. Na prática procura-se adaptar a complexidade do algoritmo de

aprendizagem quer à dificuldade do problema, quer à quantidade e qualidade de exemplos disponíveis. Para tal recorre-se a aproximações heurísticas clássicas não dependentes do algoritmo, por assim dizer, usando métodos bem conhecidos dos quais se faz a seguir uma breve revisão.

Para um conjunto de exemplos e um tamanho de amostra de treino dados existe um modelo de complexidade óptimo que corresponde a um erro de generalização mínimo. Por consequência, qualquer método de aprendizagem a partir de um conjuntos de exemplos finitos carece de mecanismos que controlem a sua complexidade admissível. Tais mecanismos incluem, por exemplo, a regularização, o decrescimento dos pesos, e vários procedimentos construtivos na topologia das redes, quer de crescimento (*greedy*) quer de decrescimento (*prunning*), controlados. Existem ainda métodos que permitem determinar a complexidade mínima de um modelo baseadas em estimativas analíticas e empíricas do risco e em técnicas de re-amostragem.

2.3.1 O compromisso viés-variância

Primeiramente, vai-se proceder a um exame do problema da generalização à luz do compromisso viés-variância. A análise da decomposição do erro quadrático médio (EQM) gerado por um modelo de aprendizagem faz-se de maneira diferente, consoante se trata de um problema de regressão, ou de um problema de classificação. No problema de regressão o objectivo básico é a determinação de um modelo, uma função real de variável vectorial, em que se considera que a variável de saída (suponha-se sem perda de generalidade ser única) está condicionada pelo vector das variáveis de entrada. Designe-se por $f(x) \in \mathbb{F}$ a função aprendida por um dado modelo com base num conjunto de treino $\mathbb{Z} = \{(x_i, y_i) \in \mathbb{X} \times \mathbb{Y},\ i = 1 \ldots m\}$. Outras funções da classe $\mathbb{F}$ poderiam ser aprendidas pelo modelo utilizando outros conjunto de treino extraídos de $\mathbb{X} \times \mathbb{Y}$, produto cartesiano do conjunto das entradas pelo conjunto das etiquetas. Supondo ser aleatório o processo de recolha de exemplos seria de esperar que as aproximações dadas por cada uma destas funções fossem, para um dado conjunto de teste, umas melhores, outras piores. A medida habitualmente usada na avaliação da eficácia do modelo é o EQM em relação ao óptimo desejado. Assim, pode determinar-se a média para todos os conjuntos de treino retirados de $\mathbb{X} \times \mathbb{Y}$ com o mesmo tamanho m calculando-se o primeiro momento da distribuição dos EQM:

$$\mathbf{E}\left[(f(\mathbf{x}) - y)^2\right] =$$

$$
\begin{aligned}
&= \mathbf{E}\left\{\left[f(\mathbf{x}) - \mathbf{E}(y) + \mathbf{E}(y) - y\right]^2\right\} \\
&= \mathbf{E}\left\{\left[\left(f(\mathbf{x}) - \mathbf{E}(y)\right)^2 + \left(\mathbf{E}(y) - y\right)^2 + 2\left(f(\mathbf{x}) - \mathbf{E}(y)\right)\left(\mathbf{E}(y) - y\right)\right]^2\right\} = \\
&= \underbrace{\mathbf{E}\left\{\left(f(\mathbf{x}) - \mathbf{E}(y)\right)^2\right\}}_{(\text{viés})^2} + \underbrace{\mathbf{E}\left\{\left(\mathbf{E}(y) - y\right)^2\right\}}_{\text{variância}}
\end{aligned}
$$

Verifica-se experimentalmente que os modelos dados por redes neuronais com elevada adaptabilidade (complexidade) geram valores de viés baixos e valores de variância elevados. Porém, a natureza do conjunto de dados $\mathbb{Z}$ também influencia os componentes do erro. Se os dados de treino contiverem muito ruído é preciso estabelecer empiricamente um equilíbrio entre a minimização a fazer-se numa e na outra componente. Este equilíbrio é facilitado em três casos especiais. primeiro, quando se conhece bem o nível de ruído dos dados, e sabe-se que este não é muito elevado. Segundo, quando o conjunto de treino tem um tamanho grande. Finalmente, se for conhecida informação *a priori* acerca da família de onde provêm as funções alvo. Uma análise desta decomposição no caso de problemas de classificação torna-se mais elaborada e pode ser vista em Duda et al. (2001) ou em Friedman (1997).

Seguidamente, vai formular-se o problema no caso da *regressão* supondo, sem perda de generalidade, que existe apenas uma só variável de entrada.

Consideremos a seguinte função de custo:

$$
\begin{aligned}
c(x, y, f(x)) &= \iint (f(x) - y)^2 \, p(x, y) \, dxdy = \\
&= \iint (f(x) - y)^2 \, p(x|y) \, p(x) \, dxdy = \\
&= \iint (f(x) - \mathbf{E}[y|x] + \mathbf{E}[y|x] + y)^2 \, p(x|y) \, p(x) \, dxdy = \\
&= \iint (f(x) - \mathbf{E}[y|x])^2 + (\mathbf{E}[y|x] + y)^2 \, p(x|y) \, p(x) \, dxdy
\end{aligned}
$$

O terceiro termo resultante do desenvolvimento do quadrado do integrando:

$$
\iint 2\left(f(x) - \mathbf{E}[y|x]\right)\left(\mathbf{E}[y|x] + y\right) p(x) \, dxdy,
$$

anula-se em consequência da integração sobre y.

Se considerarmos, por definição de valor esperado, que $\mathbf{E}[y|x] = \int y \, p(y|x) \, dy$ e que

$$
\mathbf{E}[y^2|x] = \int y^2 \, p(y|x) \, dy \, ,
$$

então, a função de custo passa a escrever-se:

$$c(x, y, f(x)) = \int (f(x) - \mathbf{E}[y|x])^2 p(x)\, dx + \int (\mathbf{E}[y|x] + y)^2 p(x)\, dx =$$
$$= \int (f(x) - \mathbf{E}[y|x])^2 p(x)\, dx + \int \left(\mathbf{E}[y^2|x] + \mathbf{E}[y\,|x|]^2\right) p(x)\, dx$$

O segundo termo do modelo representa o ruído intrínseco dos dados, é independente do modelo, pelo que no sentido desta função de custo o modelo óptimo de erro é o que satisfaz a relação $f(x) = \mathbf{E}[y|x]$, isto é, prediz a média condicional.

Considerando, agora, todos os conjuntos de treino possíveis que contêm m exemplos provenientes da mesma distribuição de probabilidades conjunta $p(x, y)$, vê-se que o primeiro integral contém uma medida que permite controlar o erro do modelo que pode ser estimada com o conjunto de treino mas trata-se de uma medida muito enviezada. Torna-se preferível determinar esta quantidade considerando a média de várias conjuntos de treino:

$$\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[(f(x) - \mathbf{E}[y|x])^2\right]$$

Se o modelo for um estimador perfeito da função de regressão $\mathbf{E}[y|x]$ este erro tem um valor nulo. Se não for este o caso, então, pode acontecer que a saída do modelo seja em média diferente da função de regressão, designando-se esta discrepância por *viés*. Pode também o modelo ser muito sensível a um conjunto particular de dados de $\mathbb{X} \times \mathbb{Y}$. Para um exemplo (x, y) a saída $f(x)$ deste modelo pode ser maior que o valor esperado para um certo conjunto de dados de treino e mais pequeno para outro, discrepância que é habitualmente designada por *variância*. Uma variância muito elevada pode ser devida a que:

$$\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[(f(x) - E[f(x)])^2\right] = \mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[(f(x) - g(x))^2\right] = \mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[\varepsilon^2\right]$$

Este termo representa de facto a variância do ruído sobre os dados e pode ser muito elevado. Constata-se assim que tem que haver um compromisso entre o viés e a variância (ver Figura 2.10). Uma função que se ajusta demasiado aos dados tende a ter uma grande variância e a originar assim uma grande erro. Pode diminuir-se a variância suavizando a função $f(x)$ mas indo-se longe de mais neste procedimento, então, o viés pode tornar-se muito grande e o erro ainda crescer mais.

2.3.2 Critérios clássicos de selecção de modelos

A estimativa do risco funcional pode fazer-se utilizando de fórmulas analíticas que exprimem o risco em função do erro amostral (erro empírico), regularizado por alguma medida da complexidade do modelo, ou , alternativamente usando-se técnicas

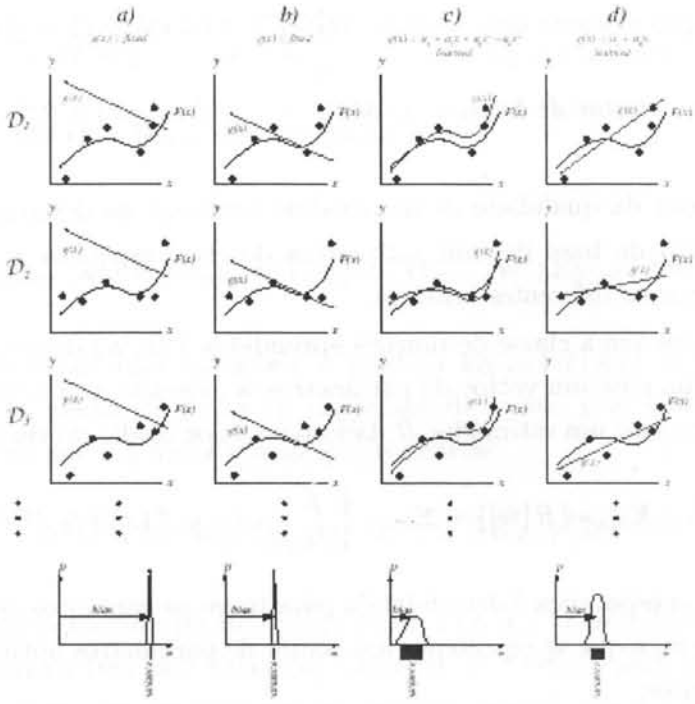


Figura 2.10: Compromisso viés-variância ilustrado no domínio da regressão (Duda et al., 2000, p.467)

de re-amostragem.

Quando se conhecem formas analíticas de determinar estimativas do risco funcional estas podem utilizar-se na selecção de modelos escolhendo a complexidade destes de forma a minimizar o valor da estimativa do *risco esperado*. A literatura estatística para modelos lineares contém várias fórmulas para a determinação de estimativas do risco esperado $\widehat{R}[f, d, m]$. As fórmulas que se reproduzem a seguir são referidas em Cherkassky et al. (1999), que também citam as referências originais, contêm pequenas modificações na notação por razões de clareza. Estas fórmulas diferem entre si pelo coeficiente de penalização do erro de treino, $\gamma(d, m)$. Este depende da complexidade do modelo, ou graus de liberdade, d, e do tamanho da amostra m :

$$\widehat{R}[f, d, m] = \gamma(d, m) \left[\frac{1}{m} \sum_{i=1}^{m} (f(x) - y)^2 \right]$$

em que $\gamma(d, m)$ é dado por uma das seguintes quatro fórmulas (propostas por vários autores):

1. *Erro final de predição* (FPE) $\quad \gamma(d, m) = \left(1 + \frac{d}{m}\right)\left(1 - \frac{d}{m}\right)^{-1}$
2. *Critério B da informação* (BIC) $\quad \gamma(d, m) = 1 + \frac{\ln \frac{d}{m}}{2} \frac{d}{m}\left(1 - \frac{d}{m}\right)^{-1}$

3. *Validação cruzada generalizada* (GCV) $\gamma(d,m) = \left(1 - \frac{d}{m}\right)^{-2}$

4. *Modelo selector de Shibata* (SMS) $\gamma(d,m) = 1 + 2\frac{d}{m}$

A avaliação da qualidade de um modelo baseia-se na determinação através de uma expressão de base de uma estimativa do risco esperado a partir da qual se possam derivar os diferentes critérios.

Suponha-se que a classe de funções aprendidas $f(x, \mathbf{w})$ depende apenas do conjunto de treino e de um vector de parâmetros $\mathbf{w}$. Para se avaliar a qualidade de um modelo procura-se um estimador $\widehat{R}$ dado pelo valor médio do risco esperado:

$$\widehat{R} = \mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R[\hat{\mathbf{w}}]\} = \mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left\{\int_{\mathbb{X}\times\mathbb{Y}} c(x, y, f(x, \hat{\mathbf{w}}))\, dP(x, y)\right\}$$

supondo que a esperança é determinada para todos os conjuntos de dados de $\mathbb{X} \times \mathbb{Y}$, de dimensão m, e que $\hat{\mathbf{w}}$ representa um vector de parâmetros óptimos para cada um destes conjuntos.

Suponha-se que $\mathbf{w}^*$ é o vector parâmetro em que se verifica um mínimo do risco esperado, isto é, $\nabla R(\mathbf{w}^*) = 0$, e que $\hat{\mathbf{w}}$ está próximo de $\mathbf{w}^*$. É, então, possível fazer-se um desenvolvimento em série de Taylor de ordem dois em redor de $\mathbf{w}^*$:

$$R(\hat{\mathbf{w}}) \approx R(\mathbf{w}^*) + \frac{1}{2}\Delta\mathbf{w}^T\mathbf{H}(\mathbf{w}^*)\,\Delta\mathbf{w}$$

em que:

$$\Delta\mathbf{w} = \hat{\mathbf{w}} - \mathbf{w}^*,$$

e $\mathbf{H}(\mathbf{w})$ é a matriz Hessiana da função de risco $R(\mathbf{w})$. Supondo que é possível fazer-se um desenvolvimento em série de Taylor do risco empírico em redor de $\mathbf{w}^*$, admitindo ainda que $\nabla R_{emp}(\hat{\mathbf{w}}) = 0$ e designando por $\mathbf{H}_{emp}$ a Hessiana de $R_{emp}(w)$ pode escrever-se:

$$R_{emp}(\hat{\mathbf{w}}) \approx R_{emp}(\mathbf{w}^*) + \frac{1}{2}\Delta\mathbf{w}^T\mathbf{H}_{emp}(\mathbf{w}^*)\,\Delta\mathbf{w}$$

Seja a igualdade algébrica:

$$\Delta\mathbf{w}^T\,\mathbf{H}\,\Delta\mathbf{w} = tr\left\{\mathbf{H}\,\Delta\mathbf{w}\,\Delta\mathbf{w}^T\right\}$$

e a aproximação:

$$\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R_{emp}[\mathbf{w}^*]\} \approx R(\mathbf{w}^*)$$

Calcula-se para os conjuntos de treino de tamanho m os valores médios $\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R[\hat{\mathbf{w}}]\}$ e de $\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R_{emp}[\hat{\mathbf{w}}]\}$:

$$\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R[\hat{\mathbf{w}}]\} \approx R(\mathbf{w}^*) + \frac{1}{2} tr\left\{\mathbf{H}(\mathbf{w}^*)\,\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[\Delta\mathbf{w}\,\Delta\mathbf{w}^T\right]\right\}$$

Suponha-se que $\mathbf{H} \approx \mathbf{H}_{emp}$ na vizinhança de $\mathbf{w}^*$:

$$\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R_{emp}[\hat{\mathbf{w}}]\} \approx R_{emp}(\mathbf{w}^*) + \frac{1}{2} tr\left\{\mathbf{H}_{emp}(\mathbf{w}^*)\,\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[\Delta\mathbf{w}\,\Delta\mathbf{w}^T\right]\right\}$$

Combinando estas duas equações, e fazendo $\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R(\hat{\mathbf{w}})\} \approx R_{emp}(\hat{\mathbf{w}})$, i.é, substituindo-se a média de todos os conjuntos de treino pela estimativa do erro empírico uma vez que é a única conhecida, obtem-se:

$$\widehat{R} = \mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\{R[\hat{\mathbf{w}}]\} \approx R_{emp}(\hat{\mathbf{w}}) + tr\left\{\mathbf{H}(\mathbf{w}^*)\,\mathbf{E}_{\mathbb{X}\times\mathbb{Y}}\left[\Delta\mathbf{w}\,\Delta\mathbf{w}^T\right]\right\} \tag{2.13}$$

Para uma dada escolha da função de custo $c(x, y, f(x, \hat{\mathbf{w}}))$ pode deduzir-se as seguintes expressões (ver por exemplo, Cibas et al. (1995)) para três critérios de selecção de modelos que apresentam um aspecto semelhante a algumas referidas anteriormente, diferindo nalguns pormenores que poderão ter a ver com certas suposições restritivas.

Por exemplo, se a função de custo $c(x, y, f(x, \hat{\mathbf{w}})) = -\frac{1}{m}\ln\mathcal{L}(w)$, em que $\mathcal{L}(w)$ é função de verosimilhança, sob certas hipóteses, o último termo da expressão (2.13) pode exprimir-se pela dimensão do espaço dos parâmetros d, isto é, a complexidade do modelo (ou graus de liberdade), obtendo-se assim o critério AIC de Akaike (1972) (referido em Cherkassky et al. (1999))

$$\widehat{R}_{AIC} = -\frac{1}{m}\ln\mathcal{L}(\hat{\mathbf{w}}) + \frac{d}{m}.$$

O critério BIC tem por base argumentos bayesianos foi proposto por Schwartz (1978) (referido em Cherkassky et al. (1999)):

$$\widehat{R}_{BIC} = -\frac{1}{m}\ln\mathcal{L}(\hat{\mathbf{w}}) + \frac{d}{2m}\ln m.$$

A partir do momento em que $\ln m > 2$ a penalização deste critério nos modelos mais complexos é superior à do critério AIC.

Supondo que a função de custo é o erro quadrático médio

$$c(x, y, f(x, \mathbf{w})) = \frac{1}{m}\sum_i (y_i - f(x_i, \mathbf{w}))^2$$

e utilizando-se a expressão assimptótica seguinte para a estimativa da variância-

covariância referida no segundo termo da expressão (2.13),

$$E_{\mathbb{X}\times\mathbb{Y}}\left[\Delta\mathbf{w}\,\Delta\mathbf{w}^T\right] \approx \frac{2\sigma_\nu^2}{m}\mathbf{H}^{-1}\left(w^\circ\right),$$

em que σ_ν^2 é a variância do ruído, obtém-se o seguinte critério FPE :

$$\widehat{R}_{FPE} = \frac{1}{m}\sum_i \left(y_i - f\left(x_i, \hat{\mathbf{w}}\right)\right)^2 + 2\sigma_\nu^2\frac{d}{m}.$$

Para finalizar vai-se fazer um esboço do critério do comprimento da descrição mínima (MDL) de Rissanen, baseando-nos em Principe et al. (2000b). Um desenvolvimento deste tópico existe, por exemplo, em McConnell (1995).

Como se viu anteriormente não basta considerar-se o ajustamento de um modelo aos exemplos, é preciso contrabalançar os seus graus de liberdade, o que pode ser feito em termos do comprimento de um código. Pode pensar-se que os exemplos se podem considerar como tendo um dado comprimento intrínseco de código, e tentar-se uma sua descrição através de um código a ser definido.

O conjunto de exemplos requer pois um certo número de *bits* para a sua descrição. Se os dados contiverem ruído aleatório, é preciso usar-se toda a amostra para descrever esses dados, e os comprimentos do código e dos exemplos vão ser o mesmo. Mas, isso não acontece se os dados forem, por exemplo, gerados por um modelo linear, para o qual apenas se precisa de dois números, o declive da recta e a sua intercepção no eixo dos y.

Quando se pesquisa um modelo que representa os dados está-se efectivamente a criar uma sua descrição, de diferentes maneiras, seja através da topologia e parâmetros do modelo, seja através do erro amostral que ele gera. Considerando o caso em que se atribui um certo número de bits para codificar o erro amostral, $C\left(E_i\right)$, e para representar os parâmetros do modelo, $C\left(M_i\right)$, então, a descrição dos dados usando um determinado modelo M_i, pode ser dada pela expressão:

$$C\left(M_i, E_i\right) = C\left(M_i\right) + C\left(E_i\right).$$

A descrição do conjunto de dados mais eficiente é aquela que minimiza todo o comprimento do código, isto é, o número total de bits para representar o erro amostral e os parâmetros do modelo (ver figura 2.11):

$$\min_i C\left(M_i, E_i\right).$$

Esta ideia é muito interessante porque alia a complexidade, ou tamanho do modelo, o valor do seu erro amostral. Usando um número muito pequeno de parâmetros

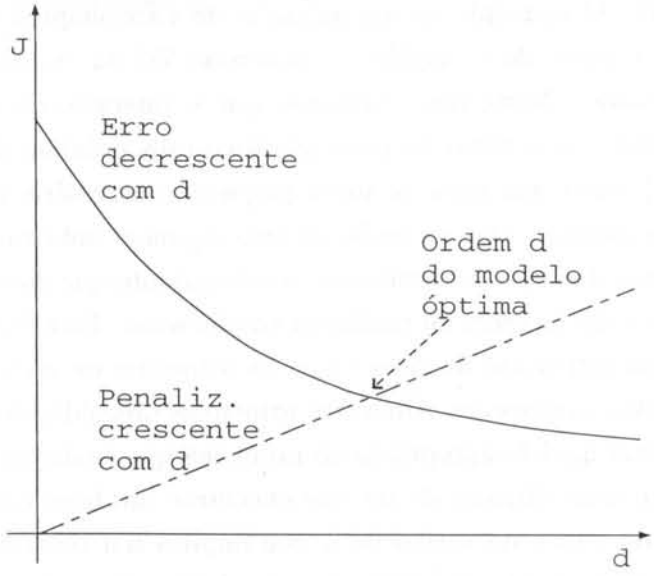


Figura 2.11: Melhor ordem do modelo em conformidade com o critério de Akaike

o erro será grande, e é preciso muitos bits para representar este erro. Por outro lado, usando-se um modelo muito complexo usam-se muitos bits para descrever o modelo mas são precisos poucos para descrever os erros. O melhor compromisso em termos de comprimento do código está num equilíbrio entre modelos simples e erros razoáveis. Uma das mais simples implementações desta ideia consiste em penalizar o erro quadrático médio no treino incluindo um termo que cresce com o aumento do tamanho do modelo, tal como foi proposto no modelo AIC de Akaike, e que se referiu acima.

2.4 Sumário

Existem ligações teóricas entre o domínio da estatística e o das redes neuronais que se procuram ilustrar neste capítulo. Por exemplo, em estatística termos como estimação, parâmetros e valor dos parâmetros, traduzem-se em linguagem das redes neuronais por "*aprendizagem*", "*pesos*" e "*conhecimento*", respectivamente. Poder-se-á dizer que as redes neuronais visam mais o objectivo da *previsão* enquanto que os modelos estatísticos visam dum modo geral o objectivo da *interpretação*. A dimensão de Vapnik-Chervonenkis de uma classe de funções é um instrumento potente na aprendizagem, que já tinha sido pressentido por Cover (1965) num enquadramento de funções lineares e que Vapnik desenvolveu, definindo as suas propriedades para problemas de classificação dicotómica, do mesmo tipo dos que estavam a ser estudados pelos métodos PAC, "provavelmente e aproximadamente correcto", de

Valiant (1984). O princípio da minimização do risco empírico é eficaz quando a razão entre o número de exemplos e a dimensão VC da classe de funções da rede neuronal é grande. Neste caso verifica-se que o intervalo de confiança tem uma largura pequena e o mínimo do risco empírico está próximo do mínimo do risco verdadeiro. Quando esta razão se torna pequena é necessário implementar no sistema de aprendizagem uma sucessão de subconjuntos imbricados do conjunto da funções implementadas, e determinar-se o subconjunto que garanta um mínimo do risco empírico e da margem de confiança combinados. Esta é a ideia do princípio da minimização estrutural do risco e que as máquinas de vectores de suporte implementam automaticamente. Uma das principais dificuldades da aprendizagem é a selecção de um modelo apropriado ao problema que se deseja tratar. Na prática está-se perante uma situação de ter que encontrar um bom compromisso entre as componentes do viés e da variância, o que implica um controlo da complexidade do modelo. A ideia de complexidade, que intuitivamente se tem quando por exemplo se compara uma modelo de regressão linear com um não-linear, foi objecto de uma formulação rigorosa devida a Geman et al. (1992) e que a teoria estatística da aprendizagem veio enriquecer.

2.5 Referências bibliográficas e notas históricas

A exposição do tema deste capítulo inclui uma selecção de conceitos básicos relativos aos modelos gerais da teoria da aprendizagem estatística. Adoptaram-se abordagens introduzidas recentemente, em particular, as devidas a (i) Vapnik (1995), Vapnik (1998), e Vapnik (1999). e (ii) Anthony (1997), Anthony (1998), Anthony and Bartlett (1999) e Anthony and Holden (1998).

Tentou-se integrar outros pontos de vista, salientando-se em particular as contribuições teóricas expostas com uma a forte preocupação didáctica devidas a: (i) Bishop (1995) sobre viés-variância e regularização (pp.332-346); sobre selecção da ordem dos modelos e dimensão VC (pp.371-380); (ii) Haykin (1999); (iii) Poggio et al. (2002), (iv) Cristianini and Shawe-Taylor (2000).

Compararam-se diferentes perspectivas: (i) que dão mais relevância ao ponto de vista estatístico, como por exemplo, Soulie (1997), Gallinari (1997), Gascuel (1997), Cibas (1996) e Cherkassky and Mulier (1998); (ii) que focam aspectos da geometria e topologia algébrica, por exemplo, Anthony and Bartlett (1999), Schölkopf et al. (1999), Smola (1998); (iii) que também consideram o ponto de vista Bayesiano, MacKay (1992c), Bishop (1995); (iv) que estabelecem pontes para as aproximações informático-teoréticas Mackay (2002), Principe et al. (2000b), (v) que são significativas da escola portuguesa de reconhecimento de padrões e redes neuronais Almeida (1987), Almeida et al. (1997), Silva (1998), Pedroso and Murata (1999), Marques

(1999) e Marques_de_Sá (2001a).

Capítulo 3

Máquina de Vectores de Suporte e Rede Neuronal

Introdução

Este capítulo tem por objectivo a revisão de aspectos relevantes relacionados com o funcionamento, propriedades e condições de convergência dos algoritmos que dão realidade à aprendizagem supervisada.

A secção 3.1 introduz o conceito de hiperplano de classificação óptimo que se pode construir com um *máquina de vectores de suporte (MVS)*, o algoritmo da optimização com restrições que permite a obtenção de margens de separação largas em problemas de classificação dicotómica. Faz-se ainda uma generalização do algoritmo da optimização com restrições ao problema da regressão linear, e não-linear, do ajustamento de rectas, curvas ou superfícies a dados experimentais.

A secção 3.2 mostra, em face da natureza de um problema "*incompletamente formulado*", como é possível encontrar uma solução baseada na suavização de funções, graças à aplicação dos princípios variacionais que estão na base da *teoria da regularização,* a qual se revê sumariamente. Nesta perspectiva, tipicamente estatística, as soluções que se baseiam na suavização dos dados parecem despertar mais interesse do que as que recorrem apenas a uma interpolação exacta dos dados Wahba (1990).

A secção 3.3 apresenta o modelo de *rede de funções de base radiais* e a forma como este modelo se integra na teoria da regularização. Faz-se também uma revisão de alguns conceitos da teoria da aproximação, nos quais o modelo de rede de *funções de base radiais* tem as suas origens.

A secção 3.4 apresenta uma perspectiva baseada na teoria estatística da aprendizagem para a construção do algoritmo de um *perceptrão* simples, assim como um resumo de alguns dos principais aspectos considerados na teoria clássica das redes de *perceptrão de multicamada.*

3.1 Máquina de Vectores de Suporte

3.1.1 Classificação

O problema da aprendizagem de uma máquina para reconhecimento de padrões pode exprimir-se formalmente como um problema de estimação de uma função vectorial $h : \mathbf{R}^n \rightarrow \{-1, 1\}$, que a máquina induz a partir de exemplos de treino. Seja $\mathbb{X}$ um conjunto pertencente ao espaço euclidiano real n-dimensional, ou a qualquer outro espaço pré-Hilbertiano dotado de um produto interno. Seja o conjunto dicotómico $\mathbb{Y} = \{-1, 1\}$ o contradomínio da função vectorial h (por vezes, há conveniência em que contradomínio seja o conjunto binário $\mathbb{Y} = \{0, 1\}$), e seja $\mathbb{Z}$ o produto cartesiano do conjunto das entradas $\mathbb{X}$ pelo conjunto das saídas $\mathbb{Y}$:

$$\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$$

Suponha-se que se tem um conjunto de treino, de tamanho m, com um número suficiente de exemplos, em que cada um é constituído por um par de vectores de entrada e saída:

$$\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\} \in \mathbb{Z}$$

Pretende-se determinar uma função h,pertencente a uma dada classe de funções $\mathbb{H}$, capaz de classificar correctamente qualquer outro exemplo desconhecido $(\mathbf{x}, y) \in \mathbb{Z}$. Portanto, para um qualquer exemplo $(\mathbf{x}, y)$ pertencente à mesma população que originou os exemplos de treino e cuja distribuição de probabilidade $P(\mathbf{x}, y)$ existe, mas é desconhecida, deverá verificar-se:

$$h(\mathbf{x}) = y.$$

Não existe necessariamente uma única função $h(\mathbf{x})$ que satisfaça a condição de classificar correctamente todos os exemplos de treino. Podem existir várias funções $h_i(\mathbf{x})$ que cumprem a especificação, no entanto, terão comportamentos diferentes quando aplicadas a exemplos desconhecidos. Isto equivale a dizer que pelo facto de durante o treino se verificar uma minimização do funcional *risco empírico*, dado pela expressão (2.6), tal não implica necessariamente que, para um conjunto de teste disjunto do de treino, se obtenha um valor mínimo para o *risco esperado*, dado pela expressão (2.5).

No capítulo anterior foi referido um aspecto da teoria da aprendizagem que se relaciona com a necessidade de restringir a escolha das funções $h(\mathbf{x})$ a uma determinada classe de funções $\mathbb{H}$, e verificou-se ser mais apropriado usar-se uma determinada medida da capacidade de $\mathbb{H}$, *a dimensão VC* da classe de funções $\mathbb{H}$. Uma das motivações em relação à máquina de vectores de suporte resulta do facto de os

limites teóricos do *risco esperado* $R[f]$ não dependerem explicitamente do número de variáveis do espaço de entradas. Recorde-se para o caso da aprendizagem de funções binárias de contradomínio $\{\pm 1\}$ que na expressão (2.12) do *risco esperado* definida numa norma L_1:

$$R_{emp}[f] - \varepsilon(m, d_{vc}, \delta) \leq R[f] \leq R_{emp}[f] + \varepsilon(m, d_{vc}, \delta)$$

os limites do risco $R[f]$ não dependem do número de variáveis do espaço de entradas, mas apenas do tamanho da amostra m, da *VC-dimensão* d_{vc}, e de uma constante $\delta \in [0,1]$, (mais propriamente, $1-\delta$ representa um nível de confiança nos limites não tendo a ver com o desempenho da máquina). Um outro facto a salientar diz respeito aos limites de $R[f]$ serem automaticamente minimizados quando se procede à maximização da *margem*, entendendo-se por margem, que se denota por γ, a distância mínima entre o hiperplano separador das classes e os exemplos de treino mais próximos desse hiperplano.

3.1.2 Hiperplano de classificação óptimo

Uma das classes de funções de decisão $\mathbb{H}$ com dimensão d_{VC} determinável é a chamada classe dos hiperplanos. Seja $\mathbb{X}$ um conjunto pertencente ao espaço euclidiano real n-dimensional, $\mathbb{Y}$ um conjunto de *"etiquetas"* e $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$ o produto cartesiano destes dois conjuntos.

Dado um conjunto de treino constituído por m exemplos *separáveis,* constituídos por pares ordenados de vectores de entrada e de saída (ou etiquetas) :

$$(\mathbf{x}_i, y_i) \in \mathbb{Z},\ i = 1, ..., m$$

Existe um vector $\mathbf{w}$ e um escalar b tal que:

$$\mathbf{w} = (w_1, w_2, ..., w_n) \in \mathbf{R}^n; \qquad b \in \mathbf{R}$$

e com os quais se pode definir uma família de hiperplanos:

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0, \tag{3.1}$$

e outra de funções de decisão:

$$h(\mathbf{x}) = sgn(\langle \mathbf{w}, \mathbf{x} \rangle + b). \tag{3.2}$$

O conjunto dos vectores de treino $(\mathbf{x}_i, y_i)$ diz-se *optimamente separado* pelo hiperplano se estiver separado sem cometer qualquer erro de classificação e se for máxima

a distância entre o hiperplano os exemplos de ambas as classes mais próximos. Uma vez que se trata de um problema separável existe um vector peso $\mathbf{w}$ e um escalar b tal que:

$$y_i \ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \ \geq 0, \quad i = 1, \ldots, m$$

Pode considerar-se uma restrição ao caso dos chamados hiperplanos *canónicos*, através de um re-escalonamento de $\mathbf{w}$ e b, tal que os pontos mais próximos do hiperplano satisfazem:

$$|\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1$$

Logo, um hiperplano separador canónico satisfaz a restrição:

$$y_i \ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \ \geq 1, \ i = 1, \ldots, m. \tag{3.3}$$

Se a distância de um exemplo $\mathbf{x}_i$ ao hiperplano $(\mathbf{w},\mathbf{b})$ for:

$$d\left(\mathbf{x}_i, (\mathbf{w},\mathbf{b})\right) = \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|},$$

então, o hiperplano óptimo obtém-se maximizando a margem γ:

$$\begin{aligned}
\gamma &= \min_{\mathbf{x}_i^-} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} + \min_{\mathbf{x}_i^+} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} = \\
&= \frac{1}{\|\mathbf{w}\|} \left(\min_{\mathbf{x}_i^-} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| + \min_{\mathbf{x}_i^+} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \right) = \\
&= \frac{2}{\|\mathbf{w}\|}
\end{aligned}$$

Tem que existir um hiperplano:

$$\langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = 0$$

para o qual, em primeiro lugar, a margem de separação entre as duas classes $\{-1, +1\}$ seja máxima, isto é, que satisfaça a condicionante:

$$\max_{\mathbf{w},b} \left(\min_i \left(\|\mathbf{x} - \mathbf{x}_i\| \ | \ \mathbf{x} \in \mathbb{X}, \ \langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = 0, \ i = 1, \ldots, m \right) \right).$$

e, em segundo lugar, a capacidade decresça quando a margem cresce.

A construção deste hiperplano foi sugerida Vapnik (1998) através da solução de um problema de optimização, no qual o hiperplano óptimo que separa os exemplos deverá satisfazer uma restrição: a minimização de uma função quadrática, na qual a constante 1/2 se destina a simplificar os cálculos e não tem propriamente um

significado particular:

$$\tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.4}$$

Esta restrição é independente de b em virtude de se verificar a equação (3.3), e havendo uma mudança em b apenas faz com que o hiperplano separador sofra uma translação paralelamente a si mesmo, a margem mantém-se constante, apenas o hiperplano deixa de ser óptimo porque fica mais próximo de uma das classes.

A implementação do princípio da minimização estrutural do risco, referido no Capítulo 2, pode verificar-se quando se minimiza o funcional (3.4). Supondo que existe um majorante A tal que $\|\mathbf{w}\| \leq A$ então:

$$d(\mathbf{x}, (\mathbf{w}, b)) \geq \frac{1}{A},$$

resultando que o número de hiperplanos possíveis passa a ser mais reduzido, uma vez que estes só podem existir nos interstícios das regiões definidas pelos círculos centrados nos pontos do conjunto de treino cujo raio é $\frac{1}{A}$. Esta diminuição da capacidade da classe dos hiperplanos pode exprimir-se em termos de d_{VC} e do raio da hiperesfera , R, que contém todos os pontos do conjunto de treino no espaço de entradas n-dimensional:

$$d_{VC} \leq \min\left(R^2 A^2, n\right) + 1$$

Portanto, ao minimizar-se $\|\mathbf{w}\| \leq A$ está-se a minimizar d_{VC}, logo, para o hiperplano óptimo verifica-se um mínimo da capacidade do conjunto das funções de $\mathbb{H}$, expressa pela dimensão VC, e esta capacidade decresce sempre que a margem cresce.

Este problema de optimização com restrições admite a seguinte formulação nas variáveis $\mathbf{w}$ e b:

$$\textbf{Minimizar} \qquad \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{3.5}$$

$$\textbf{Sujeito a} \qquad y_i\ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\ \geq 1,\ i = 1, \ldots, m \tag{3.6}$$

Esta minimização permite determinar um ponto de estacionaridade da função $\tau(\mathbf{w})$ (3.5), isto é, da função quadrática cujo argumento é a norma de $\mathbf{w}$. A restrição (3.6) assegura que a função de decisão $h(\mathbf{x}_i) = +1$ quando $y_i = +1$, e que $h(\mathbf{x}_i) = -1$, quando $y_i = -1$. A razão de se escrever na condicionante " ≥ 1", tem a ver com o facto de assim se fixar efectivamente o vector $\mathbf{w}$, não o deixando depender de uma qualquer transformação de escala $0 < \lambda < 1$, que podendo satisfazer a condicionante " > 0", no entanto, não minimizaria $\tau(\mathbf{w})$.

Note-se que $y_i\ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) / \|\mathbf{w}\|$ é igual à distância do exemplo $\mathbf{x}_i$ ao hiperplano, logo, se $\|\mathbf{w}\| = 1$, então, o membro esquerdo da expressão (3.6) seria igual

a esta distância. Então, se for possível satisfazer a restrição (3.6) para todo o $i = 1, ..., m$, com um vector $\mathbf{w}$ cujo comprimento seja mínimo maximiza-se a margem. A função $\tau(\mathbf{w})$ (3.5) designa-se *função objectivo,* enquanto que a expressão (3.6) designa-se *restrição da desigualdade*, e, conjuntamente, estas duas condicionantes definem um *problema de optimização com restrições.*

Este problema pode ser reduzido a uma *minimização* da Lagrangeana primal Marques_de_Sá (2001a):

Problema 44 (Primal)

$$\textbf{\textit{Minimizar}} \quad L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i \left(\langle \mathbf{x}_i, \mathbf{w} \rangle + b \right) - 1 \right) \tag{3.7}$$

$$\textbf{\textit{Sujeito a}} \quad \alpha_i > 0 \ , \ y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right) \geq 1, \qquad i = 1, \ldots, m \tag{3.8}$$

Os parâmetros α_i são os chamados *multiplicadores de Lagrange.* Determinando as derivadas parciais da expressão $L(\mathbf{w}, b, \alpha)$ em ordem a $\mathbf{w}$ e a b, e substituindo-as nesta mesma fórmula primal obtém-se a Lagrangeana dual de Wolfe, a qual tem de ser *maximizada* com respeito aos α_i:

Problema 45 (Dual)

$$\textbf{\textit{Maximizar}} \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \, \alpha_j \, y_i \, y_j \, \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{3.9}$$

$$\textbf{\textit{Sujeito a}} \quad \alpha_i > 0 \ , \ e \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

Notar que se a condicionante (3.8) fosse violada, então, o membro esquerdo desta seria negativo e, neste caso, a Lagrangeana $L(\mathbf{w}, b, \alpha)$, dada pela expressão (3.7) tenderia a crescer se os correspondentes valores de α_i crescessem também. Para contrariar este crescimento descontrolado os valores de $\mathbf{w}$ e b teriam de variar. Para impedir que as parcelas do somatório da Lagrangeana se tornem negativas, e arbitrariamente grandes, as mudanças a introduzir nos valores de $\mathbf{w}$ e b teriam de assegurar que, se o problema for separável, a condicionante fosse eventualmente satisfeita. Da mesma forma se pode perceber que para todos os exemplos para os quais a condicionante não se reduz à igualdade a zero, os α_i correspondentes têm de ser todos nulos.

As condições de anulamento das derivadas parciais de $L(\mathbf{w}, b, \alpha)$ com respeito às

variáveis primais $\mathbf{w}$ e b implicam que se verifique:

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.10}$$

$$\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{3.11}$$

isto é, o vector solução $\mathbf{w}^*$ exprime-se pelo somatório de um subconjunto de exemplos de treino para os quais os multiplicadores de Lagrange α_i são diferentes de zero aos quais se chama *vectores de suporte.*

As condições de complementaridade de Karush-Kuhn-Tucker da teoria da optimização:

$$\alpha_i \left[y_i \left(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^* \right) - 1 \right] = 0, \quad i = 1, ..., m \tag{3.12}$$

exprimem que todos os vectores de suporte cujos multiplicadores de Lagrange não sejam nulos *(exemplos com maior influência)* devem pertencer às margens de separação, e que os restantes vectores não desempenham qualquer papel na optimização podendo ser considerados como redundantes.

A função $h(\mathbf{x})$ óptima é pois:

$$h(\mathbf{x}) = sgn\left(\sum_{i=1}^{m} y_i . \alpha_i . \langle \mathbf{x}, \mathbf{x}_i \rangle + b^* \right). \tag{3.13}$$

em que b^* se obtém a partir da expressão de Kuhn-Tucker (3.12):

$$b^* = -\frac{1}{2} \langle \mathbf{w}^*, (\mathbf{x}_{-1} + \mathbf{x}_{+1}) \rangle$$

e em que $\mathbf{x}_{-1}$, $\mathbf{x}_{+1}$ são quaisquer vectores de suporte, cujos $\alpha > 0$, e que pertencem à classe -1 e $+1$, respectivamente.

Pode mostrar-se, Scholkopf and Smola (2002) e Schölkopf et al. (1999), que o hiperplano óptimo tem uma boa performance de generalização e é computacionalmente atractivo uma vez que pode ser construído pela resolução de um problema de optimização por *programação quadrática.* Sempre que os dados são linearmente separáveis o número de vectores de suporte tende a ser reduzido, e estes podem ser considerados como constituindo *um resumo toda a informação* contida no conjunto de treino.

Pode mostrar-se Schölkopf et al. (1999) que se verificam as seguintes relações:

$$\|\mathbf{w}^*\|^2 = \sum_{VS} \alpha_i = \sum_{VS} \sum_{VS} \alpha_i \, \alpha_j \, y_i \, y_j \, \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$d_{VC} \leq \min\left(R^2 A^2, n\right) + 1$$

$$d_{VC} \leq \min\left(R^2 \sum_{VS} \alpha_i, n\right) + 1$$

em que $\sum_{VS}$ significa um somatório estendido os vectores de suporte.

Exemplo 46 (***Classificação, 5 pontos, 2 classes***) *Vai-se fazer uma aplicação deste método na determinação do hiperplano óptimo:*

$$\langle [u, v], [x, y] \rangle + b = 0$$

que separa os pontos:

$$\begin{aligned} z_1 &= \{(0,1), -1\}, \\ z_2 &= \{(1,0), -1\}, \\ z_3 &= \{(1,1), -1\}, \\ z_4 &= \{(2,2), 1\}, \\ z_5 &= \{(2,3), 1\} \end{aligned}$$

Escreve-se a expressão geral da Lagrangeana(3.7):

$$L = \frac{1}{2} \|[u, v]\|^2 - \sum_{i=1}^{5} \alpha_i \left(y_i . \left(\langle [u, v], [x, y] \rangle + b\right) - 1\right)$$

Substitui-se os valores das coordenadas dos vectores dados:

$$\begin{aligned} L = \frac{1}{2}u^2 + \frac{1}{2}v^2 \\ -\alpha_1 \left(-1. \left(v + b\right) - 1\right) \\ -\alpha_2 \left(-1. \left(u + b\right) - 1\right) \\ -\alpha_3 \left(-1. \left(u + v + b\right) - 1\right) \\ -\alpha_4 \left(1. \left(2u + 2v + b\right) - 1\right) \\ -\alpha_5 \left(1. \left(2u + 3v + b\right) - 1\right) \end{aligned} \tag{3.14}$$

A Lagrangeana dever ser minimizada *relativamente a* $\mathbf{w} = [u, v]$ *e b:*

$$\begin{aligned} \frac{\partial L}{\partial b} &= \alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 = 0 \\ \frac{\partial L}{\partial u} &= u + \alpha_2 + \alpha_3 - 2\alpha_4 - 2\alpha_5 = 0 \end{aligned} \tag{3.15}$$

$$\frac{\partial L}{\partial v} = v + \alpha_1 + \alpha_3 - 2\alpha_4 - 3\alpha_5 = 0$$

Resolvendo a expressão (3.14) em termos de u, v *e* b :

$$\begin{aligned} L = \frac{1}{2}u^2 + \frac{1}{2}v^2 + \alpha_1 (v + b) + \alpha_1 + \alpha_2 (u + b) + \alpha_2 + \alpha_3 (u + v + b) + \alpha_3 \\ -\alpha_4 (2u + 2v + b) + \alpha_4 - \alpha_5 (2u + 3v + b) + \alpha_5 \\ = \frac{1}{2}u^2 + \frac{1}{2}v^2 + (\alpha_2 + \alpha_3 - 2\alpha_4 - 2\alpha_5) u + (\alpha_1 + \alpha_3 - 2\alpha_4 - 3\alpha_5) v \\ + (\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5) b + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5; \end{aligned}$$

Usando as equações (3.15)

$$\begin{aligned} L = \frac{1}{2}u^2 + \frac{1}{2}v^2 - u^2 - v^2 + 0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5; \\ = -\frac{1}{2}(-\alpha_2 - \alpha_3 + 2\alpha_4 + 2\alpha_5)^2 - \frac{1}{2}(-\alpha_1 - \alpha_3 + 2\alpha_4 + 3\alpha_5)^2 \\ + (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5); \end{aligned}$$

A Lagrangeana dever ser maximizada *relativamente a* $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$*:*

$$\begin{aligned} \frac{\partial L}{\partial \alpha_1} = -\alpha_1 - \alpha_3 + 2\alpha_4 + 3\alpha_5 + 1 = 0 \\ \frac{\partial L}{\partial \alpha_2} = -\alpha_2 - \alpha_3 + 2\alpha_4 + 2\alpha_5 + 1 = 0 \\ \frac{\partial L}{\partial \alpha_3} = (-\alpha_2 - \alpha_3 + 2\alpha_4 + 2\alpha_5) + (-\alpha_1 - \alpha_3 + 2\alpha_4 + 3\alpha_5) + 1 = 0 \\ \frac{\partial L}{\partial \alpha_4} = -2(-\alpha_2 - \alpha_3 + 2\alpha_4 + 2\alpha_5) - 2(-\alpha_1 - \alpha_3 + 2\alpha_4 + 3\alpha_5) + 1 = 0 \\ \frac{\partial L}{\partial \alpha_5} = -2(-\alpha_2 - \alpha_3 + 2\alpha_4 + 2\alpha_5) - 3(-\alpha_1 - \alpha_3 + 2\alpha_4 + 3\alpha_5) + 1 = 0 \end{aligned}$$

Obtém-se o seguinte sistema de equações:

$$\mathbf{H}\,\alpha = \mathbf{1}$$

numericamente igual a:

$$\begin{pmatrix} 1 & 0 & 1 & -2 & -3 \\ 0 & 1 & 1 & -2 & -2 \\ 1 & 1 & 2 & -4 & -5 \\ -2 & -2 & -4 & 8 & 10 \\ -3 & -2 & -5 & 10 & 13 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Notar, porém, que este sistema não tem solução porque a matriz do kernel $\mathbf{H}$*, dos*

coeficientes, é singular: $\det(\mathbf{H}) = 0$. *Para os problemas quadráticos com esta estrutura* $\mathbf{H}$ *é sempre singular. Raciocinando geometricamente vê-se que deve existir uma solução. Não se pode esquecer a existência das restrições :*

$$\alpha_i \geq 0,\ i = 1 \ldots 5$$

. Assim como a restrição (3.15), isto é,

$$\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 = 0.$$

*A solução pode obter-se recorrendo a um algoritmo apropriado Fletcher (2000), Luenberger (1989), Mangasarian (1994). Usou-se a função do Matlab "**qp**" para programação quadrática, que requer as seguintes matrizes de entrada:*

$$H = \begin{pmatrix} 1.0000 & 0.0000 & 1.0000 & -2.0000 & -3.0000 \\ 0.0000 & 1.0000 & 1.0000 & -2.0000 & -2.0000 \\ 1.0000 & 1.0000 & 2.0000 & -4.0000 & -5.0000 \\ -2.0000 & -2.0000 & -4.0000 & 8.0000 & 10.0000 \\ -3.0000 & -2.0000 & -5.0000 & 10.0000 & 13.0000 \end{pmatrix}$$

$$c = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 \end{pmatrix}^T$$

$$A = \begin{pmatrix} -1 & -1 & -1 & 1 & 1 \end{pmatrix}$$

$$b = (0)$$

$$LimInfAlfa = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T$$

$$LimSupAlfa = \begin{pmatrix} Inf & Inf & Inf & Inf & Inf \end{pmatrix}^T$$

$$x0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T$$

$$numRstIgual = 1$$

$$[alfa] = \mathbf{qp}(H,\ c,\ A,\ b,\ LimInfAlfa,\ LimSupAlfa,\ x0,\ numRstIgual);$$

A solução devolvida por esta função do Matlab é:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1.0 \\ 1.0 \\ 0 \end{pmatrix}$$

$$w = \begin{pmatrix} 1.0000 & 1.0000 \end{pmatrix};\ b = -3.0000$$

Hiperplano:

$$\left\langle \begin{pmatrix} 1 & 1 \end{pmatrix}, \begin{pmatrix} x_1 & x_2 \end{pmatrix} \right\rangle - 3 = 0;$$

$$x_1 + x_2 - 3 = 0$$

$|w|^2$: 2.0

Margem: $\sqrt{2.0} = 1.4142$

Soma dos alfa: 2.0000

Número Vectores de Suporte: 2 (40.0%)

Vectores de Suporte: $z_3 = \{(1,1), -1\}$, $z_4 = \{(2,2), 1\}$

$d_{VC} \leq \min\left(R^2 \sum_{VS} \alpha_i, n\right) + 1 = \min\left(2R^2, 2\right) + 1 = 3$ *uma vez que* $R > 1$ *(ver fig. 3.1)*

Verificação:

$$\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5 = 0 + 0 + 1 - 1 + 0 = 0$$

$$\begin{pmatrix} 1 & 0 & 1 & -2 & -3 \\ 0 & 1 & 1 & -2 & -2 \\ 1 & 1 & 2 & -4 & -5 \\ -2 & -2 & -4 & 8 & 10 \\ -3 & -2 & -5 & 10 & 13 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ -2 \\ 4 \\ 5 \end{pmatrix}$$

Não se obtem no membro da direita um vector coluna de 1*s uma vez que, como se viu, o sistema era singular, logo, a programação quadrática limita-se a determinar os coeficientes* α_i *que permitem uma melhor aproximação do segundo membro.*

Representação gráfica:

$$z_1 = \{(0,1), -1\},$$
$$z_2 = \{(1,0), -1\},$$
$$z_3 = \{(1,1), -1\},$$
$$z_4 = \{(2,2), 1\},$$
$$z_5 = \{(2,3), 1\};$$
$$x_2 = -x_1 + 3$$

Notas:

(1) *O suporte algébrico Maple (7.0) associado ao processador do presente texto só dispõe de facilidades para optimização linear pelo método* simplex, *razão pelo*

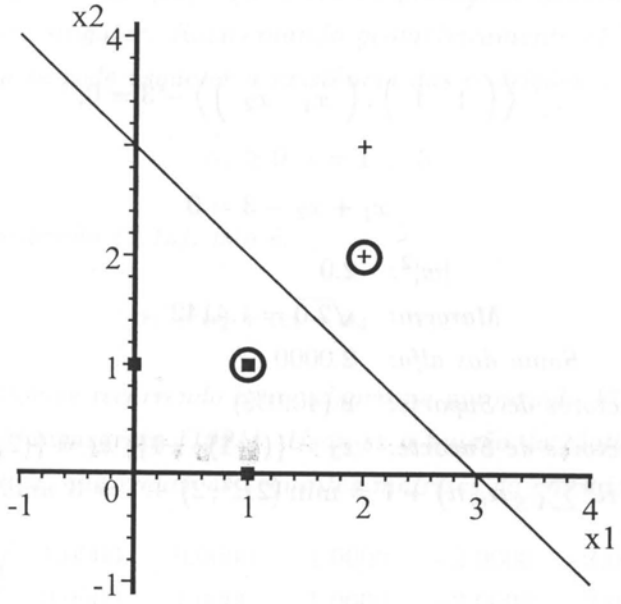


Figura 3.1: Exemplo 46: a recta a contínuo separa os quadrados das cruzes, os círculos assinalam os vectores de suporte.

qual se recorreu ao MATLAB. Este problema de optimização também pode ser resolvido por programação linear (Campbell (2000)), porém, como não foi assim que historicamente ele foi resolvido por razões de espaço omite-se esta alternativa.

(2) *A função do Matlab* **qp** *está já obsoleta. A versão do Matlab 6.0-Release-12, tem disponível nas suas ferramentas de optimização uma versão mais actualizada* **quadprog**. *Porém, para este problema muito simples é indiferente a versão utilizada, desde que se respeitem as definições dos parâmetros de entrada.*

(3) *Pode ser necessário adicionar alguma regularização de ordem zero no caso de existir um mau condicionamento da matriz do* kernel.

(4) *Num problema de classificação está-se à procura duma função que nos dê um valor superior a uma dada constante ε (no caso de hiperplanos canónicos $\varepsilon = 1$) para os exemplos positivos e um valor inferior a $-\varepsilon$ para os exemplos negativos. Os pontos onde a função toma os valores $\pm\varepsilon$ definem a ε-margem no espaço dos exemplos e da qual se vai falar extensivamente, já a seguir, na estimação por* regressão. *O que é importante aqui salientar é que quanto maior for a proximidade do hiperplano canónico (recta discriminante) em relação à normal ao segmento que une os dois pontos mais próximos de cada uma das envoltórias convexas dos dois conjuntos, tanto mais larga é a margem, e esta é a tarefa que a MVS é capaz de realizar com sucesso*

3.1.3 Classificação de padrões não separáveis linearmente

Verificando-se numa grande parte dos conjuntos de treino a existência de ruído se a MVS se sobreajustar em demasia a este ruído passa a ter uma capacidade de generalização muito pobre. O algoritmo descrito na secção anterior quando aplicado a dados não separáveis não origina nenhuma solução viável, facto que é evidenciado pelo crescimento arbitrário da função objectivo, ou do dual Lagrangeano.

Anteriormente reduziu-se o efeito dos valores extremos através da introdução da margem. Presentemente, vai-se lidar com o caso não-separável, relaxando a restrição (3.3) através da introdução de um novo custo na função objectivo primal, o qual pode ser feito com variáveis de folga. Este método de "margem suave" (*soft margin*) que se baseia na norma do erro L foi introduzido por Cortes and Vapnik (1995) e constitui um entre outros esquemas existentes na teoria das MVS para lidar com o efeito do ruído.

Continuando a supor, para simplificar, que se espera que um hiperplano possa correctamente separar os dados, introduz-se as variáveis de "folga" não-negativas $\xi_i \geq 0$, que funcionam como uma espécie de medida do grau de erro de classificação, assim como uma função de penalização $F(\xi) = \sum_{i=1}^{m} \xi_i$.

A restrição (3.3) passa a ser:

$$y_i \ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \ \geq 1 - \xi_i \ , \quad i = 1, \ldots, m$$

e a função objectivo $\tau(\mathbf{w})$ a minimizar passa a ser constituída por duas parcelas :

$$\tau(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i$$

Pode reformular-se o problema da optimização na seguinte função objectivo:

Problema 47 (Primal não-separável)

$$\textit{Minimizar} \quad L(\mathbf{w}, b, \alpha, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \, (\, y_i . (\, \langle \mathbf{x}_i, \mathbf{w} \rangle + b) - 1 + \xi_i) - r_i \sum_{i=1}^{m} \xi_i$$

$$\textit{Sujeito a} \quad \alpha_i \geq 0 \quad e \ y_i \ (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \ \geq 1 - \xi_i \ , \quad i = 1, \ldots, m$$

As derivadas parciais com respeito a $\mathbf{w}$, b *e* ξ *são:*

$$\frac{\partial L}{\partial w_\nu} = w_\nu - \sum_{i=1}^{m} \alpha_i y_i x_{i\nu} = 0,$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - r_i = 0.$$

Substituindo-se estes resultados na função objectivo primal obtém-se a mesma função objectivo (3.9)

Problema 48 (Dual não-separável)

$$\textit{Maximizar} \qquad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\textit{Sujeito a} \qquad 0 \leq \alpha_i \leq C$$

Verificando-se $r_i > 0$ e $C - \alpha_i - r_i = 0$, logo $\alpha_i \leq C$, obtendo-se a restrição $0 \leq \alpha_i \leq C$. Os exemplos cujos valores verificam $0 < \alpha_i < C$ são designados *não-limitados* e aqueles para os quais $\alpha_i = 0$, ou $\alpha_i = C$, são designados como *nos-limites*.O valor de b da função de decisão é determinado pela expressão $b = y_i - \sum_j \alpha_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, supondo que $y_i = \pm 1$. O valor óptimo de C tem que ser determinado empiricamente através dum conjunto de validação, parecendo que não está muito relacionado com as características do conjunto de treino, apesar de haver quem admita que tem um papel de parâmetro de regularização e que reflecte o conhecimento que se tem acerca do ruído nos dados (Cristianini and Shawe-Taylor (2000), Burges (1998) and Cortes and Vapnik (1995)).

3.1.4 Generalização num espaço de características de dimensionalidade elevada

Os algoritmos de construção das máquinas de vectores de suporte não se limitam à determinação de um hiperplano óptimo no espaço de entradas $\mathbb{X}$ porque muitas vezes tal não é possível. O princípio básico que preside aos algoritmos das MVS consiste na possibilidade de transformação deste mesmo espaço de entradas, através duma funcional $\Phi(\mathbb{X})$, para um outro espaço que se designa *espaço de características (feature space)* tal como mostra a figura (3.2). Esta designação que não se deve confundir com "características de entrada" utilizada na teoria do reconhecimento de padrões para designar as variáveis de entrada.

O vector de pesos (3.11) passa a ter uma expansão dada por um somatório num espaço de características e, em geral, deixará de corresponder à imagem de um único vector no espaço de entradas.

x	y
$(-1, -1)$	-1
$(-1, 1)$	1
$(1, -1)$	1
$(1, 1)$	-1

Tabela 3.1: Conjunto de treino no espaço de entradas

Uma vez que todos os exemplos aparecem como produtos internos podem substituir-se pelo *kernel* de Mercer k obtendo-se, assim, funções de decisão mais gerais:

$$h(\mathbf{x}) = sign\left(\sum_{i=1}^{m} y_i \alpha_i \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i)\rangle + b\right) \tag{3.16}$$

$$= sign\left(\sum_{i=1}^{m} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{3.17}$$

as quais, por sua vez, originam programas quadráticos. Por exemplo, no caso separável:

Problema 49 (caso separável com kernel)

$$\textbf{\textit{Maximizar}} \qquad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j \, y_i y_j \, k(\mathbf{x}_i, \mathbf{x}_j) \tag{3.18}$$

$$\textbf{\textit{Sujeito a}} \qquad \alpha_i \geq 0, \quad i = 1, \ldots, m, \quad e \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3.19}$$

Existem algumas restrições quanto ao tipo de transformações Φ e correspondentes funções *kernel* $k(\mathbf{x}_i, \mathbf{x}_j)$ que se podem usar. Por exemplo, se o *kernel* for polinomial quadrático:

$$k(\mathbf{x}_i, \mathbf{x}_j) = [\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1]^2$$

realiza uma aplicação linear de um vector do espaço bidimensional num espaço de características de dimensão seis

Exemplo 50 *Consideremos um conjunto de treino constituído por quatro pares de entradas e as respectivas etiquetas dadas pela conhecida função binária* XOR *representadas pela Tabela 3.1 A superfície de decisão neste espaço de entradas é não-linear. Pode obter-se um hiperplano de decisão efectuando uma transformação* Φ *muito simples, que consiste em acrescentar ao espaço de entradas mais uma componente, que se obtém do produto de outras duas,* $x_1 x_2$. *Constrói-se assim um* espaço

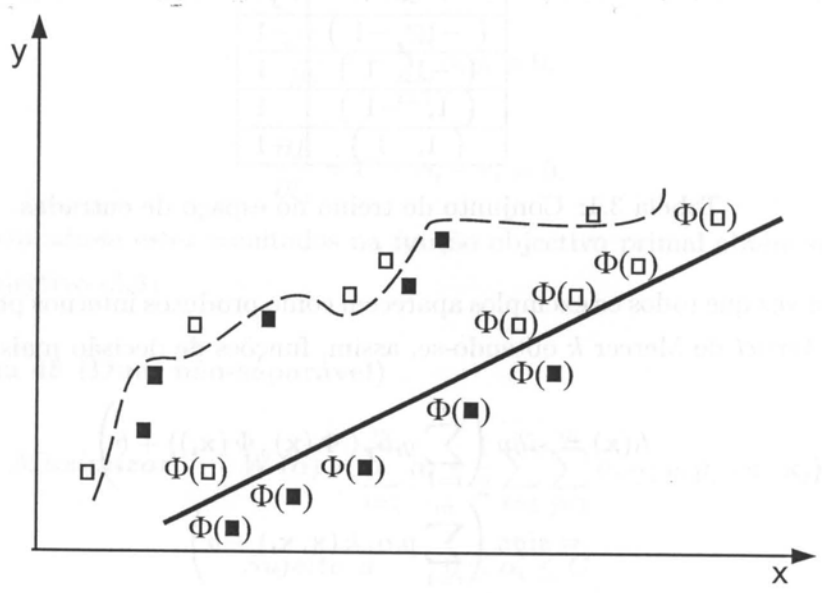


Figura 3.2: A idea das MVS é uma transformação Φ aos dados de treino mergulhando-os num *espaço de características* em que se pode determinar um hiperplano separador de margem máxima.

$\Phi(\mathbf{x})$	y
$(-1, -1, 1)$	-1
$(-1, 1, -1)$	1
$(1, -1, -1)$	1
$(1, 1, 1)$	-1

Tabela 3.2: Conjunto de treino no espaço de entradas

de características *tridimensional:*

$$\Phi : \mathbf{R}^2 \to \mathbf{R}^3$$
$$(x_1, x_2) \mapsto (x_1, x_2, x_1 x_2)$$

Os exemplos neste novo espaço estão representados na Tabela 3.2: Os novos exemplos que resultam desta transformação são agora linearmente separáveis podendo obter-se uma superfície de decisão óptima $g(x_1, x_2, x_3) = 0$, *em que* $x_3 = x_1 x_2$ *dada por:*

$$g(x_1, x_2) = (0, 0, -1)^T (x_1, x_2, x_1 x_2)$$

logo, no espaço de características os exemplos de treino são separados pela função linear:

$$g(x_1, x_2) = -x_1 x_2$$

enquanto que no espaço de entradas eles são separados pela função não-linear:

$$y = sgn\left(-x_1 x_2\right)$$

Vamos, seguidamente, aplicar o "truque" *do* kernel *para resolver este mesmo problema. Em lugar de se trabalhar no espaço de características usa-se um* kernel *apropriado, que satisfaça certas condições, e resolve-se um problema de programação quadrática dual para obter a superfície de decisão. Vamos usar, por exemplo, um* kernel *polinomial de ordem dois, que como se sabe satisfaz as condições de Mercer:*

$$K\left(\mathbf{u}, \mathbf{v}\right) = \left(1 + \langle \mathbf{u}, \mathbf{v} \rangle\right)^2 \tag{3.20}$$

Determinam-se os elementos $\mathbf{H}_{ij}$, *da linha* i *coluna* j *da matriz do* kernel $\mathbf{H}$, $i, j = 1, \ldots m$:

$$\mathbf{H}_{ij} = y_i\, y_j\, K\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

Numericamente, obtém-se:

$$\mathbf{H} = \begin{bmatrix} 9.0000 & -1.0000 & -1.0000 & 1.0000 \\ -1.0000 & 9.0000 & 1.0000 & -1.0000 \\ -1.0000 & 1.0000 & 9.0000 & -1.0000 \\ 1.0000 & -1.0000 & -1.0000 & 9.0000 \end{bmatrix}$$

Pode agora fazer-se a formulação do problema *de programação quadrática dual:*

$$\begin{aligned} \min \quad & \alpha^T \mathbf{H}\, \alpha - \mathbf{1}^T \alpha \\ s.a. \quad & \mathbf{y}^T \alpha = \mathbf{0} \\ & \mathbf{0} \leq \alpha \leq \mathbf{1} \end{aligned}$$

Usou-se a função do Matlab ***qp*** *tendo-se obtido a seguinte solução:*

$$\alpha = \begin{bmatrix} 0.125 & 0.125 & 0.125 & 0.125 \end{bmatrix}^T$$

A função discriminante $g\left(\mathbf{x}\right)$ *que permite classificar um novo vector* $\mathbf{x}$ *é pois constituída por uma primeira parcela que calcula a combinação linear dos alfas e do* kernel *do novo vector que se quer classificar* $\mathbf{x}$, *em que* $[\mathbf{x}]_1, [\mathbf{x}]_2 \neq 0$, *mais uma segunda parcela que calcula o valor do termo independente (e que neste exemplo tem um valor nulo):*

$$g\left(\mathbf{x}\right) = \sum_{i=1}^{4} \alpha_i\, y_i\, K\left(\mathbf{x}, \mathbf{x}_i\right) + \left(y_1 - \sum_{i=1}^{4} \alpha_i\, y_i\, K\left(\mathbf{x}_1, \mathbf{x}_i\right) \right)$$

Desenvolvendo esta expressão segundo as componentes do novo vector $\mathbf{x} = [x_1\ x_2]$:

$$g([x_1,\ x_2]) = 0.125\left[-(1+x_1+x_2)^2 + (1-x_1+x_2)^2\right.$$

$$\left. +(1+x_1-x_2)^2 - (1-x_1-x_2)^2\right]$$

Seja, por exemplo, $\mathbf{x} = [0.01, -0.01]$, *obtém-se*

$$classe(\mathbf{x}) = sign\{g([0.01, -0.01])\} = 1.$$

Notar que no espaço de características associado ao kernel *(3.20)*

$$K(\mathbf{u},\mathbf{v}) = (1+\langle \mathbf{u},\mathbf{v}\rangle)^2 = (1+u_1v_1+u_2v_2)^2 =$$
$$= 1 + 2v_1u_1 + 2v_2u_2 + 2v_1v_2u_1u_2 + v_1^2u_1^2 + v_2^2u_2^2$$

pode definir-se a seguinte transformação $\Phi(\mathbf{u})$*:*

$$\Phi([u_1,\ u_2]) = \left[1,\ \sqrt{2}u_1,\ \sqrt{2}u_2,\ \sqrt{2}u_1u_2,\ u_1^2,\ u_2^2\right]^T$$

O vector de pesos óptimo:

$$\begin{aligned}\mathbf{w}^* &= 0.125\left\{\left[-1,\sqrt{2},\sqrt{2},-\sqrt{2},-1,-1\right] + \left[1,-\sqrt{2},\sqrt{2},-\sqrt{2},1,1\right] + \right.\\ &\left. + \left[1,\sqrt{2},-\sqrt{2},-\sqrt{2},1,1\right] + \left[-1,-\sqrt{2},-\sqrt{2},-\sqrt{2},-1,-1\right]\right\}\\ &= \left[0,0,\left(\frac{\sqrt{2}+\sqrt{2}-\sqrt{2}-\sqrt{2}}{2^3}\right),\left(\frac{-\sqrt{2}-\sqrt{2}-\sqrt{2}-\sqrt{2}}{2^3}\right),0,0\right]\\ &= \left[0,0,0,\frac{-1}{\sqrt{2}},0,0\right]\end{aligned}$$

Finalmente, a superfície de decisão não linear *óptima:*

$$\langle \mathbf{w}^*, \Phi(\mathbf{u})\rangle + b = 0$$

$$\left[0,0,0,\frac{-1}{\sqrt{2}},0,0\right]\left[1,\ \sqrt{2}u_1,\ \sqrt{2}u_2,\ \sqrt{2}u_1u_2,\ u_1^2,\ u_2^2\right]^T + 0 = -u_1u_2$$

$$g(\mathbf{u}) = sign(-u_1u_2)$$

A transformação para um espaço de características apresenta vantagens e desvantagens. Uma primeira vantagem, caso as características tenham sido bem escolhidas, consiste em permitir no novo espaço uma separação linear das classes quando no espaço de entradas original se apresentava uma relação complexa e não linear. Da mesma forma verifica-se que o número total de características não precisa de

ser muito superior ao número de características do espaço de entradas. quanto a desvantagens há a referir que a escolha das características mais apropriadas é feita heuristicamente, por tentativa e erro, e nem sempre é evidente qual o número que se deve experimentar.

A formulação de problemas de programação quadrática com transformações características $\Phi(\mathbf{x})$ em $\mathbf{R}^k$ tem a seguinte expressão no caso linearmente não-separável em que se usam variáveis de folga ξ :

Problema 51 (não-separável, c/ transf. de características)

$$\underset{\mathbf{w}\in\mathbf{R}^k,\ b\in\mathbf{R},\ \xi\in\mathbf{R}^m}{\textit{Minimizar}} \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{1}^T\xi$$

$$\textit{Sujeito a} \quad y_i\left(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b\right) \geq 1 - \xi_i$$
$$i = 1, \ldots, m$$

A classificação de um novo exemplo $\mathbf{u}$ do qual se desconhece a etiqueta é dado pela expressão:

$$y(\mathbf{u}) = sign\left(\langle \mathbf{w}, \Phi(\mathbf{u})\rangle + b\right).$$

O "truque" do *kernel* apresenta a vantagem de permitir utilizar automaticamente as características sem ter que as especificar após a transformação. Utiliza-se a transformação dada por uma função *kernel* $K(\mathbf{u}, \mathbf{v})$,tal que $K : \mathbf{R}^n \times \mathbf{R}^n \longrightarrow \mathbf{R}$, que aplica dois vectores num único número real. Este obtém-se substituindo as componentes dos vectores numa dada expressão, como por exemplo:

$$K(u_1, u_2, v_1, v_2) = (1 + u_1v_1 + u_2v_2)^2.$$

Define-se uma matriz $n \times n$, $\mathbf{H}$, no espaço dual da programação quadrática:

$$\mathbf{H}_{ij} = y_i\, y_j\, K(\mathbf{x}_i, \mathbf{x}_j), \quad i, j = 1, \ldots m.$$

Verificando-se a condição de Mercer:

$$K(\mathbf{u}, \mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v})\rangle$$

para uma dada transformação no espaço de características Φ, então, o dual da programação quadrática é um programa convexo que permite obter uma função discriminante dada pela expressão:

$$g(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x})\rangle + b \ \textit{ em que } \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \Phi(\mathbf{x}_i) \tag{3.21}$$

1	Polinomial	$\left(1+\mathbf{x}^T\mathbf{y}\right)^p$
2	Perceptrão	$tanh\left(\beta_0\mathbf{x}^T\mathbf{y}+\beta_1\right)$
3	Exponencial	$e^{\gamma\mathbf{x}^T\mathbf{y}}-1$
4	Gaussiano RBF	$e^{\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|^2}$
5	Exponencial RBF	$e^{\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|}$
6	Spline	$\sum_{r=0}^{k} x^r y^r + \sum_{s=0}^{m} (x-\tau_s)_+^k (y-\tau_s)_+^k$
7	Mistura	$\left(1+\mathbf{x}^T\mathbf{y}\right)^p e^{\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{y}\|}$
8	Dirichlet $(1-d)$	$\frac{\sin\left[\left(m+\frac{1}{2}\right)(\mathbf{x}-\mathbf{y})\right]}{2\sin\left[\frac{1}{2}(\mathbf{x}-\mathbf{y})\right]}$

Tabela 3.3: Exemplos de funções kernel

Notar que para o *kernel* verificar a igualdade:

$$K(\mathbf{u},\mathbf{v}) = \langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle$$

a condição de Mercer estipula que o *kernel* tem que ser:

(i) simétrico: $K(\mathbf{u},\mathbf{v}) = K(\mathbf{v},\mathbf{u})$;

(ii) integrável para todas as funções f, tal que f^2 é integrável no quadrado $[a,b]^2$, e se verifica ser positivo neste quadrado o integral $\iint K(\mathbf{u},\mathbf{v})\, f(\mathbf{u})\, f(\mathbf{v})\, d\mathbf{u}\, d\mathbf{v} \geq \mathbf{0}$.

Sendo conhecido o *kernel* $K(\mathbf{u},\mathbf{v})$, e o vector óptimo dos multiplicadores de Lagrange α obtidos como solução da programação quadrática do dual, então, pode facilmente escrever-se, sem ter que determinar explicitamente $\Phi(\mathbf{x})$ a partir da expressão (3.21), a função discriminante $g(\mathbf{x})$:

$$\begin{aligned} g(\mathbf{x}) &= \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b \\ &= \sum_{i=1}^{m} \alpha_i y_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle + b \\ &= \sum_{i=1}^{4} \alpha_i\, y_i\, K(\mathbf{x},\mathbf{x}_i) + b \end{aligned}$$

em que o termo independente se pode determinar para qualquer j tal que $0 < \alpha_j < C$, pela expressão:

$$b = y_j - \sum_{i=1}^{m} \alpha_i\, y_i\, K(\mathbf{x}_j,\mathbf{x}_i)$$

A tabela (3.3) mostra alguns exemplos das principais funções *kernel* usadas em MVS e que satisfazem a condição de Mercer. O kernel (7) também é conhecido por *kernel* "*série de Fourier*".

3.1.5 Regressão linear com função de custo insensitiva

A generalização dos algoritmos das *MVS* classificadoras ao caso da *MVS* regressores, capazes de fazer a estimação de funções arbitrárias de contradomínio real, requer que se encontre uma maneira de reter as características particulares deste problema. No caso da classificação existia um número discreto de valores possíveis no contradomínio da função, agora, passa a existir um número infinito de valores. Primeiramente, introduziu-se (Vapnik (1995)) para simplificar a solução deste problema a função de custo *ε-insensitiva:*

$$|y - f(x)|_\epsilon = \max\{0\,,\, |y - f(x)| - \epsilon\}$$

a qual despenaliza os erros inferiores a um dado valor positivo de ε, escolhido a priori (figura 3.3). Existem razões justificativas desta escolha. No problema da classificação quando se media o custo ocasionado por um dado exemplo existia uma grande zona em que esse custo tinha valor nulo. Sempre que um exemplo estivesse no lado correcto da superfície de decisão, e não tocasse na margem, ele não contribuía com qualquer custo para a função objectivo. Por consequência ele não era portador de qualquer "*informação*" acerca da posição da superfície de decisão, uma vez que esta era determinada pela minimização da própria função objectivo. E por esta razão este exemplo não aparecia na expansão dada pela *MVS* para a solução. Uma função de custo para a estimação por regressão deve também ter uma zona insensitiva, e é por esta razão que se usa a função ε-insensitiva para penalizar os custos.

O desenvolvimento do algoritmo da regressão processa-se de forma bastante análoga ao caso que se viu acima para a classificação. Uma vez mais se faz a estimativa de classes de funções lineares, usa-se $\|\mathbf{w}\|^2$ como regularizador e reescreve-se tudo em termos de produtos internos. Mais à frente faz-se a generalização ao caso não-linear. Determinam-se as funções afins:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \text{ em que } \mathbf{w}, \mathbf{x} \in H \text{ e } b \in \mathbf{R} \tag{3.22}$$

baseadas em exemplos independentes e identicamente distribuídos (iid):

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{H} \times \mathbf{R}$$

O conjunto $\mathbb{H}$ é um espaço de produto interno no qual vivem os transformados dos vectores de entrada. Por outras palavras, $\mathbb{H}$ é o espaço de características induzido por um *kernel*, o qual, na presente exposição se simplifica utilizando-se a própria função produto interno.

O objectivo da aprendizagem é a determinação de uma função f com um valor

baixo para a funcional de risco:

$$R[f] = \int c(f, \mathbf{x}, y)\, dP(\mathbf{x}, y)$$

o que é equivalente a dizer que o erro de teste e o erro amostral devem ser o mais reduzidos possível. Porém, este funcional $R[f]$ não se pode determinar uma vez que se desconhece a distribuição de probabilidades $P(\mathbf{x}, y)$. Portanto, em sua substituição usa-se a minimização da funcional de risco empírico com regularização:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \cdot R_{emp}^{\varepsilon}[f] \tag{3.23}$$

em que;

$$R_{emp}^{\varepsilon}[f] = \frac{1}{m}\sum |y_i - f(x_i)|_{\epsilon}$$

mede o erro de treino ε-insensitivo e C é uma constante que determina um compromisso com o termo $\|\mathbf{w}\|^2$ penalizador da complexidade. Numa palavra, a minimização da expressão (3.23) capta a ideia essencial transmitida pela teoria estatística da aprendizagem, quando esta estabelece que se pode obter um risco pequeno se existir um controlo simultaneamente do erro de treino e da complexidade do modelo, obtido com um modelo simples explicativo dos dados.

Um valor pequeno de $\|\mathbf{w}\|^2$ corresponde a uma função linear (3.22) no espaço de características aproximadamente plana. Este mesmo regularizador foi usado na classificação, contudo, correspondia a uma margem ampla. Pode então colocar-se a questão de saber porque é que a regressão com *VS* requer um parâmetro adicional ε, o qual não foi necessário para as *MVS* classificadoras.

Definição 52 (ε-margem) *Sejam* $(\mathbb{E}, \|.\|_{\mathbb{E}})$, $(\mathbb{G}, \|.\|_{\mathbb{G}})$ *espaços normados, e* $\mathbb{X} \subset \mathbb{E}$. *Define-se* ε-margem *duma função* $f : \mathbb{X} \to \mathbb{G}$ *pela expressão*

$$m_{\varepsilon}(f) = \inf\left\{\|x - x'\|_{\mathbb{E}} \quad \textit{tal que } x, x' \in \mathbb{X}, \quad e \quad \|f(x) - f(x')\|_{\mathbb{G}} \geq 2\varepsilon\right\}$$

A minimização de (3.23) é equivalente à seguinte optimização com restrições:

$$\underset{\mathbf{w}\in\mathbb{H}, \xi^{(*)}\in\mathbf{R}^m, b\in\mathbf{R}}{\textbf{Minimizar}} \quad \tau\left(\mathbf{w}, \xi^{(*)}\right) = \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{3.24}$$

$$\begin{aligned} \textbf{Sujeito a} \quad & (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i \leq \varepsilon + \xi_i, \\ & y_i - (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \leq \varepsilon + \xi_i^*, \\ & \xi_i^{(*)} \geq 0; \quad i = 1, \ldots, m \end{aligned}$$

Escreve-se a Lagrangeana a partir da função objectivo e das restrições introdu-

zindo um par de variáveis duais, o que permite obter uma função que pode ter um ponto de estacionaridade com respeito às variáveis primais e duais na solução. Seja pois a Lagrangeana:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) \tag{3.25}$$
$$- \sum_{i=1}^{m}\alpha_i\left(\varepsilon + \xi_i + y_i - (\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\right)$$
$$- \sum_{i=1}^{m}\alpha_i^*\left(\varepsilon + \xi_i^* - y_i + (\langle \mathbf{w}, \mathbf{x}_i\rangle + b)\right)$$

em que as variáveis duais $\eta_i^{(*)}$, $\alpha_i^{(*)}$têm que obedecer a restrições de positividade:

$$\eta_i^{(*)},\ \alpha_i^{(*)} > 0 \tag{3.26}$$

Segue-se das condições de estacionaridade da teoria da optimização que as derivadas parciais da Lagrangeana L com respeito às variáveis primais ($\mathbf{w}$,b, ξ_i, ξ_i^*) têm de ser nulas:

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \tag{3.27}$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\,\mathbf{x}_i = 0 \tag{3.28}$$

$$\frac{\partial L}{\partial \xi_i^{(*)}} = \frac{C}{m} - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{3.29}$$

Substituindo (3.27), (3.28), e (3.29) na expressão (3.25), obtém-se o seguinte problema de optimização

Problema 53 (Dual da regressão)

$$\textit{Maximizar}\ W(\alpha, \alpha^*) = \sum_{i,j=1}^{m}(\alpha_i^* - \alpha_i)\left(\alpha_j^* - \alpha_j\right)\langle \mathbf{x}_i, \mathbf{x}_j\rangle - \tag{3.30}$$
$$- \varepsilon\sum_{i=1}^{m}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i(\alpha_i^* - \alpha_i)$$

$$\textit{Sujeito a}\ \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \quad \text{e}\ \ \alpha_i, \alpha_i^* \in [0, C/m]\ \ \text{e}\ \ i = 1, \ldots, m$$

Ao determinar-se (3.30) eliminam-se as variáveis duais η_i e η_i^* pela condição

(3.29). A equação (3.28) pode reescrever-se da seguinte forma:

$$\mathbf{w} = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, \mathbf{x}_i$$

e portanto pode obter-se a expansão familiar segundo os vectores de suporte, que exprime sem ambiguidade $\mathbf{w}$ e $f(\mathbf{x})$ como uma combinação linear de um subconjunto de exemplos $\mathbf{x}_i$:

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \langle \mathbf{x}_i, \mathbf{x} \rangle - b \tag{3.31}$$

Como se pode notar verifica-se tal como no algoritmo da classificação que a função de regressão se pode exprimir como uma função de produtos internos entre os exemplos. Como a determinação de $f(\mathbf{x})$ não requer o conhecimento explícito de $\mathbf{w}$ pode-se assim estender esta expressão à utilização de funções *kernel*.

A determinação da parcela b da expressão (3.35) obtém-se através da utilização das condições de Karush-Kuhn-Tucker (KKT), as quais estipulam que no ponto correspondente à solução os produtos entre as variáveis duais e as restrições devem ser nulos:

$$\begin{aligned} \alpha_i (\varepsilon + \xi_i + y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) &= 0 \\ \alpha_i^* (\varepsilon + \xi_i^* - y_i + (\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) &= 0 \end{aligned} \tag{3.32}$$

assim como,

$$\begin{aligned} \left(\frac{C}{m} - \alpha_i\right) \xi_i &= 0 \\ \left(\frac{C}{m} - \alpha_i^*\right) \xi_i^* &= 0 \end{aligned} \tag{3.33}$$

Notar que: (i) Só os exemplos $(\mathbf{x}_i, y_i)$ aos quais correspondem $\alpha_i^{(*)} = C/m$ podem cair fora do "tubo" ε-insensitivo que rodeia a função f , (i.é, $\xi_i^{(*)} > 0$); (ii) Como se tem que $\alpha_i \alpha_i^* = 0$ nunca pode acontecer existirem um conjunto de variáveis duais α_i e α_i^* que possam ser simultaneamente não nulas; (iii) Para $\alpha_i^{(*)} \in (0, C/m)$ tem-se que $\xi_i^{(*)} = 0$ assim como o segundo factor de (3.32) deve anular-se. Por consequência b pode calcular-se da seguinte forma:

$$\begin{aligned} b &= y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - \varepsilon \quad para \quad \alpha_i \in (0, C/m) \\ b &= y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle + \varepsilon \quad para \quad \alpha_i^* \in (0, C/m) \end{aligned} \tag{3.34}$$

Pode reparar-se que da expressão (3.32) resulta que os multiplicadores de Lagrange, α_i, α_i^*, podem ser ambos diferentes de zero apenas no caso de os exemplos

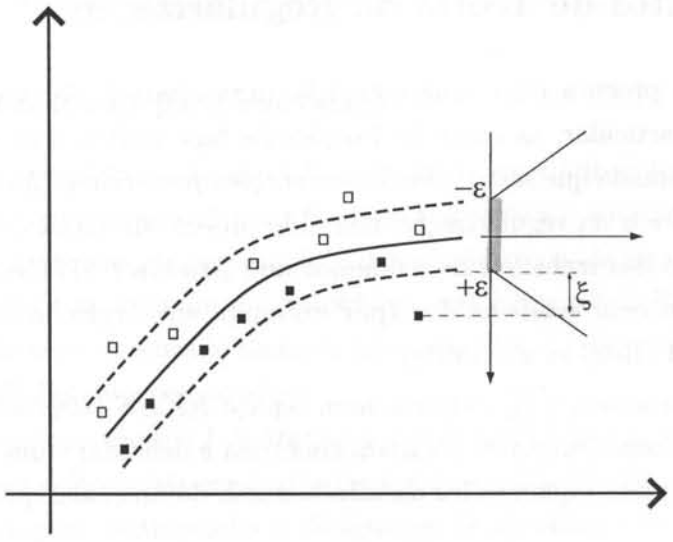


Figura 3.3: Regressão com MVS com uma precisão desejada $\pm\varepsilon$. O gráfico da direita representa a função de custo $\varepsilon - insensitiva$.

serem exteriores ao ε-tubo, $|f(\mathbf{x}_i) - y_i| > \varepsilon$, porque se forem interiores têm que se anular em virtude da segunda condição KKT ser satisfeita. Por consequência, obtém-se um desenvovimento esparso de $\mathbf{w}$ em função de $\mathbf{x}_i$. Os exemplos que não têm coeficiente nulos são os chamados **vectores de suporte**. É geometricamente aceitável que os pontos interiores ao tubo não tenham qualquer contribuição para a solução, podendo-se retirá-los e mesmo assim obter a mesma solução, podendo então dizer-se que eles não são portadores de qualquer "*informação*" sobre a solução.

Pode fazer-se uma generalização ao caso *não linear* substituindo-se em (3.31) o *kernel* trivial $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ pelo *kernel* não linear $k(\mathbf{x}_i, \mathbf{x}_j)$:

$$\begin{aligned} \textbf{Maximizar } W(\alpha, \alpha^*) = & \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\, k(\mathbf{x}_i, \mathbf{x}_j) - \\ & - \varepsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i) \end{aligned} \tag{3.35}$$

Obtém-se $f(\mathbf{x})$ através de uma expansão em vectores de suporte dada por :

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, k(\mathbf{x}_i, \mathbf{x}) - b \tag{3.36}$$

3.2 Elementos de Teoria da Regularização

A presente secção procura fazer uma transição para o estudo das redes neuronais artificiais e, em particular, as redes de funções de base radiais e as redes de perceptrão de multicamada que serão revistas em secções posteriores. As relações entre a teoria das MVS e a da regularização têm sido objecto de estudo, em particular desde a publicação dos trabalhos de Tikhonov and Arsenin (1977), e encontram-se actualmente muito bem estabelecidas (por exemplo, em Evgeniou et al. (2000a), Vapnik (1998) and Girosi et al. (1995)).

Para uma dada norma $\|\cdot\|_K^2$ definida num espaço $RKHS$ (*Reproducing Kernel Hilbert Space*) por uma função K positiva, simétrica e definida, e um λ fixo, seja o esquema de aproximação que resulta da minimização do funcional quadrático:

$$\min H(f) = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(\mathbf{x}_i))^2 - \lambda \|f\|_K^2 \tag{3.37}$$

Formulações dadas por expressões do tipo (3.37) constituem formas especiais da teoria da regularização para resolver problemas "*mal formulados*" e em particular para resolver problemas de aproximação de relações funcionais entre $\mathbf{x}$ e y dado um número finito de exemplos $D_m = \{\mathbf{x}_i, y_i\}_{i=1}^m$. Na teoria da estabilização clássica o termo acerca dos dados é uma função de custo em L_2 para o risco empírico, enquanto que o segundo termo, designado estabilizador, é usualmente um funcional $\Omega(f)$com certas propriedades (Evgeniou (2000)).

O estabilizador na expressão (3.37) efectivamente restringe f a pertencer ao $RKHS$ definido por K. É possível mostrar que a função que minimiza o funcional (3.37) tem a forma:

$$f(\mathbf{x}) = \sum_{i=1}^{m} c_i K(\mathbf{x}, \mathbf{x}_i) \tag{3.38}$$

em que os coeficientes c_i dependem dos dados e satisfazem o seguinte sistema linear de equações:

$$(K + \lambda)\,\mathbf{c} = \mathbf{y}$$

em que $\mathbf{I}$ é matriz identidade e em que se define $(\mathbf{y})_i = y_i$, $(\mathbf{c})_i = c_i$, e $(K)_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

O esquema de aproximação (3.38) admite uma interpretação muito simples em termos de rede com várias unidades numa camada escondida, e utilizando várias funções kernel podem-se obter diversas variantes de redes neuronais (Evgeniou (2000)). Os tópicos que se referem a seguir são aspectos muito simples que permitem uma compreensão elementar do conceito de espaço $RKHS$ (*Reproducing Kernel Hilbert Space*), porém reconhece-se a sua incomplitude, uma vez que não é possível entrar-se

na perspectiva mais abrangente dada pela teoria espectral.

3.2.1 Funcionais para suavização

Viu-se que o problema da aprendizagem pode exprimir-se formalmente como um problema de aproximação de uma função a partir de um conjunto de dados de treino. Porém, este problema de aproximação de uma função pode ter um número infinito de soluções,constituindo assim um "problema mal formulado". Para ser possível obter uma solução com significado torna-se necessário impor algumas restrições adicionais ao espaço $\mathbb{H}$ das soluções possíveis.

O matemático francês J. S. Hadamard (1865-1963) introduziu em 1923 no campo das equações diferenciais em derivadas parciais a designação de "*problema mal formulado*" ao qual contrapunha a designação de "*problema bem formulado*" sempre que a solução única do problema existisse, fosse robusta em relação ao ruído, quer dizer dependesse continuamente dos dados iniciais.

A escolha do espaço de hipóteses $\mathbb{H}$ deve sempre ter em consideração algum conhecimento *a priori* sobre o problema, sem esquecer que diferentes a escolhas de $\mathbb{H}$ vão corresponder a diferentes técnicas de aproximação. Uma das restrições ao espaço $\mathbb{H}$ consiste em que as funções têm que ter *suavidade*. Formula-se a hipótese de que as funções subjacentes a um dado conjunto de exemplos devem ser "*regulares e suaves*", no sentido em que as respostas dadas pela máquina de aprendizagem a dois exemplos próximos um do outro sejam igualmente próximas uma da outra, isto é, a suavidade deve entender-se no sentido de se verificarem pequenas oscilações, ou variações nas respostas da máquina.

Um exemplo clássico de soluções regularizadas é a solução de um sistema sobredeterminado de m equações a n incógnitas, em que $m > n$,

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

em que se pretende determinar $\mathbf{x}$ quando se conhece $\mathbf{A}$ e $\mathbf{y}$. De entre uma infinidade de soluções possíveis a solução regularizada mais comum é a que se obtém no sentido dos mínimos-quadrados, em que se minimiza:

$$\|e\|^2 = (\mathbf{y} - \mathbf{A}\mathbf{x})^T (\mathbf{y} - \mathbf{A}\mathbf{x}).$$

O problemas básico de saber como medir a suavidade e como garantir que a aproximação obtida seja suave é resolvido pela minimização dum funcional de suavização $\Phi(h(x))$, que aplica um conjunto de funções $h(x)$ no conjunto dos números reais positivos $\mathbb{R}^+$. Uma vez que os funcionais de suavização devem ter valores elevados para funções não são suaves, e valores baixos para as que são suaves, para caracteri-

zar uma "*des-suavização*" pode utilizar-se as derivadas de $h(x)$. Por exemplo, Verri and Poggio (2001) (class 3) citam o caso do integral do quadrado da primeira (ou segunda, etc.) derivada de uma função $h(x)$:

$$\Phi(h) = \int_{-\infty}^{+\infty} \left(\frac{d}{dx} h(x) \right)^2 dx$$

pode constituir um exemplo de funcional suavizador (e que pode ser considerado como uma seminorma ou pseudo-norma).

3.2.2 *Kernel* da transformada de Fourier

Pode considerar-se estas mesmas funções no domínio de Fourier. A transformada de Fourier $\tilde{h}(\omega)$ de uma função $h \in L_1$ é uma função complexa, que generaliza a série de Fourier complexa, e pode considerar-se como sendo uma representação do conteúdo de informação (período, amplitude e fase) existente num sinal $h(x)$. Tem a seguinte expressão:

$$\tilde{h}(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} h(x) e^{-i\omega x} dx$$

Por sua vez a transformada inversa de Fourier restitui o original $h(x)$ quando se sabe a função transformada $\tilde{h}(x)$:

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \tilde{h}(\omega) e^{i\omega x} d\omega.$$

A transformada de Fourier da primeira derivada de $h(x)$ é $j\omega\tilde{h}$. Aplicando a relação de Parseval ao produto $h'.h'$, obtém-se:

$$\Phi(h) = \int_{-\infty}^{+\infty} \left(h'(x)\right)^2 dx = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \omega^2 \left| \tilde{h}(\omega)^2 \right| d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\left| \tilde{h}(\omega)^2 \right|}{\tilde{K}(\omega)} d\omega.$$

em que $\tilde{k}(\omega) = 1/w^2$ é uma dada função simétrica positiva, decrescente para zero quando ω tende para infinito. Pode também dizer-se que $\tilde{k}(\omega)$ constitui uma "*função kernel*" Verri and Poggio (2001) (class 3).

A transformada inversa de Fourier de uma qualquer função simétrica positiva $\tilde{K}(\omega)$ que decresce monotonicamente para zero, quando ω tende para infinito, também pode ser considerada uma função *kernel*. Por exemplo:

$$K(x) = \frac{-|x|}{2} \Longleftrightarrow \tilde{k}(\omega) = \frac{1}{w^2}$$

Uma outra possível função *kernel* é, por exemplo:

$$K(x) = e^{-x^2/2\sigma^2} \Leftrightarrow \tilde{K} = e^{-\omega^2\sigma^2/2}$$

Se atendermos à expansão em série de potências o espaço de hipóteses para este *kernel* Gaussiano consiste num conjunto de funções quadráticas integráveis cujas derivadas de ordem ímpar são quadraticamente integráveis.

3.2.3 Regularização da função aproximada

Restringindo a atenção ao caso em que o funcional de suavização se pode escrever utilizando a transformada de Fourier de uma função *kernel* $\tilde{K}$, vamos resolver o problema de determinar o mínimo de

$$\sum_{i=1}^{m} (y_i - h(x_i))^2$$

sujeito à condicionante

$$\frac{1}{2\pi} \int \frac{\left|\tilde{h}(\omega)\right|^2}{\tilde{K}(\omega)} dw \leq \mu$$

para um dado parâmetro $\mu > 0$.

A teoria dos multiplicadores de Lagrange recomenda que se determine primeiramente a função h_λ que minimiza sem restrições o seguinte funcional:

$$F(h) = \sum (y_i - h(x_i))^2 + \lambda \frac{1}{2\pi} \int \frac{\left|\tilde{h}(\omega)\right|^2}{\tilde{K}(\omega)} d\omega \tag{3.39}$$

para cada $\lambda > 0$, e depois determina-se o valor único de λ para o qual a função $f_\lambda$satisfaz a restrição

$$\frac{1}{2\pi} \int \frac{\left|\tilde{h}(\omega)\right|^2}{\tilde{K}(\omega)} d\omega = \mu \, .$$

Notar que:

(*i*) O funcional suavizador de um modo geral pode exprimir-se em termos de um operador condicionante $\mathbf{P}$ e de uma norma, usualmente L_2:

$$\Phi(h) = \|\mathbf{P}\, h(x)\|^2 = \int |\mathbf{P}\, h(x)|^2 \, dx.$$

(*ii*) A escolha deste operador tem uma grande importância na concepção de uma rede *RBF* uma vez que ele determina a função de activação (base);

(*iii*) O problema passa a ser um problema de procura da função h_λ que aproxima os exemplos e minimiza o funcional (3.39), em que λ é um número positivo pequeno,

designado parâmetro de regularização e que é proporcional ao grau de ruído nos exemplos;

(*iv*) O funcional (3.39) pode ser visto como consistindo de duas parcelas, a primeira que se refere à proximidade de $h_\lambda$em relação aos exemplos, i.é, o *erro empírico*, e a segunda parcela que força a suavização da função h_λ, e se chama por vezes *estabilizador*.

(*v*) Uma classe de funções $h(x)$ que minimiza o funcional $F(h)$ tem a seguinte forma geral

$$h(x) = \sum_i w_i G(\mathbf{x}; \mathbf{x}_i) + \sum_j a_j \gamma_j(\mathbf{x}),$$

em que $G(\cdot;\cdot)$ é uma função condicionalmente positiva e definida, chamada *função de Green* do operador diferencial $\hat{\mathbf{P}}\mathbf{P}$ de ordem m centrado em x_i ($\hat{\mathbf{P}}$ é o chamado operador adjunto);

(*vi*) a combinação linear

$$\sum_j a_j \gamma_j(\mathbf{x})$$

é um gerador do espaço nulo do operador $\mathbf{P}$ constitui em muitos casos uma base no espaço dos polinómios de ordem $m-1$.

Um resultado de grande importância referido em Verri and Poggio (2001) (class 3), precursor da subsecção que se segue, redescoberto por vários investigadores Duchon (1977), Meinguet (1979), Wahba (1977), Madych and Nelson (1990), Poggio and Girosi (1989), Girosi (1992), é o seguinte teorema cuja demonstração se omite dado o nível de detalhe de cálculo funcional que requer:

Teorema 54 *Seja $\tilde{K}(\omega)$ a transformada de Fourier de uma função kernel $K(x)$, a solução geral do problema da função aproximada é uma função f_λ que minimiza o funcional (3.39):*

$$\sum (y_i - f(x_i))^2 + \lambda \frac{1}{2\pi} \int \frac{\left|\tilde{f}(w)\right|^2}{\tilde{K}(w)} dw$$

a qual tem a forma

$$f_\lambda(x) = \sum \alpha_i K(x - x_i)$$

para alguns valores apropriados dos coeficientes α_i, $i = 1, ..., m$.

3.2.4 *Kernel* Reprodutivo no Espaço de Hilbert

Um dos aspectos relevantes diz respeito à compreensão de como se pode relacionar entre si o espaço de hipóteses, com o conhecimento a priori e com o funcional de suavização, através da interpretação da solução regularizada em termos da solução cuja norma é mínima num espaço Euclideano apropriado.

Se o *kernel* definir uma seminorma deve adicionar-se um termo correspondente ao funcional de suavização que dá conta do espaço de nulidade não trivial. Suponhamos que se pretendia determinar os valores próprios e os vectores próprios de um *kernel* $\mathbf{K}$ dado por uma matriz $n \times n$. Isto é, sendo $\mathbf{0} \neq \mathbf{u} \in \mathbb{R}^n$, e $\lambda \in \mathbb{R}$, pretende-se encontrar a solução da equação:

$$K\mathbf{u} = \lambda \mathbf{u}.$$

Notar que para qualquer λ, o espaço de vectores que satisfaz esta equação é um subespaço linear de $\mathbb{R}^n$. Interessa apenas o caso especial em que $\mathbf{K}$ se define como (positivo) semipositivo, isto é, $\mathbf{K}$ é uma matriz simétrica e todos os valores próprios são (estritamente) positivos. Como se sabe qualquer a matriz simétrica pode ser diagonalizada e escrever-se na forma

$$\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum \lambda_p \mathbf{u}_p \mathbf{u}_p^T$$

em que $\mathbf{\Lambda}$ é uma matriz diagonal de dimensão n e $\mathbf{U}$ uma matriz ortogonal $n \times n$ em que as colunas são vectores próprios u_p mutuamente ortogonais que correspondem aos valores próprios λ_p. Sempre que os valores próprios são estritamente positivos a matriz $\mathbf{K}$ permite definir um novo produto interno em $\mathbb{H} = \mathbb{R}^n$.

Fazendo o desenvolvimento de cada vector nesta nova base $u_1, ..., u_n$, pode definir-se um para o vector

$$\mathbf{f} = \sum_{p=1}^{n} c_p \mathbf{u}_p$$

e o vector

$$\mathbf{g} = \sum_{p=1}^{n} d_p \mathbf{u}_p$$

um novo produto interno entre $\mathbf{f}$ e $\mathbf{g}$:

$$\langle \mathbf{f}, \mathbf{g} \rangle_K = \sum_{p=1}^{n} \frac{c_p d_p}{\lambda_p} = \mathbf{c}^T \mathbf{\Lambda}^{-1} \mathbf{d} = \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{g}.$$

A nova norma para $\mathbf{f} = \sum_{p=1}^{n} \mathbf{c}_p \mathbf{u}_p$ passa a ser:

$$\|\mathbf{f}\|_K^2 = \sum_{p=1}^{n} \frac{c_p^2}{\lambda_p}$$

O espaço de Hilbert $n-$dimensional munido do produto interno induzido pelo kernel K é um **Kernel Reprodutivo no Espaço de Hilbert** e será designado por $\mathbb{H}_K$.

A propriedade reprodutiva é devida à presença do inverso de K na definição do produto interno. Calculando o produto interno em H_K entre a $i-$ésima linha e qualquer vector $\mathbf{f}$ resulta a $i-$ésima componente de $\mathbf{f}$. Esta abordagem devida

a (Verri and Poggio,2001, class04) carece de mais desenvolvimento, o qual não é possível incluir.

3.2.5 Teorema de Mercer

Seja a função $K(s,t) \in L^2([a,b] \times [a,b])$ um *kernel* de um operador definido positivo, então, pode escrever-se como uma série uniformemente convergente, para todo o s e $t \in [a,b]$

$$K(s,t) = \sum_{p=1}^{\infty} \lambda_p \Phi_p(s) \Phi_p(t)$$

em que $\Phi_p(s)$ são as funções próprias (*eigenfunctions*), ortogonais duas a duas, e λ_p são os correspondentes valores próprios da equação integral

$$\int_a^b K(s,t) \Phi(s) ds = \lambda \Phi(t).$$

A *figura* 3.2 ilustra a ideia de os dados de treino serem sujeitos a uma aplicação não linear sobre um espaço de características, de dimensão igual ou superior ao espaço de entradas, utilizando uma transformação Φ, determinando-se neste novo espaço um hiperplano separador (de margem máxima no caso das máquinas de VS). É para se conseguir uma maior facilidade de cálculo, sem ter que explicitamente levar a aplicação para o espaço de características, que se define uma função kernel de Mercer $K(s,t) = \langle \Phi_p(s) \Phi_p(t) \rangle$ (Schölkopf et al. (1999)). Os modelos de geradores de hipóteses aproximadoras de conjuntos de exemplos dos tipos *MVS*, *FBR*, ou *PMC*, usam o formalismo das equações integrais (de Fredholm), dos problemas variacionais tratados sob o ponto de vista unificado de problemas de optimização com recurso aos *Kernel Reprodutivos no Espaço de Hilbert.*

3.3 Redes de Funções de Base Radiais (*FBR*)

Este tipo de redes pertencentes a uma numerosa classe de esquemas de aproximação são exemplos particulares de redes regularizadas, nas quais o estabilizador está dotado de simetria radial, isto é, satisfaz a relação:

$$\Phi[f(\mathbf{x})] = \Phi[f(\mathbf{R}\mathbf{x})]$$

para uma dada matriz de rotação $\mathbf{R}$. Esta escolha reflete um pressuposto a priori de que todas as variáveis têm a mesma relevância, não havendo portanto direcções privilegiadas.

3.3.1 Problema da aproximação de funções

A origem das *FBR*s radica nos métodos de interpolação exacta de conjuntos de dados multidimensionais. Recorde-se que existem dois teoremas sobre aproximação de funções, um devido a Weierstrass e outro a Cybenko, que desempenharam um papel muito importante no desenvolvimento da teoria da aproximação (Verri and Poggio (2001), class07).

Teorema 55 (Weierstrass, 1885) *O conjunto de todos os polinómios $P_{[a,b]}$ é denso em $C[a,b]$, por outras palavras, dada uma função $h \in C[a,b]$ e um $\varepsilon > 0$, existe um polinómio p para o qual se verifica que $|p(x) - h(x)| < \epsilon$ para todo $x \in [a,b]$.*

Teorema 56 (Cybenko, 1989) *Seja σ uma qualquer função sigmoidal e I_n um cubo n-dimensional $[0,1]^n$, então, qualquer somatório finito da forma*

$$f(\mathbf{x}) = \sum_{i=1}^{m} c_i \sigma\left(\langle \mathbf{w}_i, \mathbf{x} \rangle + b\right)$$

é denso em $C[I_n]$. Por outras palavras, dada uma função $h \in C[I_n]$ e um $\varepsilon > 0$, existe um somatório $f(x)$, com a forma acima indicada tal que, $|f(x) - h(x)| \leq \epsilon$ para todo o $x \in I_n$.

O teorema seguinte afirma que um qualquer conjunto de exemplos que inclua os valores alvo y pode ser associado a uma representação com *FBR*s constituída por uma combinação linear de um conjunto de funções de base $K(\cdot)$ em que para cada exemplo do conjunto de treino existe uma função da forma $K(\|x - p_i\|)$.

Teorema 57 (Powell, 1987) *Seja () K uma função de Função de Base Radial, e I_n um cubo $n-$dimensional $[0,1]^n$, então, qualquer somatório finito da forma*

$$f(\mathbf{x}) = \sum_{i=1}^{m} c_i K\left(\mathbf{x} - \mathbf{p}_i\right)$$

é denso em $C[I_n]$. Por outras palavras, dado uma função $h \in C[I_n]$ e $\varepsilon > 0$, existe um somatório $f(x)$, da forma acima indicada, para a qual $|f(x) - h(x)| \leq \epsilon$ para todo o $x \in I_n$ Powell (1987).

3.3.2 Interpolação exacta com ruído

O problema da interpolação exacta pode formular-se da seguinte forma. Dado um conjunto $\{(x_i, y_i) \,|i = 0, \ldots, n\}$, em que x_i são pontos todos distintos, pretende-se obter uma função $h(x)$ tal que $y_i = h(x_i)$.

Admitindo-se que possam existir alguns erros nos dados, pode aliviar-se a definição através da exclusão da hipótese de ser possível a igualdade, mas mantendo-se uma certa aproximação (por exemplo no sentido dos mínimos quadrados). Depois, pode usar-se $h(x)$ para estimar o valor alvo, y, correspondente a um ponto desconhecido x_{novo} que não pertença ao conjunto de dados inicial mas tal que $min(x_i) < x_{novo} < max(x_i)$.

Na interpolação clássica utilizam-se séries de funções elementares, como por exemplo polinómios simples, polinómios de Lagrange, funções definidas positivas e semi-positivas.

As condições de interpolação do teorema (57) podem escrever-se na forma matricial, se $\mathbf{t}$ e $\mathbf{w}$ forem, respectivamente, matrizes do vector de alvos e pesos, e $\mathbf{K}$ for uma matriz quadrada de elementos tais que $K_{mn} = k\left(\|\mathbf{x}_m - \mathbf{x}_n\|\right)$, então:

$$\mathbf{K}\mathbf{w} = \mathbf{t}$$

Existindo a matriz inversa $\mathbf{K}^{-1}$, o que se verifica de uma maneira geral sempre que os pontos de entrada são distintos Micchelli (1986) obtém-se a solução:

$$\mathbf{w} = \mathbf{K}^{-1}\mathbf{t}$$

No contexto da interpolação exacta muitas das propriedades da função de interpolação são relativamente insensíveis às variações da função $k(.)$, razão pela qual uma das funções de base mais usadas é a precisamente a função localizada Gaussiana.:

$$K(z) = \exp\left(-\frac{z^2}{2\sigma^2}\right)$$

na qual o parâmetro σcontrola as propriedades de suavidade da função de interpolação. Facilmente se generaliza este resultada a redes com várias saídas Bishop (1995)).

No caso do ruído nos dados ser muito elevado a interpolação exacta não dá em geral uma boa generalização. Pode também não *ter boa adaptação* (*rescaling*) se o conjunto de dados for muito grande devido ao número de funções de base ser igual ao número de pontos de dados. Para obstar a estes problemas introduziram-se as seguintes modificações:

1. Redução do número de funções de base, que passou a ser muito menor do que o número de pontos de dados.

2. Permissão para os centros coincidirem com os centros de aglomerações de pon-

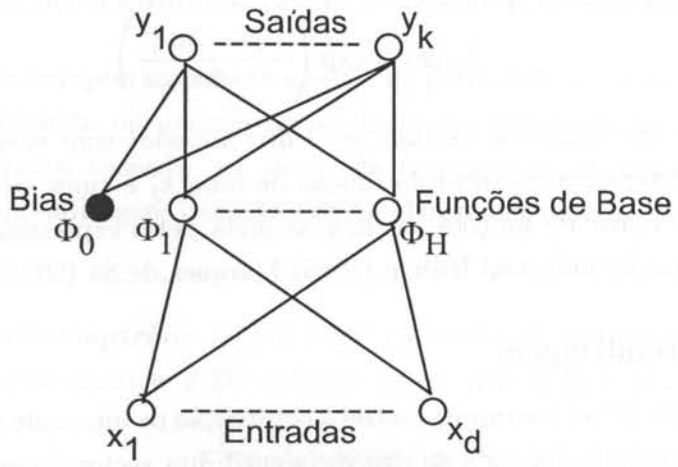


Figura 3.4: Arquitectura de uma rede de funções de base radial

tos de dados.

3. Individualização da largura σ de cada uma das funções de base .

4. Inclusão de um *bias* na combinação linear das funções de base.

Com a introdução destas modificações na fórmula de interpolação exacta obtém-se as seguintes expressões para as saídas:

$$y_k(\mathbf{x}) = \sum_{j=1}^{m} w_{kj} k_j(\mathbf{x}) + w_{k0} = \tag{3.40}$$

$$= \sum_{j=0}^{m} w_{kj} k_j(\mathbf{x}) \tag{3.41}$$

A expressão (3.41) resulta da absorção dos *bias* w_{k0} no símbolo de somatório, através da inclusão de uma função extra de base k_0 cujo valor de activação é constante e igual a 1.

A figura 3.4 representa a arquitectura de uma RN com FBR correspondente à expressão (3.40). A camada escondida contém as funções de base radial. As conexões entre a função de base k_j e as entradas $i = 1 \ldots d$ representam os pesos μ_{ji} do vector μ_j. Os pesos w_{kj} estão representados por conexões entre as funções de base e as unidades de saída. Os *bias* estão representados por pesos das conexões entre a função de base $k_0 = 1$.

No caso de se usarem funções de base Gaussianas temos que:

$$k_j(x) = \exp\left(-\frac{\|\mathbf{x} - \mu_j\|^2}{2\sigma_j^2}\right)$$

em que $\mathbf{x}$ é um vector de entrada $m - dimensional$ com elementos x_i e μ_j é o vector que determina o centro da função de base k_j e cujos elementos são μ_{ji}. A sobreposição linear de funções localizadas dada pelas expressão (3.40) é capaz de uma aproximação universal Bishop (1995) Marques_de_Sá (2001a).

3.3.3 Aprendizagem

Como se acaba de ver é sempre possível a construção de uma rede neuronal utilizando unidades escondidas baseada na distância entre um vector de entrada e um vector protótipo definido para cada unidade. As redes de Função de Base Radiais podem tem um algoritmo de aprendizagem que se processa em duas fases e torna o processo de aprendizagem substancialmente mais rápido que o das MVS e PMC. Uma descrição mais pormenorizada encontra-se no capítulo seguinte. Muito sucintamente, pode aqui antecipar-se que numa primeira fase, treina-se a camada escondida de uma forma não supervisada usando apenas os dados de entrada e ignorando os dados alvo. Na segunda fase, treina-se a camada de saída como se fosse um perceptrão de uma única camada supervisado. No fim do treino a rede de FBR utiliza sempre as mesmas bases para gerar discriminantes lineares generalizados e aproximar a função para a qual foi treinado.

3.4 Perceptrão

Relembrando a analogia entre neurónio e perceptrão, sabe-se que um neurónio recebe muitos estímulos provenientes de outros neurónios à sua volta. Estes estímulos podem ser de excitação, ou de inibição, e quando a combinação deles todos atinge um determinado limiar esse neurónio entra em acção (*fires*). Pode interpretar-se o modelo matemático designado *perceptrão* desta mesma forma, com a diferença de que os estímulos são vectores $\mathbf{x}$ n-dimensionais de números reais. Cada uma das componentes x_i deste vector estímulo tem associado um determinado *peso* w_i, um número real que poderá ser positivo, dizendo-se neste caso que se trata de um estímulo *excitatório*, ou negativo em que o estímulo é *inibitório*. Se a soma pesada dos estímulos:

$$\sum_{i=1}^{n} w_i x_i = \mathbf{x}^T \mathbf{w}$$

for superior a um dado limiar b então o perceptrão fica activo e pode dizer-se que o vector de entrada $\mathbf{x}$ pertence a uma dada classe, $\mathbb{A}$, e no caso alternativo diz-se que ele pertence a uma outra classe, $\mathbb{B}$.

3.4.1 Descrição segundo a teoria estatística da aprendizagem

Com base numa interpretação deste modelo de perceptrão é possível construirem-se definições matemática de perceptrão originais, e explicações do problema de optimização dos pesos baseadas na teoria da programação matemática e na teoria estatística da aprendizagem Bennett and Bredensteiner (1998).

Definição 58 (Perceptrão) *Seja* $\mathbf{x} \in \mathbf{R}^n$ *um vector de entrada a ser classificado. Um perceptrão com pesos* $\mathbf{w} \in \mathbf{R}^n$ *e limiar* $b \in \mathbf{R}$ *pode definir-se como*

$$\begin{aligned} \mathbf{x}^T\mathbf{w} - b > 0 &\implies \mathbf{x} \in \mathbb{A} \\ \mathbf{x}^T\mathbf{w} - b < 0 &\implies \mathbf{x} \in \mathbb{B} \end{aligned} \tag{3.42}$$

Se acontecer $\mathbf{x}^T\mathbf{w} = b$ *a classe de* $\mathbf{x}$ *ficaria indefinida, pelo que, por convenção, vai supor-se que neste caso* $\mathbf{x} \in \mathbb{B}$.

Geometricamente, a aprendizagem de um perceptrão corresponde à determinação de um plano que separa os dois conjuntos dicotómicos $\mathbb{A}$ e $\mathbb{B}$. O perceptrão determina o plano separador $\mathbf{x}^T\mathbf{w} = b$, em que o vector $\mathbf{w}$ é um vector perpendicular ao plano, e $\frac{|b|}{\|\mathbf{w}\|}$ é a distância Euclideana deste plano à origem. Supondo que as coordenadas dos vectores de entrada de $\mathbb{A}$ são dadas por uma matriz $\mathbf{A}$ de tamanho $m \times n$, e que as coordenadas dos vectores de entrada de $\mathbb{B}$ são dadas por uma matriz $\mathbf{B}$ de tamanho $k \times n$. A *definição* 58 de modelo de perceptrão permite saber que existe um perceptrão que classifica correctamente dois conjuntos finitos se e só se existirem $\mathbf{w}$ e b que satisfazem as desigualdades:

$$\mathbf{Aw} > \mathbf{1}b \quad e \quad \mathbf{Bw} < \mathbf{1}b,$$

em que $\mathbf{1}$ é um vector de unidades com as dimensões apropriadas. Se existirem $\mathbf{w}$ e b diz-se que os dois conjuntos são linearmente separáveis. É claro que $\mathbf{w}$ e b são os pesos e o limiar, respectivamente, do perceptrão separador. Pode, por exemplo, imaginar-se no espaço $\mathbf{R}^2$ dois conjuntos de pontos e uma recta a separá-los, em que os pontos de $\mathbb{A}$ pertencem ao semi-plano definido pela recta tal que $\mathbf{x^T w} > b$, e os pontos de $\mathbb{B}$ pertencem ao outro semi-plano tal que $\mathbf{x^T w} < b$. Considere-se agora cada uma das envoltórias convexas dos conjuntos $\mathbb{A}$ e $\mathbb{B}$, constituídas pelas áreas interiores de cada um dos dois polígonos fechados definidos pelos segmentos de recta que unem os pontos mais periféricos de cada um destes conjuntos. Relembre-se que uma envoltória convexa (*convex hull*) é um conjunto que se pode exprimir em termos de combinações lineares c e d de todos os pontos de um dado conjunto. Por

BIBLIOTECA F.E. PORTO

exemplo, para o conjunto $\mathbb{A}$ a combinação convexa pode definir-se:

$$\mathbf{c}^T = A_1 u_1 + \ldots + A_m u_m = \mathbf{A}^T \mathbf{u}, \text{ em que } \mathbf{u} \in \mathbf{R}^m, \mathbf{u} \geq 0, \sum_{i=1}^{m} u_i = \mathbf{1}^T \mathbf{u} = 1.$$

Para o conjunto $\mathbb{B}$ a combinação convexa pode definir-se:

$$\mathbf{d}^T = B_1 u_1 + \ldots + B_m u_m = \mathbf{B}^T \mathbf{v}, \text{ em que } \mathbf{v} \in \mathbf{R}^m, \mathbf{v} \geq 0, \sum_{i=1}^{m} v_i = \mathbf{1}^T \mathbf{v} = 1$$

Pode então dizer-se que para um perceptrão classificar correctamente dois conjuntos separáveis, $\mathbb{A}$ e $\mathbb{B}$, é uma condição necessária e suficiente que nenhum ponto possa ser expresso como uma combinação convexa simultaneamente de $\mathbb{A}$ e $\mathbb{B}$, portanto, pode-se expressar o seguinte problema de optimização:

Problema 59 (classificação correcta)

$$\left.\begin{array}{c} \mathbf{A}^T \mathbf{u} = \mathbf{B}^T \mathbf{v} \\ \mathbf{1}^T \mathbf{u} = \mathbf{1}^T \mathbf{v} \\ \mathbf{u} \geq 0 \;\; \mathbf{v} \geq 0 \end{array}\right\} \textit{não pode ter solução} \tag{3.43}$$

À primeira vista pode parecer que as duas caracterizações (3.42) e (3.43), sobre a capacidade de aprendizagem de um perceptrão não se relacionam entre si. É possível, usando o Teorema de Gordan da Alternativa, demonstrar-se que ambas resolvem o mesmo problema subjacente (ver Bennett and Bredensteiner (1998)). Mais, estas duas formulações apesar de estarem em dois espaços diferentes podem resolver o mesmo problema subjacente, ambas podem ser usadas para definir a separabilidade linear de dois conjuntos e ambas constituem uma das muitas facetas que a *dualidade* pode ter em geometria. Neste caso o problema dual da determinação do plano separador consiste no problema de determinar se duas envoltórias convexas se intersectam. A seguir aborda-se o problema da aprendizagem no caso linearmente separável, o qual uma vez compreendido facilmente se generaliza ao caso linearmente não separável.

Intuitivamente vê-se que construindo o segmento de recta que une os dois pontos mais próximos das envoltórias convexas de dois conjuntos $\mathbb{A}$ e $\mathbb{B}$, o seu plano bissector, ortogonal a este segmento de recta, é o "melhor" plano separador. O sentido a dar ao adjectivo "melhor" é o de que este plano é aquele que melhor generaliza, produz classificações correctas de pontos desconhecidos pertencentes à vizinhança das envoltórias convexas $\mathbb{A}$ e $\mathbb{B}$.

O problema da determinação dos dois pontos mais próximos nas envoltórias convexas pode exprimir-se como um problema de optimização:

BIBLIOTECA F.E. PORTO

Problema 60 (dos dois pontos mais próximos)

$$\begin{array}{c} \min_{u,v} \frac{1}{2} \left\| \mathbf{A}^T\mathbf{u} - \mathbf{B}^T\mathbf{v} \right\|^2 \\ \text{s.a. } \mathbf{1}^T\mathbf{u} = 1 \;\; \mathbf{1}^T\mathbf{v} = 1 \\ \mathbf{u} \geq 0 \;\; \mathbf{v} \geq 0 \end{array} \tag{3.44}$$

Pode construir-se um perceptrão $(\mathbf{w}, b)$ a partir dos resultados do *problema* (3.44). O vector dos pesos $\mathbf{w}$ sendo normal ao plano separador, é exactamente o vector cujas extremidades são os dois pontos mais próximos das envoltórias convexas. Suponhamos que $\bar{\mathbf{u}}$ e $\bar{\mathbf{v}}$ são soluções óptimas de (3.44), então, os dois pontos mais próximos são

$$\mathbf{c} = \mathbf{A}^T\bar{\mathbf{u}} \mathbf{e} \mathbf{d} = \mathbf{B}^T\bar{\mathbf{v}}.$$

Os pesos do perceptrão são dados pela diferença entre estes dois pontos:

$$\mathbf{w} = \mathbf{c} - \mathbf{d} = \mathbf{A}^T\bar{\mathbf{u}} - \mathbf{B}^T\bar{\mathbf{v}}.$$

O limiar b é a distância da origem ao ponto médio do segmento de recta definido pelos dois pontos mais próximos, os quais são as extremidades de $\mathbf{w}$, logo, o limiar é dado por:

$$b = \left(\frac{\mathbf{c}+\mathbf{d}}{2}\right)^T \mathbf{w} = \frac{\bar{\mathbf{u}}^T\mathbf{A}\mathbf{w} - \bar{\mathbf{v}}^T\mathbf{B}\mathbf{w}}{2}.$$

Recorrendo à dualidade pode transformar-se este problema num problema de optimização no espaço do perceptrão. Pode mostrar-se que o *problema* (3.44) tem um dual que se pode exprimir pelo seguinte *problema*:

Problema 61 (dual)

$$\begin{array}{c} \min_{w,\alpha,\beta} \frac{1}{2} \|\mathbf{w}\|^2 - (\alpha - \beta) \\ \textit{s.a.} \quad \mathbf{A}\mathbf{w} - \alpha\mathbf{1} \geq 0 \\ -\mathbf{B}\mathbf{w} - \beta\mathbf{1} \geq 0 \end{array} \tag{3.45}$$

Uma interpretação simplificada deste *problema* (3.45) pode resultar considerando que todos os pontos de $\mathbb{A}$ são condicionados a estar no "lado maior que" do $\alpha-$plano $\mathbf{x}^T\mathbf{w} = \alpha$, e que todos os pontos de $\mathbb{B}$ são condicionados a estar no "lado menor que" do $\beta-$plano $\mathbf{x}^T\mathbf{w} = \beta$. (ver *Figura* 3.5).

O $\alpha-$plano é um suporte do plano $\mathbb{A}$ e o $\beta-$plano é um suporte do plano $\mathbb{B}$. A distância entre estes dois planos de suporte paralelos e portanto entre as duas envoltórias convexas é

$$\frac{\alpha - \beta}{\|\mathbf{w}\|} \tag{3.46}$$

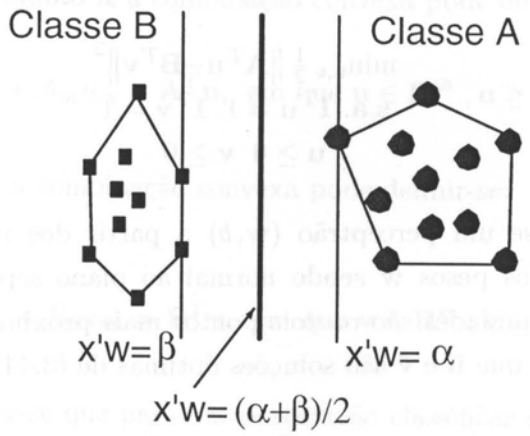


Figura 3.5: Maximização da distância entre dois planos de suporte paralelos através da resolução do problema dual.

A função objectivo de (3.45) minimiza $\|\mathbf{w}\|^2$ e maximiza $\alpha - \beta$. Portanto, o *problema* (3.45) maximiza a distância entre os dois planos paralelos. Sejam $\hat{\mathbf{w}}$, $\hat{\alpha}$ e $\hat{\beta}$ uma solução de (3.45). Para problemas de classes separáveis verifica-se $\hat{\alpha} - \hat{\beta} > 0$. Os dois planos de suporte paralelos estão representados na *Figura* 3.5. O plano separador final é o plano a meia distância entre estes dois planos:

$$\mathbf{x}^T \hat{\mathbf{w}} = \frac{\hat{\alpha} + \hat{\beta}}{2}.$$

Os *problemas* (3.44) e (3.45) são ambos problemas de optimização por programação quadrática com restrições lineares. O *problema* (3.44) tem restrições mais simples e o número de variáveis depende apenas do número total de pontos, enquanto que o *problema* (3.45) pode oferecer um enquadramento unificador para outras modos de construir um perceptrão. Existem duas simplificações que se podem fazer para transformar o *problema* (3.45). Numa, acrescenta-se uma nova restrição $\|\mathbf{w}\|^2 = 1$ e maximiza-se $\alpha - \beta$, e na outra, mantém-se constante $\alpha - \beta$ num dado valor positivo e minimiza-se $\|\mathbf{w}\|^2$. Da primeira simplificação resulta o seguinte *problema*:

Problema 62 (dual simplificado - I)

$$\begin{aligned} &\max_{w,\alpha,\beta} (\alpha - \beta) \\ &\textit{s.a.} \quad \mathbf{A}\mathbf{w} - \alpha\mathbf{1} \geq 0 \\ &\qquad -\mathbf{B}\mathbf{w} + \beta\mathbf{1} \geq 0 \\ &\qquad \|\mathbf{w}\|^2 = 1 \end{aligned} \tag{3.47}$$

Os problemas de optimização acima formulados aplicam-se apenas ao caso de as classes serem separáveis, tornando-se necessários alguns cuidados, semelhantes aos referidos mais acima na subsecção dos MSV classificadoras, para descrever os problemas no caso linearmente não separável.

3.4.2 Máquinas de aprendizagem lineares

As técnicas usadas na classificação binária originais de Rosenblatt (1959) eram *técnicas lineares*, que utilizavam uma função real, $h : \mathbb{X} \to \mathbf{R}$, de tal modo que um dado vector $\mathbf{x}$, pertencente ao espaço de entradas $\mathbb{X} \subseteq R^{\mathbf{n}}$, seria classificado na classe positiva se $h(\mathbf{x}) \geq 0$, e na classe negativa no caso contrário. Se $h(\mathbf{x})$ for uma função linear na variável $\mathbf{x} \in R^{\mathbf{n}}$ pode escrever-se:

$$h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \sum_{i=1}^{m} w_i.x_i + b$$

em que $(\mathbf{w}, b) \in \mathbf{R}^n \times \mathbf{R}$ são os parâmetros que controlam a função $h(\mathbf{x})$, e a regra de decisão é dada por

$$sgn\,(h(\mathbf{x}))$$

convencionando-se que $sign(0) = 1$. Neste método de aprendizagem os parâmetros $(\mathbf{w}, b)$ têm que ser aprendidos com bases no conjunto dos exemplos de treino.

Pode interpretar-se geometricamente este tipo de hipóteses considerando uma partição do espaço das entradas $\mathbb{X}$ em dois subespaços criados pelo hiperplano de decisão, discriminante linear, definido pela equação

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0.$$

Entende-se por hiperplano um sub-espaço afim de dimensão $n - 1$ que realiza uma partição do espaço $\mathbb{X}$ em dois semi-espaços cada um correspondente a entradas que originam classificações numa e só numa das duas classes distintas. Na *figura* 3.6 o hiperplano está representado pela recta a negro, a região superior supõe-se ser positiva, e a inferior negativa. O vector $\mathbf{w}$ define uma determinada direcção ao hiperplano, uma vez que é perpendicular a este. O valor b significa a distância percorrida pelo hiperplano na translação até á origem, paralelamente a si mesmo. Portanto, verifica-se que para representar todos os hiperplanos possíveis em $\mathbf{R}^n$ é necessário envolverem-se $n + 1$ parâmetros livres. Este tipo de classificador é geralmente conhecido sob a designação de *perceptrão,* apesar de ele ser equivalente ao *discriminante linear* cuja teoria foi desenvolvida em 1936 por Sir Ronald Fisher (1890-1962).

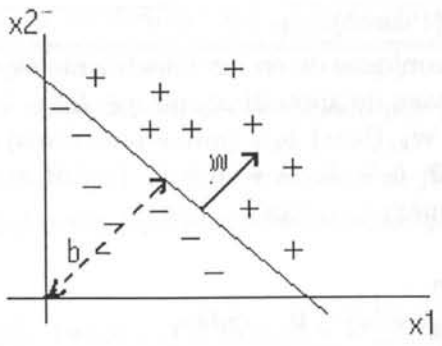


Figura 3.6: Um hiperplano de separação $(\mathbf{w}, b)$ para um conjunto bidimensional

O primeiro algoritmo iterativo de aprendizagem de um perceptrão aplicado a classificadores lineares foi proposto em 1956 por Rosenblatt.

Se os dados forem linearmente separáveis *teorema* (66), devido a Novikoff Novikoff (1962), garante a convergência do algoritmo para a solução.

Em Cristianini and Shawe-Taylor (2000) utiliza-se a designação de *margem funcional* de um exemplo $(\mathbf{x}_i, y_i)$ com respeito ao hiperplano o valor de γ_i dado pela expressão:

$$\gamma_i = y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b \right).$$

Notar que nesta expressão uma classificação correcta do exemplo $(\mathbf{x}_i, y_i)$ implica que a margem γ_i seja sempre positiva. A distribuição da margem funcional de um hiperplano $(\mathbf{w},b)$ relativamente a um dado conjunto de treino $\mathbb{Z}$ é a distribuição das margens dos exemplos de $\mathbb{Z}$, e o seu mínimo designa-se por margem funcional do hiperplano $(\mathbf{w},b)$ em relação ao conjunto de treino $\mathbb{Z}$.

O algoritmo (*primal*) indicado na tabela (3.4) permite determinar o hiperplano óptimo $(\mathbf{w}^*,b^*)$, $\mathbf{w}^* \in \mathbf{R}^n$ e $b^* \in \mathbf{R}$, tal que $\gamma_i > 0$, $i = 1, \ldots, m$. Inicializa-se um vector peso $\mathbf{w}_0 = \mathbf{0}$, e um valor para o limiar (*bias*) $b = 0$. Sempre que um exemplo de treino é mal classificado pelos pesos correntes $(\mathbf{w},b)$ estes vectores são actualizados. Este algoritmo converge garantidamente desde que o hiperplano que classifique correctamente os exemplos de treino exista e, neste caso, diz-se que são linearmente separáveis. Pode mostrar-se que o número de iterações depende da margem γ.

Muitas vezes em lugar da margem funcional tem mais interesse a *margem geométrica* que se define pela distância euclideanas dos exemplos ao hiperplano normalizado $\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}, \frac{b}{\|\mathbf{w}\|}\right)$. Neste caso designa-se por *margem de um conjunto de treino* $\mathbb{Z}$ o valor máximo da sua margem geométrica em relação a todos os hiperplanos possíveis. O único hiperplano para o qual se verifica este valor máximo chama-se *hiperplano de margem máxima.*

Algoritmo 3.1(Primal)

Entradas : um conjunto de treino linearmente separável: $\{\mathbf{x}_i, y_i\}, i = 1 \ldots m$
uma taxa de aprendizagem: $\eta \in R^+$
Saídas : (pesos) $\mathbf{w}_k$, (bias) b_k e (número de erros) k.
Iniciar : $\mathbf{w}_0 \leftarrow \mathbf{0}$; $b_0 \leftarrow 0$; $k \leftarrow 0$
$M \leftarrow \max_{1 \leq i \leq m} \|\mathbf{x}_i\|$
Repetir
Para $i = 1$ até m
Se $y_i (\langle \mathbf{w}_k, \mathbf{x}_i \rangle + b_k) \leq 0$, *então,*
$\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \eta\, y_i\, \mathbf{x}_i$
$b_{k+1} \leftarrow b_k + \eta\, y_i\, M^2$
$k \leftarrow k + 1$
Fim de se;
Fim de para;
Até que não se verifiquem mais erros no ciclo *Para.*
Terminar

Tabela 3.4: Forma primal do algoritmo do perceptrao

Teorema 66 (da convergência, Novikoff, 1962) *Seja um dado conjunto de vectores de treino* $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$, *em que* $\mathbf{x}_i \in \mathbf{R}^n$ *e* $y_i \in \{-1, 1\}$, *linearmente separáveis. Existem os parâmetros do hiperplano óptimo* $\mathbf{w}^* \in \mathbf{R}^n$ *e* $b^* \in \mathbf{R}$, *tal que:*

$$(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*)\, y_i > 0,\ i = 1, \ldots, m,$$

que um perceptrão pode aprender num número finito de iterações. Designando por M *o maior dos vectores de treino:*

$$M = \max_{1 \leq i \leq m} \|x_i\|,$$

e por γ margem de um conjunto de treino *(o dobro da distância do hiperplano separador óptimo ao vector de entrada mais próximo), tem-se que o número de iterações necessárias admite como majorante*

$$\tau \leq \left(\frac{2M}{\gamma} \right)^2 \tag{3.50}$$

Para exemplos não linearmente separáveis o algoritmo não converge originando valores de $\mathbf{w}$ oscilantes que mudam a hipótese $\mathbf{w}_i$ sempre que aparecer um exemplo mal classificado.

Pode demonstrar-se (teorema Freund & Schapire Freund and Schapire (1998)) que existe um limite para o número de erros cometidos neste caso. Existe uma equivalência com as máquinas de vectores de suporte na medida em que se enquadra no conceito de margem atrás referido.

Primeiramente, precisa-se de definir uma quantidade ξ_i que exprima em que medida um exemplo falha em cair dentro de uma margem que dista do hiperplano de um distância b. Se $\xi_i > b$ então, $\mathbf{x}_i$ é mal classificado por $(\mathbf{w}, b)$

Definição 67 (variável de folga) *Define-se variável de folga da margem relativamente a um exemplo* $(\mathbf{x}_i, y_i)$ *a um hiperplano* $(\mathbf{w}, b)$ *e margem alvo* $\gamma > 0$*:*

$$\xi_i = \xi\left((\mathbf{x}_i, y_i), (\mathbf{w}, b), \gamma\right) = \max\left(0, \gamma - y_i\left(\langle \mathbf{w}, \mathbf{x}\rangle + b\right)\right)$$

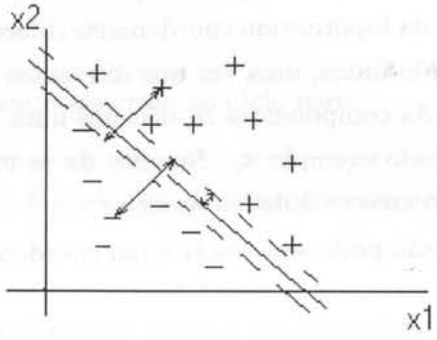


Figura 3.7: Margem de um hiperplano de separação $(\mathbf{w}, b)$ para um conjunto bidimensional.

A *figura* 3.7 mostra as distancias ξ_i de dois pontos (-) e (+) do conjunto de exemplos, assim como a margem γ, igual à distância que separa uma das rectas a traço interrompido e o hiperplano separador, a traço contínuo.

Teorema 68 (*Freund & Schapire, 1998*) *Seja um conjunto de treino não trivial, sem exemplos duplicados e em que* $\|x_i\| \leq M$ *. Seja* $(\mathbf{w}, b)$ *um qualquer hiperplano com* $\|\mathbf{w}\| = 1$*, e para* $\gamma > 0$ *define-se a medida:*

$$D = \sqrt{\sum_{i=1}^{m} \xi_i^2} = \sqrt{\sum_{i=1}^{m} \xi\left[(\mathbf{x}_i, y_i), (\mathbf{w}, b), \gamma\right]^2}$$

Então, o número de erros τ *na primeira execução do ciclo "Para" do algoritmo primal do perceptrão (Tabela (3.4)) para os exemplos de* $\mathbb{Z}$ *é limitado por:*

$$\left(\frac{2(M + D)}{\gamma}\right)^2.$$

O funcionamento do algoritmo do perceptrão pode ver-se como um vector inicial de pesos arbitrário $\mathbf{w}_0$ ao qual se vão adicionando, ora exemplos positivos mal classificados, ora subtraindo exemplos negativos mal classificados. Portanto, se $\mathbf{w}_0 = \mathbf{0}$,

o vector de pesos final será uma combinação linear dos exemplos de treino:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \, y_i \, \mathbf{x}_i.$$

Uma vez que o sinal do coeficiente de $\mathbf{x}_i$ é dado pela etiqueta y_i os α_i são valores positivos proporcionais ao número de vezes que a classificação errada de $\mathbf{x}_i$ tenha causado uma actualização do vector peso. Os valores de α_i serão tanto maiores quanto maior for o número de erros causados.

Dada um conjunto de treino $\mathbb{Z}$ pode pensar-se o vector α como sendo uma representação alternativa da hipótese em coordenadas diferentes, ou *duais*. Este desenvolvimento não é contudo único, uma vez que diferentes α podem corresponder a uma mesma hipótese $\mathbf{w}$. As componentes α_i dão-nos uma ideia intuitiva do conteúdo da informação de um dado exemplo $\mathbf{x}_i$. No caso de os exemplos serem não separáveis estes coeficientes vão crescer indefinidamente.

A função de decisão pode ser reescrita em coordenadas duais da seguinte forma:

$$\begin{aligned} h(\mathbf{x}) &= sgn\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right) = \\ &= sgn\left(\sum_{i=1}^{m} \alpha_i \, y_i \, \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right) \end{aligned}$$

e o algoritmo do perceptrão pode ser expresso também nesta forma dual como se indica na Tabela 3.5. Notar que a taxa de aprendizagem afecta apenas a escala dos hiperplanos, e não o algoritmo cujo vector inicial é nulo, pelo que se omitiu no algoritmo. Esta formulação alternativa do algoritmo do perceptrão e a sua função de decisão tem propriedades interessantes. Por exemplo, o facto de os pontos que custam mais a aprender irem ter valores de α_i mais elevados pode usar-se para ordenar os dados em função do seu conteúdo de informação. Como se vê os conceitos de margem, a distribuição da margem, a representação dual, que foram referidos nas secções anteriores relativas a MVS, existem subjacentes no conceito de perceptrão simples.

Notar ainda que:

(i) Uma vez que o número de actualizações é igual ao número de erros, e que cada actualização implica que se adicione 1 exactamente a um dos seus componentes verifica-se que:

$$|\alpha| \leq \left(\frac{2M}{\gamma}\right)^2,$$

em que o segundo membro é o majorante dado por (3.50). Pode portanto ver-se a norma L_1 de α como uma medida da complexidade da função alvo na representação

Algoritmo 3.2 (Dual)

Entradas : um conjunto de treino linearmente separável: $\{\mathbf{x}_i, y_i\}, i = 1 \ldots m$
Saídas : (α, b) permitem definir a função $h(\mathbf{x})$
Iniciar : $\alpha \leftarrow \mathbf{0}$; $b \leftarrow 0$ (*multiplicador e bias*)
$M \leftarrow \max_{1 \leq i \leq m} \|\mathbf{x}_i\|$
Repetir
Para $i = 1$ até m
Se $y_i \left(\sum_{j=1}^{m} \alpha_j y_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle + b \right) \leq 0$ *então*
$\alpha_i \leftarrow \alpha_i + 1$
$b \leftarrow b + y_i M^2$
Fim de se
Fim de para
Até que não se verifiquem mais erros no ciclo *para.*
Terminar

Tabela 3.5: Forma dual do algoritmo do perceptrao

dual;

(*ii*) Os exemplos de treino que entram no algoritmo podem ser vistos como constituindo os elementos da matriz

$$\mathbf{G} = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle)_{i,j=1}^{m},$$

que se designa por *matriz de Gram.*

3.4.3 Regras clássicas de aprendizagem

Os métodos clássicos de aprendizagem das redes com muitas unidades de perceptrão interligadas são extensivamente utilizados no Capítulo 4, e o conhecimento dos algoritmos de aprendizagem têm um interesse primordial. Porém, por necessidade de abreviar, faz-se aqui uma abordagem sumária da aprendizagem de perceptrões simples pelos métodos de optimização baseados nas derivadas da função de custo, portanto, em técnicas de optimização sem restrições. A história das regras de aprendizagem usadas nestes métodos pode dizer-se que começa desde os primórdios da aprendizagem automatizada, com Rosenblatt (1958), com a proposta de um procedimento muito simples, e que se descreve a seguir, para heuristicamente se minimizar o erro, efectuando-se correcções sucessivas nos pesos das sinapses dos perceptrões. Considera-se o caso já anteriormente abordado da aprendizagem de uma função binária, em que se pretende optimizar um vector de pesos que cause no perceptrão a produção de uma saída correcta, $+1$ ou -1, em conformidade com a etiqueta de cada exemplo de treino. Refere-se a seguir dois algoritmos para resolver este problema de aprendizagem : (*i*) A regra do perceptrão; (*ii*) A regra do delta. Esta última é uma

variante do método dos mínimos quadrados. Qualquer uma destas regras, ou algoritmos, é convergente para uma determinada hipótese aceitável, sob determinadas condições, está na base da aprendizagem de redes com vários perceptrões Bishop (1995), Marques_de_Sá (2001a).

A regra de treino do perceptrão. Uma forma de aprender um vector de pesos consiste em começar com valores casuais das suas componentes e, depois, iterativamente experimentar em todos os exemplos corrigindo o vector de pesos sempre que este classificar erradamente um dado exemplo. Repete-se este processo tantas vezes quantas forem necessárias até o perceptrão classificar todos os exemplos de treino correctamente. Em cada iteração as componentes do vector peso são corrigidas de acordo com a chamada *regra do perceptrão*, a qual revê o peso w_i associado à i-ésima entrada do vector $\mathbf{x}$, que designamos pelo símbolo $[\mathbf{x}]_i$. Supõe-se que $i = 1, \ldots, n$; quer dizer existem n sinápses ou variáveis de entrada.

A regra do perceptrão estipula que:

$$w_i = w_i + \Delta w_i$$

em que os incrementos de cada sinápse Δw_i são dados por:

$$\Delta w_i = \eta \left(y - f\left(\mathbf{x}\right) \right) [\mathbf{x}]_i .$$

em que o factor η é uma constante positiva, muito pequena ($\simeq 0.1$), designada *passo de aprendizagem*, e que tem o papel de moderar a variação dos pesos em cada iteração.

Se um dado exemplo de treino ($\mathbf{x}$,y) é classificado pelo perceptrão correctamente, então, o erro é nulo, $y - f\left(\mathbf{x}\right) = 0$, logo, os incrementos Δw_i deverão ser também todos nulos

$$\Delta w_i = 0, \; i = 1, \ldots, n,$$

e o vector peso não precisa de ser actualizado.

Se o perceptrão classificou um dado exemplo $f\left(\mathbf{x}\right) = -1$ enquanto a etiqueta respectiva indica $y = +1$, para corrigir esta situação algumas componentes do vector peso deverão ser corrigidas no sentido de aumentar o valor de $\langle \mathbf{w}, \mathbf{x} \rangle$. Por exemplo, se a componente $[\mathbf{x}]_i > 0$, então, aumentando w_i provoca-se uma contribuição no bom sentido, o percetrão passa a classificar correctamente esse exemplo. Note-se que a regra do perceptrão faz com que de facto Δw_i seja um incremento positivo porque todos os seus factores η, $(y - f\left(\mathbf{x}\right))$, e a componente i do vector $\mathbf{x}$, $[\mathbf{x}]_i$, são números reais positivos (por exemplo, $\Delta w_i = 0.1\left(1 - (-1)\right) 0.8 = 0.16$).

Se se verificar a situação de classificação errada em que $f\left(\mathbf{x}\right) = 1$ enquanto a etiqueta é $y = -1$, então, as componentes do vector peso positivas deverão decrescer

provocando uma diminuição do valor de $f(\mathbf{x})$. Ao fim de um número finito de iterações, limitado pelo majorante dado pelo *teorema* 66, de Novikoff, verifica-se uma convergência certa deste processo para uma classificação correcta de todos os exemplos, sempre que estes sejam linearmente separáveis.

A regra do delta e descida do gradiente. A regra anterior não permite determinar um vector de pesos que classifique correctamente os exemplos no caso destes *não serem separáveis.* Para ultrapassar esta dificuldade concebeu-se a presente regra que, mesmo quando os exemplos não são separáveis, é convergente, para a melhor aproximação possível que se ajusta às etiquetas dos exemplos de treino. A ideia chave que está por trás da regra do delta é a busca de uma solução no espaço das hipóteses possíveis de vectores de pesos, utilizando o chamado método da "descida do gradiente." O seu algoritmo pode ser mais facilmente entendido considerando a aprendizagem de um perceptrão sem polarização (*bias*), isto é, um perceptrão linear cuja saída se exprime apenas por:

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle .$$

Supondo que se utiliza a função de custo quadrática:

$$c(\mathbf{x}, y, f(\mathbf{w}, \mathbf{x})) = c(\mathbf{w}) = \frac{1}{2} \sum_i (y - f(\mathbf{w}, \mathbf{x}))^2 \tag{3.51}$$

define-se o *erro,* ou *custo total,* $c(\mathbf{w})$, como metade da soma dos quadrados dos erros, estendida a todos os exemplos de treino. É claro que o valor da função custo vai depender dos exemplos, mas e estes se mantiverem invariantes basta caracterizar o custo como uma função exclusiva do vector de pesos, $\mathbf{w} = w_1, w_2, ..., w_n$, em que n designa o número de variáveis de entrada (sinápses).

O algoritmo da descida do gradiente pode tornar-se mais fácil de perceber quando se pode vizualizar o espaço de hipóteses dos vectores de pesos. No caso deste espaço ser bidimensional, $\mathbf{w} = (w_0, w_1)$, pode-se ver na *figura* 3.8 Mitchell (1997) o plano horizontal representando o *espaço das hipóteses* e o eixo vertical representando os valores correspondentes da função custo $c(\mathbf{w})$. Obtém-se, assim, a chamada *superfície de erro (ou de custo)*, e como a função de erro é quadrática e os perceptrões são lineares, isto é, sem função de activação σ, então a superfície de erro é uma parabolóide com um único mínimo global. Estas simplificações destinam-se apenas a clarificar a exposição sem perda de generalidade. Para um dado conjunto de treino existe uma dada superfície parabólica. A procura da solução pela descida do gradiente consiste em determinar um vector de pesos que satisfaz:

$$\mathbf{w}^* = \min c(\mathbf{w}) .$$

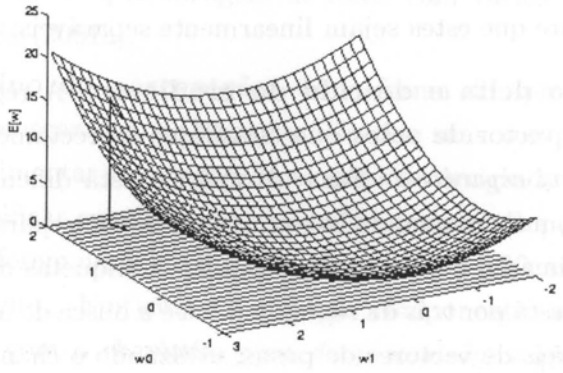


Figura 3.8: Descida do gradiente

A procura deste mínimo de $c(\mathbf{w})$ pela descida do gradiente começa com um vector $\mathbf{w}$ arbitrário inicial o qual se vai actualizando com pequenos passos em cada iteração. Esta actualização verifica-se sempre na direcção e sentido em que a descida for mais íngreme ao longo da superfície de erro, tal como se ilustrada na *figura* 3.8. Este processo de prosseguir até se ter a certeza de chegar a um mínimo global de $c(\mathbf{w})$.

Para determinar a direcção para a qual a descida é mais íngreme ao longo da superfície de erro utiliza-se a derivada de $c(\mathbf{w})$ em ordem a $\mathbf{w}$, chamada *gradiente* de c em relação a $\mathbf{w}$:

$$\nabla c(\mathbf{w}) = \left[\frac{\partial c}{\partial w_0}, \ldots, \frac{\partial c}{\partial w_n}\right]$$

em que $\nabla c(\mathbf{w})$ é também um vector cujas componentes são as derivadas parciais de c em relação às componentes do vector de peso w_i. Quando interpretado como um vector no espaço dos pesos, o gradiente especifica a direcção que produz o mais rápido crescimento de c, e o vector simétrico do gradiente $-\nabla c(\mathbf{w})$ especifica a direcção do mais rápido decrescimento. Assim tem-se a seguinte regra:

$$\mathbf{w} = \mathbf{w} + \Delta\mathbf{w}$$

em que os incrementos $\Delta\mathbf{w}$ são dados por:

$$\Delta\mathbf{w} = -\eta \nabla c(\mathbf{w}).$$

Pode reescrever-se esta regra da forma:

$$w_i = w_i + \Delta w_i$$

na qual os incrementos de cada sinápse Δw_i são dados por:

$$\Delta w_i = -\eta \frac{\partial c}{\partial w_i} \tag{3.52}$$

e em que se destaca que a descida mais íngreme se consegue através da actualização de cada uma das componentes w_i do vector de pesos $\mathbf{w}$ proporcionalmente a $\frac{\partial c}{\partial w_i}$. Pode obter-se um algoritmo prático que iterativamente actualiza os pesos em conformidade com a expressão (3.52) diferenciando a expressão (3.51)

$$\begin{aligned}
\frac{\partial c}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{j=1}^{m} \left(y_j - f\left(\mathbf{x}_j\right)\right)^2 = \\
&= \frac{1}{2} \sum_{j=1}^{m} \frac{\partial}{\partial w_i} \left(y_j - f\left(\mathbf{x}_j\right)\right)^2 = \\
&= \frac{1}{2} \sum_{j=1}^{m} 2 \left(y_j - f\left(\mathbf{x}_j\right)\right) \frac{\partial}{\partial w_i} \left(y_j - f\left(\mathbf{x}_j\right)\right) = \\
&= \sum_{j=1}^{m} \left(y_j - f\left(\mathbf{x}_j\right)\right) \frac{\partial}{\partial w_i} \left(y_j - \langle \mathbf{w}, \mathbf{x}_j \rangle - b\right)
\end{aligned}$$

logo, se designarmos por $[\mathbf{x}_j]_i$ a $i - \acute{e}sima$ componente do $j - \acute{e}simo$ exemplo de treino obtém assim uma expressão para $\frac{\partial c}{\partial w_i}$, em função das entradas $\mathbf{x}_j$ e saídas $f\left(\mathbf{x}_j\right)$ do perceptrão assim como das etiquetas dos exemplos y_j.

$$\frac{\partial c}{\partial w_i} = \sum_{j=1}^{m} \left(y_j - f\left(\mathbf{x}_j\right)\right) \left(-\left[\mathbf{x}_j\right]_i\right) \tag{3.53}$$

Substituindo o resultado (3.53) na expressão (3.52) obtém-se a regra de actualização do vector peso dado pela descida do gradiente:

$$\Delta w_i = \eta \sum_{j=1}^{m} \left(y_j - f\left(\mathbf{x}_j\right)\right) \left(-\left[\mathbf{x}_j\right]_i\right) \tag{3.54}$$

Em resumo, o algoritmo da descida do gradiente para o treino de perceptrões com unidades de processamento linear executa o seguinte procedimento. Primeiramente, determinam-se as saídas do perceptrão para todos os exemplos de treino. Seguidamente, calculam-se os incrementos Δw_i para cada uma das componentes do vector peso pela fórmula (3.54) e actualizam-se as componentes do vector peso w_i

adicionado-lhes os respectivos incrementos Δw_i. Finalmente, repete-se este ciclo até os incrementos serem tão pequenos quanto se queira. No exemplo ilustrado na *figura* 3.8 em que a superfície de erro contém um único mínimo global a convergência é garantida para um vector que minimiza o erro, independentemente de os exemplos serem ou não serem separáveis linearmente, desde que seja usado para passo de aprendizagem η um valor baixo, tipicamente $0.05 \leq \eta \leq 0.75$ Bishop (1995), Haykin (1999), Marques_de_Sá (2001a).

3.4.4 Perceptrão de Multicamada (PMC)

Um perceptrão de multicamada normal é composto por um encadeamento de perceptrões de uma única camada (isto é, perceptrões que podem ter várias unidade de processamento e igual número de saídas). Existe uma camada de nodos de entrada, uma camada de nodos de saída e uma ou mais camadas intermediárias. As camadas interiores designam-se "*camadas escondidas*" uma vez que elas não são directamente observáveis a partir das entradas ou saídas do *PMC*. Cada nodo tem uma resposta dada por uma função $f(\langle \mathbf{w}, \mathbf{x} \rangle)$, em que $\mathbf{x}$ é o vector de activações proveniente das saídas da camada anterior, $\mathbf{w}$ é um vector de pesos, e f é uma função limitada não decrescente e não linear como por exemplo uma sigmóide. Um dos pesos actua como polarização (*bias*) em virtude de uma conexão a uma entrada de valor constante. Os nodos em cada uma das camadas estão completamente conectados a todos os nodos da camada precedente e da camada seguinte. Não existem conexões entre unidades da mesma camada, nem conexões entre uma camada e uma camada mais atrás que não seja a precedente, assim como não existem "atalhos" que passam á frente conexões intermediárias. Esta é a estrutura mais comum, embora seja possível usar-se rectro-alimentação nas chamadas redes neuronais recorrentes, as quais não são abordadas nesta tese.

Por exemplo, num PMC com n entradas, p unidades de processamento na camada escondida, funções de activação sigmoidais $\sigma(\cdot)$ e k saídas reais, tem a forma canónica:

- A saída da $j-ésima$ unidade escondida, z_j é obtida tomando uma combinação linear das entradas x_i e passando-a através da função de activação:

$$z_j = \sigma\left(\sum_{i=0}^{n} w_{ji}^{(1)} x_i\right)$$

em que $w_{ji}^{(1)}$ indica o peso da conexão entre o perceptrão j da camada (1) e a entrada i.

- As saídas da rede y_k podem obter-se:

$$y_k = \sigma\left(\sum_{j=0}^{p} w_{kj}^{(2)} z_j\right)$$

em que $w_{kj}^{(2)}$ indica o peso da conexão entre o perceptrão j da camada (1) e o perceptrão k da camada (1), neste exemplo é a camada de saída.

- A função completa da rede é dada por:

$$y_k = \sigma\left(\sum_{j=0}^{p} w_{kj}^{(2)} \sigma\left(\sum_{i=0}^{n} w_{ji}^{(1)} x_i\right)\right)$$

- Ou, em notação matricial:

$$\mathbf{y} = \sigma\left(\mathbf{w}^{(2)} \sigma\left(\mathbf{w}^{(1)} \mathbf{x}\right)\right)$$

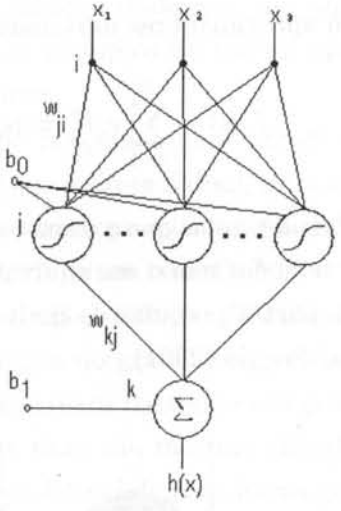


Figura 3.9: PMC com três entradas e uma camada escondida

A *figura* 3.9 mostra um PMC com três entradas, uma camada escondida de elementos de processamento sigmoidais e uma saída linear. Notar que w_{ji} é um peso que afecta a conexão que chega ao nó j proveniente do nó i, e que b_i , $i \in \{0, 1\}$, são os valores dos limiares, ou polarizações (*bias*) de cada uma das camadas. A saída da rede exprime-se por:

$$\mathbf{y} = \sigma\left(\mathbf{w}^{(2)} \left(\mathbf{w}^{(1)} \mathbf{x}\right)\right)$$

A função $\sigma(\cdot)$ é em geral um membro da família de funções sigmoidais, por

exemplo, a família $\sigma(\mathbf{x}) = \tanh(\beta \mathbf{x})$.

3.4.5 Capacidade de representação

Se uma rede PMC tiver ma única camada, as unidades de processamento forem funções em degrau de Heaviside, as variáveis de entrada forem reais e tiver uma única saída, então, ela usa um hiperplano para separar o espaço de entrada em dois subconjuntos (*figura* 3.10). Se uma rede tiver duas camadas escondidas, e as restantes características forem iguais à rede anterior, então, já é capaz de criar regiões de decisão convexas em que a unidade de saída faz a conjunção das saídas das unidades da camada escondida. Se a rede tiver três camadas escondidas é capaz de representar regiões de decisão arbitrárias.

Os *PMCs* tal como as redes de *FBR* , e outros métodos de aproximação clássicos, como por exemplo as séries de Fourier, os desenvolvimentos polinomiais e as técnicas de aproximação "*ridge*", são casos especiais das técnicas de classificação e aproximação que permitem aproximar funções contínuas com a aproximação que se desejar, daqui se designarem aproximadores universais:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i H(\mathbf{x}, p_i)$$

O teorema de Novikoff ao estipular que o número de erros não depende da dimensionalidade, condição que também tem o seu equivalente nas MVSs, representa uma direcção muito promissora para a pesquisa de limites de generalização de esquemas de aprendizagem Verri and Poggio (2001).

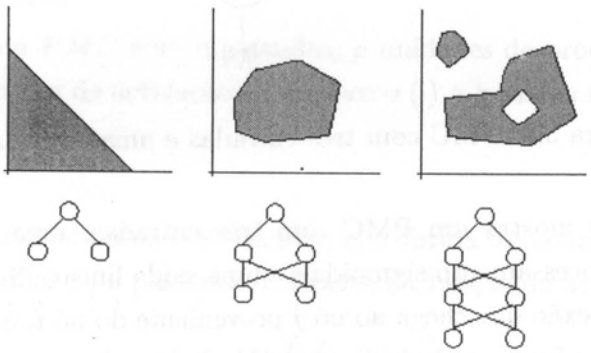


Figura 3.10: Capacidade de representação de três arqutecturas de redes se as unidades de processamento forem em degrau de Heaviside.

3.5 Sumário

As MVS classificadoras maximizam a distância entre a superfície de decisão (hiperplano) e as duas classes de dados, isto é, maximizam a margem de separação entre as classes. O hiperplano de decisão é determinado apenas pelos vectores de suporte. A ideia genial das MVS consiste em usar uma função não linear que transforma os exemplos de treino para um espaço linear de características, com grande dimensionalidade podendo ser infinita daqui se designar por espaço de Hilbert.

A utilização de um kernel K a substituir o produto interno no espaço de entradas $\mathbb{X}$ corresponde a transformar os dados para um espaço de produto interno $\mathbb{H}$ dado pela transformação:

$$\Phi : \mathbb{X} \longrightarrow \mathbb{H},$$

e a determinar-se em $\mathbb{H}$ o produto interno, pela fórmula:

$$K\left(x, x'\right) = \left\langle \Phi\left(x\right), \Phi\left(x'\right) \right\rangle$$

. Desta forma todos os cálculos são feitos implicitamente. O preço que se paga por esta elegância é que a solução dada por um algoritmo kernel obtém-se pelo desenvolvimento em série dos exemplos de treino, após estes serem transformados para o espaço de características.

A aproximação dada por máquinas com *kernel* utiliza a formulação da programação quadrática, usando-se sempre o dual, uma vez que a construção do primal requer que o espaço de características seja necessariamente conhecido, o qual pode inclusivamente ter uma número quase "infinito" de componentes.

Nas redes neuronais dos tipos FBR e PMC a optimização do funcional de custo pode ser obtida pela regra do gradiente e rectropropagação, a qual implica a resolução de problemas não convexos contendo um grande número de mínimos locais possíveis e cuja convergência para um mínimo global é sempre problemática. Nos métodos de aprendizagem com *kernel* de uma forma geral, e nas MVS em particular descritas neste capítulo, a optimização processa-se por programação matemática, os problemas de optimização incidem sobre conjuntos convexos e as soluções são globais, exactas e únicas.

3.6 Referências bibliográficas e notas históricas

O método de estimação com vectores de suporte de funções indicatrizes e funções reais, as demonstrações de muitos dos teoremas mais importantes encontram-se descritas pelo pena do seu criador, nos capítulos décimo e décimo primeiro de Vapnik (1998). Para uma introdução às MVS para reconhecimento de padrões consultou-se

Burges (1998) e Campbell (2000). Existe um consenso de que o algoritmo de treino de classificadores de margem óptima foi pela primeira publicado por Boser et al. (1992). O algoritmo das MVS para conjuntos não separáveis com margem suave é devido a Cortes and Vapnik (1995). A descrição do método das MVS adoptada nesta tese tem por base o primeiro capítulo, "*Introdução à aprendizagem por vectores de suporte*", de Schölkopf et al. (1999), o sexto capítulo de Cristianini and Shawe-Taylor (2000) e alguns desenvolvimentos introduzidos em Schölkopf (1997), Smola (1998), e, mais recentemente, em Scholkopf and Smola (2002). Obviamente que se pode encontrar em Haykin (1999) uma abordagem didática às MVS, pp.319-349.

A melhor fonte de informação sobre a teoria da regularização parecem ser as imensas publicações sobre este tópico produzidas nos anos 1990-2000 no MIT-CBLC (*Center for biological and computational learning*). Em particular, serviram de inspiração nesta tese muitos dos materiais e bibliografia disponibilizados nas páginas da rede internet (web) para apoio aos cursos de 2000, 2001 e 2002 devidos a Verri and Poggio (2001) e Poggio et al. (2002). Em particular, são de relevância os textos Girosi et al. (1995), Niyogi and Girosi (1996), Evgeniou et al. (2000b), Evgeniou (2000). Uma abordagem didática sobre este tema pode ser encontrada em Bishop (1995) (pp.338-346) e, mais recentemente, em Cherkassky and Mulier (1998) e Kecman (2001).

A perspectiva dos algoritmos primal e dual do perceptrão é inspirada em Cristianini and Shawe-Taylor (2000) e Bennett and Bredensteiner (1998). Os fundamentos da teoria da optimização com restrições e programação quadrática podem ser encontrados emTavares and Correia (1986), Mangasarian (1994), Fletcher (2000) ou Luenberger (1989).

A bibliografia sobre espaços de características induzidos por Kernel compreende os clássicos Mercer (1909), Aronszajn (1950), Courant and Hilbert (1953), e os textos modernos especificamente motivados pelas MVS, Wahba (1990), Guyon et al. (1993), Wahba (1999), Burges (1999).

A bibliografia sobre redes neuronais artificiais, PMC e redes FBR é muito extensa, tendo sido em particular seguidos nesta tese os textos deLippmann (1987), Hertz et al. (1991), Bishop (1995), Ripley (1996), Haykin (1999), Principe et al. (2000a), Marques (1999) e Marques_de_Sá (2001a). Desenvolvimentos sobre aceleração no treino das redes neuronais podem ser encontrados em Almeida (1987) e Silva (1998). Textos de síntese contendo os tópicos com mais interesse prático desenvolvidos na década de 90 no domínio das redes supervisadas podem obter-se em Reed and Marks (2000) e Duda et al. (2001).

O algoritmo sobre o perceptrão foi originalmente discutido por Rosenblatt (1959). A noção do ajustamento da grandeza das conexões em neurónios biológicos com base nas correlações entre as entradas e saídas foi inicialmente investigado por Hebb

(1949) quando ao tentar explicar como uma rede de células vivas de cérebro e adaptavam a estímulos diferentes, sugeriu que as conexões usadas com mais frequência deveriam tornar-se mais fortes, enquanto que as que as que não eram usadas enfraqueceriam. Uma obra clássica acerca das potencialidades e limitações dos perceptrões simples é devida a Minsky and Papert (1969), a qual gerou um aparentemente inexplicável abandono da investigação em neuro-computação até aos princípios da década de 1980. O algoritmo de treino da rectropropagação para redes de multicamada e unidades sigmoidais teria sido descoberto e redescoberto por Werbos (1974) e Rumelhart et al. (1986). A propriedade de aproximação universal das redes neuronais foi demonstrada de muitas formas, nomeadamente por Cybenko (1989).

Uma aplicação à análise automática do sono em que se refere o problema da determinação da arquitectura de redes de multicamada é devida a Lengellé and Schaltenbraud (1997).

Capítulo 4

Biometria Fetal, Análise Exploratória de Dados e Regressão

Introdução

Este capítulo tem por objectivo apresentar o problema da previsão do peso fetal, as principais estatísticas de um conjunto de dados ecográficos, algumas inferências obtidas a partir de uma análise exploratória destes dados, e alguns tópicos sobre o modelo de regressão linear, não-linear e bayesiano. Estes temas que como se verá irão influenciar a concepção da solução do problema da previsão do peso fetal baseada em modelos não-lineares e semi-paramétricos de rede neuronal que se apresentam nos capítulos seguintes.

Na secção 4.1 faz-se uma revisão de literatura sobre biometria fetal aplicada ao diagnóstico do crescimento fetal referindo-se trabalhos anteriores sobre modelos estimadores do peso fetal baseados quer na regressão linear generalizada, quer na regressão não-linear por rede neuronal. Conclui-se esta secção com um descrição da metodologia usada na colheita de dados.

Na secção 4.2 faz-se uma análise exploratória da matriz de dados: tipo de experiência, análise multivariável das correlações, modelização simples do peso fetal, análise das componentes principais, análise exploratória das relações entre as variáveis, o efeito dos erros nas medidas (homocedasticidade versus heterocedasctcidade). Esta secção apresenta uma perspectiva de análise exploratória de dados que se baseia na utilização de um package estatístico corrente com o qual se procurou compreender melhor as relações entre as variáveis, quer para o conjunto completo de casos (414 casos), quer para um subconjunto de casos que se caracterizam pelo seu peso fetal baixo (46 casos).

Na secção 4.3 revêem-se alguns tópicos específicos do modelo de regressão, tais como, a esperança condicional, variância condicional, modelo de regressão multilinear clássico, modelo de regressão linear generalizado, modelo não-linear

Na secção 4.4 apresenta-se sumariamente os conceitos de modelo de regressão bayesiana, máxima verosimilhança, aprendizagem bayesiana, assim como o conceito de fronteira de decisão que surge naturalmente quando se tentar fazer uma interpretação dos diagramas de dispersão de duas ou mais variáveis categorizadas segundo classes de pesos fetais.

Este é um capítulo chave para a génese e desenvolvimento de ideias que tiveram uma grande influência na abordagem do problema da previsão do peso fetal.

4.1 Biometria fetal pelo método ecográfico

Desde a introdução da ecografia nos finais de 1950 que esta tem vindo a ser um instrumento de diagnóstico muito útil em obstetrícia. Presentemente usam-se equipamentos conhecidos como "*scanners*" em tempo real, com os quais é possível ter-se uma imagem contínua do feto em movimento num ecrã de computador. Usa-se para este efeito ondas sonoras de muito alta frequência, entre 3.5 a 7.0 MHz, emitidas por um transdutor colocado em contacto com o abdómen materno e que se desloca, como se fosse uma lanterna, para observar uma dado aspecto do útero. Uma sequência repetida de feixes de ultra-sons fazem varrimentos no feto, em fatias delgadas, que são reflectidas e regressam ao mesmo transdutor. Os tecidos muito densos, como por exemplo os ossos, reflectem os ultra-sons fortemente (diz-se que são "*ecógenos*"), aparecendo a branco no ecrã. A informação obtida a partir das diferente reflexões é reconstituída numa imagem que se apresenta num ecrã. As imagens não são fotografias uma vez que resultam de algoritmos de reconstrução de imagens que podem ser sujeitos a diversas regulações dos parâmetros electrónicos do ecógrafo. As medidas que se podem obter por este processo constituem o ponto de partida para uma avaliação da idade gestacional, do tamanho e crescimento do feto, e outros diagnósticos relevantes para o bem estar do feto e da mãe.

Em Sarramon and Grandjean (1998) (p.27) as autoras referem que em 1976 apenas 11% das mulheres grávidas francesas faziam pelo menos uma ecografia, em 1981 esta proporção passou para 82%, com mais de dois exames ecográficos em 50% dos casos. Em 1995, 90% das mulheres grávidas francesas fizeram pelo menos duas ecografias, sendo apenas 1% a percentagem das que não fizeram nenhuma ecografia.

Está amplamente aceite na comunidade de especialistas em ginecologia e obstetrícia fazer-se a estimação do peso fetal em grávidas " normais" a partir de fórmulas de regressão linear generalizada. Uma destas fórmulas mais correntes utiliza as medidas ecográficas do perímetro abdominal PAB e do comprimento do fémur CFE,

tendo sido desenvolvida nos anos 80 por Hadlock et al. (1985) :

$$\log(PFE) = 1.304 + 0.05281 \times (PAB) + 0.1938 \times (CFE) - 0.004 \times (PAB) \times (CFE)\,\mathrm{g} \tag{4.1}$$

Por exemplo, para um caso em que se observaram os seguintes valores de $PAB = 32.4\,\mathrm{cm}$, $CFE = 6.7\,\mathrm{cm}$, e em que o peso do recém nascido foi $3260\,\mathrm{g}$, esta fórmula apresenta a seguinte estimativa:

$$\begin{aligned} PFE_{HAD} &= 10^{1.304+0.05281\times(32.4)+0.1938\times(6.7)-0.004\times(32.4)\times(6.7)} \\ &= 2787\,\mathrm{g} \end{aligned}$$

Uma outra fórmula devida a Shepard (1982) utiliza também duas medidas preditivas, o PAB e diâmetro biparietal DBP:

$$\log(PFE_{HAD}) = 3 - 1.7492 + 0.166 \times (DBP) + 0.046 \times (PAB) - 0.002646 \times (PAB) \times (DBP)\,\mathrm{g} \tag{4.2}$$

Por exemplo, para o mesmo caso anterior $PAB = 32.4\,\mathrm{cm}$ e $DBP = 9.5\,\mathrm{cm}$ esta fórmula apresenta a seguinte estimativa:

$$\begin{aligned} PFE_{SHE} &= 10^{3-1.7492+0.166\times(9.5)+0.046\times(32.4)-0.002646\times(32.4)\times(9.5)} \\ &= 3190\,\mathrm{g} \end{aligned}$$

Um texto recente D'Ottavio and Mandruzzato (1998) recomenda uma outra fórmula devida a Hadlock, mas, diferente da anterior, em que se utilizam três medidas preditivas, o PAB, o comprimento do fémur CFE e o perímetro cefálico PCE :

$$\log(PFE) = 1.326 + 0.0107 \times (PCE) + 0.0438 \times (PAB) + 0.158\,(CFE) - 0.00326 \times (PAB) \times (CFE)\,\mathrm{g} \tag{4.3}$$

Por exemplo, para o mesmo caso anterior, $PAB = 32.4\,\mathrm{cm}, CFE = 6.7\,\mathrm{cm}$ e $PCE =$

34.0 cm esta fórmula apresenta a seguinte estimativa:

$$PFE_{DOT} = 10^{1.326+0.0107\times(34.0)+0.0438\times(32.4)+0.158(6.7)-0.00326\times(32.4)\times(6.7)}$$
$$= 2883\,\text{g}$$

Uma fórmula simples recomendada como sendo precisa para fetos cujo pesa seja inferior a 2.5 kg, devida a Woo and Wan (1986), utiliza três medidas preditivas, PAB, CFE e DBP:

$$PFE_{WOO} = 1.4 \times (DBP) \times (CFE) \times (PAB) - 200 \tag{4.4}$$

Por exemplo, para o mesmo caso anterior $PAB = 32.4\,\text{cm}, CFE = 6.7\,\text{cm}$ e $DBP = 9.5\,\text{cm}$ esta fórmula apresenta o seguinte resultado:

$$PFE_{WOO} = 1.4 \times (9.5) \times (6.7) \times (32.4) - 200 = 2687\,\text{g}$$

A título de curiosidade refere-se que experiências com redes neuronais dos tipos PMC e FBR e máquinas de vectores de suporte MVS (*Torch*), todas treinadas com 210 casos e incluindo 5 medidas ecográficas como entradas, estimaram este caso, que não pertencia ao conjunto de treino, com os pesos fetais de 3297 g, 3188 g e 3194 g, respectivamente. Aparentemente este caso foi fácil para as redes neuronais, que quase acertaram nos 3260 g do peso real, outros casos não o serão tanto assim. No entanto tal como se verá neste e capítulos seguintes, em média, as redes neuronais têm um poder preditivo superior a estas fórmulas.

Existem muitas outras fórmulas, por exemplo, num estudo multicêntrico Chauhan et al. (1998) citam-se 26 fórmulas. Este estudo multicêntrico tinha por objectivo a comparação das precisões (*accuracies*) das estimativas dos pesos fetais por ecografia entre crianças nascidas com 24 a 34 semanas de idade gestacional em três centros hospitalares, cada um tendo feito uma recolha de dados em 171 casos com aleatoriedade controlada. Tal como este estudo demonstra a previsão do peso pré-natal constitui uma componente importante da prática clínica em obstetrícia e ginecologia, uma vez que os fetos que não se tenham tido um desenvolvimento adequado podem ter uma taxa de mortalidade perinatal mais elevada e, nomeadamente, apresentam um risco mais elevado de virem a ter problemas neurológicos Sereno et al. (2000a).

Em D'Ottavio and Mandruzzato (1998) (p.1215) refere-se a existência de situações (como por exemplo diabetes e anterioridade de parto por cesariana) em que tem importância clínica a existência de pelo menos uma boa estimativa do peso fetal num dado momento. As fórmulas correntes que utilizam diversos parâmetros ecográficos, apesar de numerosas, não são infalíveis, podendo ocorrer erros da ordem $\pm 15\%$, mesmo usando os melhores métodos. Para além disto todas as fórmulas

tendem a sobreestimar o peso no casos de pesos fetais baixos (isto é, $\leq 1500\,\mathrm{g}$) e subestimar os pesos fetais elevados (isto é., $\geq 4000\,\mathrm{g}$).

Em Pereira-Leite and Bernardes (1998) (p.1290) refere-se que o diagnóstico antenatal de macrosomia fetal constitui um tema clínico e médico-legal importante, porque pode detectar pelo menos duas situações teoricamente previsíveis: o trauma feto-maternal devido a factores mecânicos e a morte perinatal devido a alterações metabólicas. E, acrescentam os autores (p.1292), que apesar de tantos desenvolvimentos tecnológicos ocorridos nas últimas décadas em obstetrícia, o diagnóstico antenatal ainda continua a ser significativamente impreciso, e concluem sugerindo ser necessário mais investigação no que diz respeito à colheita de dados e à sua análise, tendo em vista a estimativa do crescimento fetal. Esta investigação deve incluir a utilização, por exemplo, de imagiologia por ressonância magnética e modelos matemáticos não-lineares, isto é, redes neuronais. Na opinião dos autores (p.1291), ao contrário dos modelos de regressão os modelos de rede neuronal podem considerar diferentes padrões de crescimento de diferentes grupos de fetos (por exemplo, de crescimento retardado, prematuro, macrosómico).

As estimativas obtidas por palpação abdominal e pela produção de hormona fetal mostraram-se ter uma precisão reduzida. Herrero (1999) compararam as estimativas do peso fetal do seu próprio feto feitas pela parturiente com as estimativas do médico baseadas na palpação abdominal tendo encontrado não existirem diferenças entre ambas as estimativas em termos de do erro absoluto médio.

As fórmulas tradicionais de estimação de peso fetal dadas pelas expressões (4.1), (4.2), (4.3) e (4.4) utilizam pelo menos duas medidas ecográficas, das quais o perímetro abdominal PAB é sempre utilizado, e servem para satisfazer a necessidade de obter rápida e facilmente um estimativa do peso fetal. Warsof and A (1977) num estudo pioneiro em 85 parturientes, cujos fetos apresentavam idades gestacionais compreendidas entre 17 e 41 semanas, sujeitas a medidas ecográficas 48 horas antes do nascimento, em que se fez a análise estatística e a regressão linear múltipla mostrou que : (i) o peso fetal é uma função logarítmica dos parâmetros corporais e o perímetro abdominal tem a melhor correlação com $\log_{10}$ do peso fetal; (ii) a regressão linear utilizando duas medidas ecográficas (o PAB e o diâmetro biparietal DBP) apresenta um desvio padrão de 106 g por cada kg de peso fetal .

A utilização de métodos de rede neuronal artificial começa a aparecer referida a a partir dos princípios de 1990. Farmer et al. (1992), por exemplo, referem que o peso fetal obtido pelas fórmulas clássicas dos métodos de regressão, com base em medidas ecográficas, em fetos com suspeitas de macrosomia apresenta um erro de previsão aproximadamente de 10%, o qual é clinicamente inaceitável. Estes autores experimentaram modelos de rede neuronal treinados com 146 casos, em que se omitem alguns cuidados recomendados pela teoria da regularização, que se descreve no

capítulo 3 desta dissertação, assim como apresentavam um controlo da complexidade dos modelos insuficiente. Em testes de estimação do peso fetal em 102 casos de fetos macrosómicos este estudo revela que o erro relativo médio teria sido de 4.7%, sendo estas estimativas mais significativas estatisticamente ($p = 0.001$) do que as que foram obtidas com os modelos de regressão. Estes investigadores chamam a atenção para o facto de as previsões do modelo neuronal se degradarem significativamente quando não existiam casos suficientes para o treino, o que se verificava, nos casos macrossómicos deste estudo, nas duas faixas de pesos extremos (inferiores a 3800 g e superiores a 4800 g).

Num outro estudo com modelos clássicos Chauhan et al. (1998) comparou as precisões das estimativas dos pesos fetais obtidas com medidas ecográficas, em fetos com 24 a 34 semanas de idade gestacional, em três centros hospitalares e concluiu que "*as estimativas ecográficas de peso fetal em crianças nascidas antes do tempo, tal como foram obtidas utilizando 26 fórmulas caracterizam-se por ampla gama de valores de precisão, ... [e] para muitas destas fórmulas as precisões das estimativas diferem bastante de um centro para outro.*"

Para terminar esta subsecção acrescentam-se ainda alguns comentários breves sobre a medida ecográfica, em modo *Doppler,* do fluxo sanguíneo. A medida IRU é o *índice de resistência umbilical,* o qual foi medido e constava em todos os casos da base de dados, obtida segundo as recomendações do protocolo que a seguir se refere. O índice de pulsatilidade IPU estava omisso em muitos casos. Em Mandruzzato et al. (1994) (p.399-403), refere-se que estes índices permitem obter uma medida da velocidade do fluxo sanguíneo, podendo ser usados massivamente no diagnóstico de atraso do crescimento intrauterino ($IUGR$, "*intrauterine growth retardation*") ou hipoxémia. No caso de uma gravidez normal referem estes autores que se verifica existir uma carência de estudos sobre a anormalidade para cada idade gestacional. A interpretação que se sugere nesta dissertação sobre este tema é de que, para um dado especialista em obstetrícia, um dado aparelho de ecografia, e uma técnica de medida sistematicamente reproduzida, os índices IPU ou IRU poderão constituir um melhores preditor do peso fetal do que as sua fracas correlações com o peso fetal, $-.43$ e $-.40$, tal como se mostra na tabela 4.6, permitem acreditar.

4.1.1 Metodologia da colheita de dados

O conjunto de dados foi recolhido com base num protocolo que tinha sido definido anteriormente ao início do projecto de investigação da estimativa ecográfica do peso fetal com modelos de rede neuronal. Este projecto insere-se no âmbito de um estudo multicêntrico que compreendeu vários hospitais Portugueses, nos quais se procedeu à colheita de dados em conformidade com este protocolo, o qual depois de analisado por obstetras e investigadores do INEB (Instituto de Engenharia Biomédica da Fa-

culdade de Engenharia do Porto), foi aceite como realista para a prática clínica dos especialistas de obstetrícia que iam participar na colheita de dados. Os objectivos deste protocolo eram a definição dos principais parâmetros ecográficos, critérios de inclusão de parturientes e métodos de medição:

> "**1. Introdução e objectivos**. Este projecto tem como objectivo desenvolver um método de redes neuronais, utilizando parâmetros ecográficos (perímetro cefálico, diâmetro biparietal, perímetro abdominal, comprimento do fémur, índice de Pourcelot e índice de pulsatilidade), capaz de estimar o peso fetal de forma mais precisa que as actuais fórmulas de regressão linear, que apresentam um grau de erro considerável, sobretudo nas casos de muito alto ou baixo peso.
>
> **2. Material.** Os critérios de inclusão são: (1) Todas as grávidas que derem consentimento [informado] ressalvando-se o disposto no ponto 5; (2) Gravidez única; (3) Dados ecográficos colhidos nos 7 dias anteriores ao parto (procurar-se-á que o intervalo entre a colheita de dados ecográficos e o parto seja o mais curto possível); (4) Grávidas com idade gestacional calculada de acordo com a data da última menstruação nas mulheres com ciclos regulares que não tenham feito contracepção hormonal nos últimos três meses. Nos casos em que a idade gestacional ecográfica difere da amenorreia de 7 dias até às 13 semanas, ou 10 dias até 22 semanas considera-se a idade gestacional ecográfica; (5) Ausência de malformações fetais nos segmentos a medir.
>
> **3. Métodos**. O diâmetro biparietal será medido da tábua externa à tábua externa oposta. O perímetro cefálico e perímetro abdominal serão medidos pela parte externa do segmento a medir usando ou o método da elipse, ou o do contorno (*trace*). O comprimento do fémur é medido na diáfise. O índice de Pourcelot e índice de pulsatilidade serão medidos na ansa do cordão mais próxima do topo fetal, sendo as medidas efectuadas no mesmo ciclo cardíaco. Será preenchida uma folha para cada grávida em que constarão a sua identificação, os dados ecográficos e os dados do recém-nascido."

A ficha de dados compreendia, para além das datas do exame ecográfico e do nascimento da criança, três blocos de dados: (*i*) Maternos: idade, peso altura, idade gestacional em semanas e dias; (*ii*) Ecográficos: perímetro cefálico, diâmetro biparietal, perímetro abdominal, comprimento do fémur, índice de pulsatilidade e de resistência umbilical; (*iii*) Sobre o recém-nascido: peso (*gramas*), comprimento (cm) e perímetro cefálico (cm).

Em finais de 1999 existia uma base dados com 280 casos, a qual permitiu realizar as primeiras experiências e, em Junho de 2000, o número de casos cresceu para 354 com os quais se procedeu aos ensaios comparativos entre as performances de redes neuronais dos tipos PMC e FBR, cujos resultados que foram apresentados em reuniões científicas internacionais e se relatam no quinto capítulo. A base de dados final ficou estabelecida em finais de 2000 compreendendo 414 casos, os quais serviram de base para as experiências com métodos de *kernel,* implementados em modelos de máquina de vectores de suporte (MVS) que se relatam no quinto capítulo desta dissertação.

4.2 Análise exploratória de dados

4.2.1 Tipo de experiência

Começando pela definição de experiência, adoptou-se aquela que surge mais naturalmente no contexto de um estudo estatístico e do protocolo acima referido, podendo dizer-se, então, que esta consiste numa recolha de observações, quantitativas ou qualitativas, realizada por meio de operações bem definidas e sob condições também bem definidas, recolha esta que deve ser concebida de forma a permitir uma avaliação não subjectiva da existência e dos tipos de relações que possam existir entre os dados. O propósito da experiência consiste em transpor essas relações para um modelo, o mais simples possível, num processo de criação de hipóteses, ou verificação entre uma ou mais hipóteses alternativas Chatfield and Collins (1989) (p.34).

A recolha de um conjunto de dados para efeitos de construção de um modelo de regressão, segundo Neter et al. (1996) (p.329), varia muito com a natureza do estudo, o qual pode ser de quatro tipos: (*i*) Uma experiência controlada, (*ii*) uma experiência controlada com variáveis suplementares, (*iii*) um estudo observacional confirmatório ou (*iv*) um estudo observacional exploratório. O presente estudo parece satisfazer melhor as condições do terceiro e quarto tipo, a de um estudo *observacional*, uma vez que parece ter havido um controlo insuficiente sobre as variáveis explicativas. Por exemplo, dos 414 dados recolhidos apenas 39 casos se podem considerar apropriados para um estudo de fetos com peso fetal baixo (num estudo sobre as diferenças de pesos fetais entre crianças negras nascidas em África e E.U.A. e crianças brancas nascidas nos E.U.A. David and Collins (1997) considera pesos fetais baixos de risco os pesos inferiores a 2500 g). Adicionalmente, como não são referidas quaisquer cláusulas sobre aleatoriedade da amostragem, e a cláusula do protocolo 2.1, sobre o voluntariado da mãe, condição éticamente respeitável, mas do ponto de vista estatístico sugere precaução no que diz respeito à distribuição, dependência e heterogenidade da amostragem Spanos (1999) (p.13).

No que diz respeito ao carácter confirmatório, ou exploratório, do presente es-

tudo, parece que ele poderá ter algumas características que podem ser comuns a estas duas variantes. Estudo *confirmatório* na medida em que se usam medidas ecográficas como variáveis preditivas que já se sabe, em face de outros estudos prévios como os de Hadlock et al. (1985), D'Ottavio and Mandruzzato (1998), por exemplo, que são boas preditoras do peso fetal. Estudo *exploratório* na medida em que se testa a utilização de outras variáveis sobre as quais pouco se sabe quanto à sua contribuição para a previsão do peso fetal, tais como, por exemplo, os dados de ecografia *Doppler* dados pelo índice de resistência umbilical.

4.2.2 Análise multivariável

Os objectivos de uma análise de dados são, por um lado, a descoberta de regularidades que possam existir no comportamento de duas ou mais variáveis, por outro, a realização de testes rudimentares sobre hipóteses alternativas de associação entre duas ou mais variáveis, incluindo a determinação da existência de diferenças entre duas ou mais classes que se distinguem pelos seus perfis multivariáveis. Este tipo de análise tem uma considerável importância porque permite melhorar o conhecimento *a priori* de um dado problema, a necessidade de afinação de alguns parâmetros dos modelos multilinear de regressão, de rede neuronal, assim como a escolha do *kernel* para uma dada função de aproximação.

O tipo de análise multivariável que parece ter mais interesse consiste em determinar o grau de covariação que pode existir entre o peso fetal (eventualmente, o comprimento fetal) e as variáveis explicativas, biométricas maternas, idade gestacional e ecográficas. A análise das possíveis associações entre o peso fetal e estas variáveis deveria permitir "*explicar*" as variações do peso fetal de forma a encontrar-se uma função, ou fórmula, com a qual fosse possível obterem-se estimativas do peso fetal com um grau de confiança apropriado. Este problema pode ser interpretado segundo dois ponto de vista provenientes da: (*i*) Estatística computacional, como um problema de *regressão*, (*ii*) Matemática computacional, como um problema de interpolação, isto é, *aproximação* com ou sem ruído.

A matriz dos dados em bruto completa tem dimensões 414×17, isto é, 414 linhas, as quais correspondem aos casos ou exemplos e 17 colunas, que são as variáveis ou grandezas que foram medidas. Por exemplo, as colunas 8, 9, 10, e 11, da tabela 4.2 contêm dados sobre as cinco principais medidas biométricas obtidas por ecografia; as colunas 14, 15 e 16 contêm dados sobre os/as recém-nascidos/as.

Esta matriz pode ser vista como contendo 414 linhas constituídas por pontos no espaço $17-$*dimensional* dos casos individuais, ou como contendo 17 colunas consistindo em pontos no espaço $414-$*dimensional* das variáveis. Pode estar-se interessado em estudar as relações entre os indivíduos, ou entre as variáveis. Neste último caso comparam-se os vectores coluna, como, por exemplo, no cálculo dos coeficientes de

1	2	3	4	5	6	7
Caso	HOS	IDM	PEM	ALM	IGE	DIF
1	4	14	66	152	28	0
2	1	99	62	165	33	0
3	2	148	60	158	26	7
4	4	29	70	160	30	0
5	4	8	68	158	31	2
...	...	...	...	...	...	...
205	4	56	75	166	41	0
206	4	48	63	164	40	0
207	4	52	68	169	38	1
208	3	8	62	159	39	0
209	2	121	79	159	39	0
...	...	...	...	...	...	...
410	1	8	105	166	41	0
411	2	116	81	170	40	1
412	1	95	75	150	40	0
413	4	31	87	150	40	1
414	3	16	78	159	39	0

Tabela 4.1: Matriz dos Dados em Bruto (var- 1-7, extracto)

correlação. Pode pretender-se estudar comparativamente entre si os casos individuais, utilizando-se, neste caso e em geral, uma medida apropriada da "*distância*" entre os casos, os quais são definidos pelos vectores linha num espaço com a dimensão do número de variáveis, como por exemplo em análise de escalonamento multidimensional, ou em algumas variantes da análise factorial.

Correlações

Pode proceder-se a uma análise comparativa dos vectores coluna através da matriz dos coeficientes de correlação. Cada elemento desta matriz é essencialmente o produto interno dos dois vectores coluna das variáveis correlacionadas, depois destas terem sido previamente normalizadas de forma a que a média das observações em cada coluna seja nula e a variância unitária.

Cada coeficiente de correlação pode interpretar-se como sendo o cosseno do ângulo entre dois vectores coluna num espaço com a dimensionalidade do número de casos. Por exemplo, se neste espaço os vectores estiverem muito próximos, então, o "*ângulo*" entre eles é muito pequeno, o seu cosseno é muito próximo de $+1$ (ou, -1), valor indicativo de uma correlação positiva (ou, negativa).

1	8	9	10	11	12	13	14	15	16	17
Caso	DBP	PCE	PAB	CFE	IPU	IRU	PFE	COM	PCR	SEX
1	6.0	23.0	19.6	3.9	1.93	1.00	580	31.0	22.5	1
2	7.7	27.9	21.0	4.6	1.76	1.00	860	33.0	27.5	2
3	6.6	23.2	21.4	4.7	1.76	0.69	1010	37.0	25.0	0
4	6.9	25.8	23.7	5.0	1.76	0.69	1160	37.0	26.0	2
5	7.7	27.9	25.1	5.4	0.78	0.69	1280	41.0	31.0	2
...	...	...	...	...	...	...	...	...	...	...
205	9.5	35.4	32.5	7.4	0.72	0.48	3240	49.0	35.0	1
206	9.7	33.5	33.6	7.0	0.79	0.53	3240	45.0	35.0	2
207	8.6	32.3	34.9	7.3	0.85	0.53	3250	48.0	36.0	2
208	8.9	33.1	34.1	7.3	1.50	0.50	3250	50.0	33.0	0
209	9.4	33.7	32.6	7.1	1.00	0.67	3255	51.0	35.5	0
...	...	...	...	...	...	...	...	...	...	...
410	9.7	35.9	34.0	7.6	0.56	0.43	4380	50.0	37.0	1
411	9.6	32.2	35.7	7.7	0.80	0.57	4420	53.0	38.5	0
412	10.1	35.9	39.3	7.3	0.87	0.52	4515	53.0	38.0	1
413	10.2	36.1	40.3	7.5	0.65	0.49	4820	53.5	39.0	1
414	9.6	34.8	37.3	7.8	0.60	0.50	4900	50.5	37.0	0

Tabela 4.2: Matriz dos Dados em Bruto (var. 8-17, extracto)

Coluna	**Símbolo**	**Descrição**
1	Caso	Número do caso (ou da linha)
2	HOS	Hospital
3	IDM	Identificação da mãe
4	PEM	Peso da mãe
5	ALM	Altura da mãe
6	IGE	Idade gestacional
7	DIF	Dias de intervalo ecografias e o nascimento
8	DBP	Diâmetro biparietal
9	PCE	Perímetro cefálico
10	PAB	Perímetro abdominal
11	CFE	Comprimento do fémur
12	IPU	Indice de pulsatilidade
13	IRU	Indice de resistência umbilical
14	PFE	Peso (fetal) do recém-nascido
15	COM	Comprimento do recém-nascido
16	PCR	Perímetro cefálico do recém-nascido
17	SEX	Sexo (1-masculino, 2-feminino, 0-em falta)

Tabela 4.3: Variáveis recolhidas

Medida	Objecto	Variável	Símbolo
ecografia bidimensional	feto	diametro biparietal	DBP
ecografia bidimensional	feto	perímetro cefálico	PCE
ecografia bidimensional	feto	perímetro abdominal	PAB
ecografia bidimensional	feto	comprimento fémur	CFE
Modo doppler	feto	índice de resistencia	IRU
Modo doppler	feto	índice pulsatilidade	IPU
directa	recém-nascido	peso	PFE
directa	recém-nascido	comprimento	COM
directa	recém-nascido	perímetro cefálico	PCE

Tabela 4.4: Métodos de medida

	PEM	ALM	IGE	DIF
PEM	1.00			
ALM	.22	1.00		
IGE	.12	.03	1.00	
DIF	.03	.01	−.14	1.00
DBP	.22	.18	.61	−.09
PCE	.22	.16	.61	−.08
PAB	.27	.09	.60	−.07
CFE	.21	.15	.67	−.10
IPU	−.20	−.08	−.33	.07
IRU	−.16	−.10	−.30	.09
PFE	.34	.14	.65	−.07
COM	.27	.13	.66	−.05
PCR	.23	.11	.59	−.07

Tabela 4.5: Matriz das correlações (I)

	DBP	PCE	PAB	CFE	IPU	IRU	PFE	COM	PCR
PEM									
ALM									
IGE									
DIF									
DBP	1.00								
PCE	**.82**	1.00							
PAB	.63	.69	1.00						
CFE	.63	.68	.70	1.00					
IPU	−.37	−.40	−.39	−.43	1.00				
IRU	−.30	−.35	−.34	−.40	.68	1.00			
PFE	**.73**	**.74**	**.85**	**.76**	**−.45**	**−.40**	1.00		
COM	.69	.70	.76	.75	−.44	−.37	**.85**	1.00	
PCR	.73	**.78**	.71	.67	−.43	−.36	.79	.74	1.00

Tabela 4.6: Matriz das correlações (II)

4.2.3 Um modelo probabilístico simples para a densidade dos pesos fetais

Presentemente, vai proceder-se a uma *análise exploratória de dados* a qual segundo Tuckey (1977), Murteira (1993) e Martinez and Martinez (2002) tem como primeiro objectivo de análise a descoberta de aspectos ou padrões com maior interesse através da representação dos dados de maneira a aparecerem destacados esses aspectos ou padrões.

Uma das ferramentas mais utilizadas quando os dados não são muito numerosos é um gráfico *caule-e-folhas*. A figura 4.1 representa pesos à nascença (fetal) neste tipo de gráficos, em que os dígitos dominantes do caule representam a variável peso fetal do recém-nascido, expressos em hectogramas, 10^{-1} kg, e as folhas representam decigramas, 10^{-2} kg . Este gráfico permite obter-se um aspecto global da distribuição dos pesos fetais dos recém nascidos da amostra, por ordem crescente, os pesos mais baixos estão representados na parte de cima e os mais altos na de baixo.

A escolha do número de linhas deste gráfico parece ser razoável, uma vez que permite destacar a separação em subgrupos, como, por exemplo, o grupo dos pesos médios, suponhamos entre 3100 g *a* 3190 g , que corresponde aos pesos de moda mais elevada, assim como visualizar a heterogenidade dos dados, por exemplo, os pesos baixos distribuem-se por uma maior gama de valores do que os pesos altos.

Observam-se assimetrias, como se vê a curva apresenta uma assimetria negativa, desviada para a esquerda, mas pode-se admitir que uma amostra com dimensão maior se possa obter polígonos de frequência dos pesos dos recém-nascidos simétricas e mais próximas de uma distribuição gaussiana.

Um estudo com o título "*Serão os Bebés Normais?* Clemons and Pagano (1999)", em que se utilizava uma amostra de 4 017 264 certificados de nascimento de cidadãos norte-americanos, observavam-se 269 casos de pesos altos, superiores a 6.3 kg. Neste estudo observou-se ainda que o peso mais baixo encontrado foi de 227 g, o mais pesado foi de 8164 g, a percentagem dos casos que nasceram com uma idade gestacional entre as 37 e as 43 semanas foi de 87.1%, e que o peso à nascença médio destes últimos casos foi de 3432 g com desvio padrão de 482 g. Concluiu-se que para estes casos, de idade gestacional "*normal*", a distribuição dos pesos se ajustava bem a uma distribuição gaussiana. Fora daquela gama de idades gestacionais, os recém-nascidos prematuros e a diabetes gestacional eram provavelmente responsáveis pelos desvios significativos em relação à distribuição gaussiana, pelo que os autores recomendavam que nestes casos se procura um ajustamento com uma mistura de gaussianas.

Um outro estudo, com o título "*Diferenças de Peso à Nascença entre Crianças Negras nascidas nos E.U.A., Negras nascidas em África e Brancas nascidas nos*

E.U.A.", David and Collins (1997) refere que os pesos médios à nascença numa amostra com 43 322 crianças nascidas de mãe negra era de 3089 g, numa outra amostra constituída por 3135 crianças nascidas de mãe originária de África era de 3333 g, e numa terceira amostra de 44 046 crianças nascidas de mãe branca era de 3446 g. Reporta-se neste estudo que a incidência de pesos à nascença baixos (i.é. inferiores a 2500 g) para o primeiro grupo era 13.2%, para o segundo era 7.1% e, para o terceiro grupo era 4.3%. O estudo colocava a questão de saber até que ponto estes pesos mais baixos à nascença das crianças negras nascidas nos E.U.A. poderiam estar relacionados com factores genéticos ou com factores sociais. Concluíam os autores que, apesar de os dados disponíveis se deverem considerar insuficientes para se obter uma conclusão definitiva, as crianças negras nascidas de mãe proveniente de África apresentavam, não só um peso médio *superior,* mas também uma distribuição dos pesos diferente da que era apresentada pelas crianças negras que haviam nascido de mãe residente nos E.U.A.

Voltando ao estudo dos dados representados no histograma da figura 4.1 não se observam valores anormalmente altos ou baixos porque foram previamente excluídos da matriz de dados. Existem dois valores centrais em torno dos quais se concentram outros. Em torno de 1, 7 kg parece existir uma maior aglomeração de valores de pesos baixos, e em torno de 3, 9 kg uma maior aglomeração de pesos altos. Este gráfico ainda sugere alguma evidência para se considerar que os pesos fetais podem ser interpretados como pertencendo ao tipo *observacional*, uma vez que não se observam a característica de aleatoriedade típica de amostras com um grande número de casos como acontece nos dois estudos referidos acima.

Não quer dizer que os dados tenham qualquer defeito ou enviesamento de amostragem, mas tão somente que os modelos paramétricos clássicos podem não constituir bons modelos. Neste caso, os modelos não-paramétricos de rede neuronal poderão ser mais apropriados. Esta questão é mais importante do que pode parecer à primeira vista uma vez que a justificação de se tentar melhorar a estimativa dos pesos fetais, em alternativa aos modelos estatísticos clássicos, não é imediata, antes pelo contrário pode ser controversa e, nesta primeira fase do nosso trabalho, constituiu uma preocupação constante.

O problema da modelização de dados observacionais *versus* dados experimentais, tal como aparece abordado em Spanos (1999) em relação ao estado actual da investigação estatística, requer que se dê atenção aos seguintes três aspectos mais relevantes:

> (*i*) " A literatura tradicional nos últimos 50 anos foca-se quase exclusivamente em procedimentos e métodos que são apropriados para dados experimentais (inquérito e experimentação por amostragem)".

(*ii*) " O enquadramento tradicional da modelização não se tem afastado muito do que Fisher (1935) moldou para o delineamento experimental, mesmo com as reformulações dos anos 1940 e 1950 Fisher (1973), continuando a prevalecer a perspectiva de ajustamento de curvas na tradição dos mínimos quadrados."

(*iii*) " Estas duas tradições pressupõem um enquadramento da modelização no contexto da qual os dados observados se interpretam como uma realização de um fenómeno observável que se pode ver realmente como um sistema aproximadamente isolado (seja graças à intervenção humana seja à divina), Spanos (1999), p.xi)".

Este autor contrapõe a necessidade de "*se restabelecer um novo equilíbrio, através de um enquadramento da modelação estatística, o qual deverá ser mais apropriado aos dados observacionais ... [e] ser visto com um mesmo espírito e na perspectiva que existi na escola Biométrica (de Galton, Karl Pearson e Yule), que foi criada em finais do século XIX e princípios do século XX*". Acrescenta ainda este autor que "*este enquadramento alternativo torna-se absolutamente necessário porque os dados observacionais constituem a regra e não a excepção em campos como o da econometria.*"

Pode dizer-se que trazendo esta perspectiva para a biometria, tal como este investigador sugere, o problema da estimação do peso fetal, com a concepção experimental que foi referida, tende a enquadrar-se mais numa perspectiva *observacional* atendendo à evidência que se mostra existir ao longo deste capítulo

Não obstante as considerações que se acabam de fazer, por razões de simplicidade, e numa primeira abordagem, um dos modelos probabilísticos mais simples que se pode admitir para a lei de distribuição do peso fetal, representado pela variável X, baseia-se na família de densidades $p\left(\left(x;\theta\right),\theta\in\Theta\right)$ definidas num intervalo de valores de pesos fetais (em kg, seja de 1.0 *a* 4.5 kg), e um vector de parâmetros θ que pode variar num conjunto compacto Θ de valores admissíveis. A lei de densidade de distribuição $p\left(x;\theta\right)$ que se admite corresponder ao modelo mais simples é a lei normal, ou gaussiana, representada por $N_k\left(\cdot\right)$. Para este modelo especifica-se, em conformidade com a teoria desenvolvida em Spanos (1999), os seguintes elementos descritivos:

1. • **Mecanismo gerador estatístico**: $X_k=\mathbf{E}\left(X_k\right)+\varepsilon_k,\ k\in\mathbb{N}$
   • **Modelo probabilístico** dado por uma família de funções densidade de probabilidade (gaussianas):

$$\mathbb{N}=\left\{p\left(x|\mu,\sigma\right)\right\}=\left\{\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right\}\right\} \tag{4.5}$$

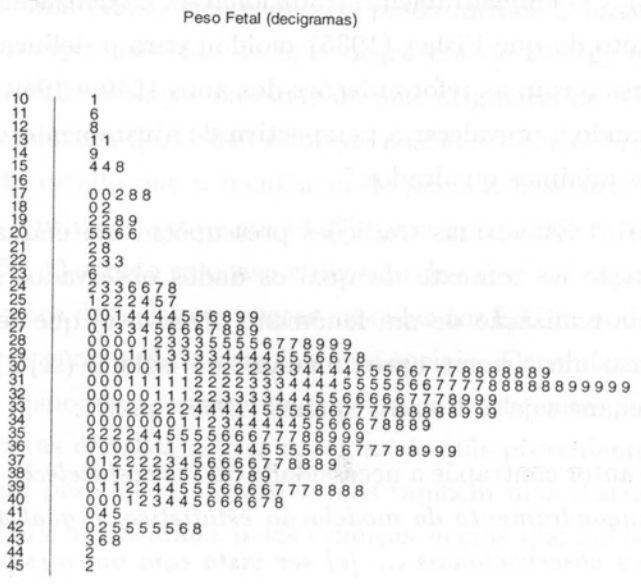


Figura 4.1: Diagrama de Caule-e-folhas de pesos à nascença (" fetais")

Por exemplo, podem usar-se como valores dos parâmetros na expressão 4.5, $\mu = 3.225\,\text{kg}$ e $\sigma^2 = 0.629^2 = 0\,.394\ \text{kg}^2$. Estes valores figuram na tabela das estatísticas descritivas e correlações com o PFE (Tabela 4.7, linha PFE), na qual se sumarizam os principais parâmetros, de *localização* (média) e *dispersão* (mínimo, máximo, e desvio padrão).

1. • **Modelo de amostragem**: $\mathbf{X} := (X_1, X_2, \ldots X_n)$ é uma amostra aleatória.

O modelo para ser válido tem de compreender também os seguintes três pressupostos probabilísticos:

$$[1]: \ X_k \sim N_k(\cdot)\,, \text{ para todos } k = 1, 2, \ldots n. \tag{4.6}$$

$$[2]: (X_1, X_2, \ldots X_n)' \text{ são independentes} \tag{4.7}$$

$$[3]: (X_1, X_2, \ldots X_n) \text{ são identicamente distribuídos} \tag{4.8}$$

A tabela 4.7 apresenta os principais valores de localização que se podem utilizar na construção de outros modelos simples.

Existem testes para uma verificação da inexactidão das especificações (*miss-pecification testing*), que permitem esclarecer sobre a adequação dos pressupostos probabilísticos básicos sobre a distribuição estatística, a dependência e a heterogenidade usada na definição de um modelo estatístico, expressões 4.6, 4.7 e 4.8, e que

	Min	Máx	μ	σ	$\rho_{(\cdot\ ,\ PFE)}$	Unidade
PEM	51.0	120.0	74.647	11.5627	.34	kg
ALM	130.0	180.0	160.145	6.3706	.14	cm
IGE	26.0	42.0	38.715	2.2303	.65	$\times 7$ d
DIF	0.00	7.0	0.862	1.3806	$-.07$	d
DBP	6.0	10.6	9.165	0.5649	.73	cm
PCE	23.0	37.6	32.783	1.8402	.74	cm
PAB	19.6	40.3	33.129	3.0696	.85	cm
CFE	3.9	8.0	7.135	0.5315	.86	cm
IPU	0.410	2.100	0.958	0.2422	$-.45$	
IRU	0.300	1.360	0.605	0.0964	$-.40$	
PFE	580	4900	3224.918	628.5639	1.00	g
COM	31.0	54.0	48.589	2.9462	.85	cm
PCR	22.5	39.0	34.300	1.9769	.79	cm

Tabela 4.7: Estatísticas descritivas e Correlações com o PFE

estão subjacentes a um dado modelo estatístico postulado com base num conjunto de exemplos disponíveis $\mathbf{x} = (x_1, x_2, \ldots, x)$ (Spanos (1999) p.729).

Gráficos de probabilidade normal. Utilizou-se este tipo de gráficos para avaliar se a proveniência de alguns dados será uma distribuição normal. Para testar se a variável regressanda peso fetal do recém-nascido obedecerá a uma lei normal usou-se uma função específica do MATLAB ('*normPlot* (peso fetal)').

O gráfico que se obteve representado na figura 4.2 tem três elementos. Os sinais + mostram a probabilidade empírica *versus* o valor dos dados para cada caso da amostra. A recta a traço contínuo, a qual não é visível nalgumas partes deste gráfico porque está sobreposta aos símbolos + aglomerados na parte central, passa no percentis 25 e 75 dos dados e representa um ajustamento linear robusto (isto é, insensível aos extremos da amostra). A linha a traço interrompido prolonga a linha a traço contínuo para os extremos da amostra.

A escala do eixo dos y não é uniforme, uma vez que representa probabilidades que vão de zero a um. A distância entre as marcas da escala no eixo dos y está em correspondência com a distância entre os quantis de uma distribuição normal. Os quantis estão mais próximos na vizinhança da mediana (probabilidade= 0.5) e alargam-se simetricamente à medida que se afastam desta. Comparando a distância vertical de parte inferior do gráfico até à probabilidade 0.25 com a distância de 0.25 até 0.50, verifica-se esta desigualdade. O mesmo se verifica se compararmos a distância do topo do gráfico até à probabilidade 0.75 com a distancia de 0.75 até 0.50.

Como todos os pontos de dados que estão assinalados com o símbolo + caem na vizinhança da recta a traço contínuo, entre as probabilidade 0.05 e 0.95, **o pres-**

suposto de normalidade é razoavelmente aceitável, apesar de se verificarem desvios mais acentuados na extremidade inferior dos pesos baixos.

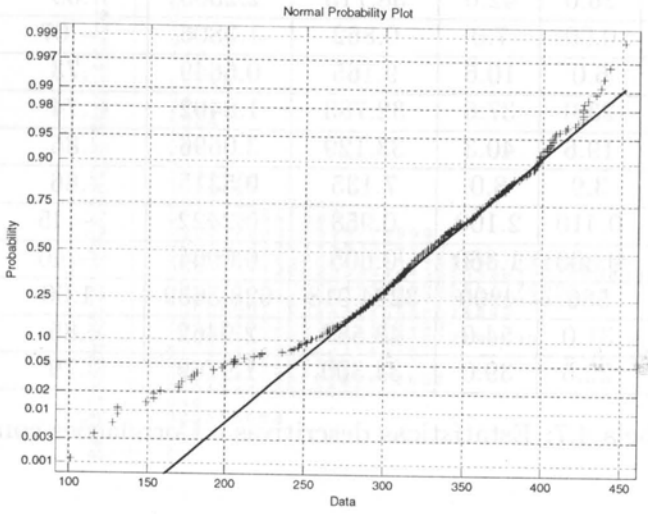


Figura 4.2: Teste de normalidade aos pesos fetais ($m = 414$)

4.2.4 Análise das componentes principais

Uma das dificuldades inerente na análise exploratória de dados com mais de três variáveis é a visualização da relação existente entre estas variáveis. Por vezes há vantagem em tentar reduzir o número de variáveis através da procura de relações (lineares) entre as variáveis, isto é, pode acontecer que várias variáveis estejam a medir a mesma grandeza biométrica. Por exemplo, as variáveis DBP e PCE medem grandezas que poderão estar correlacionadas. Admitindo ser aproximadamente elíptica a secção recta do crânio, o diâmetro menor e o perímetro da elipse estão relacionados por uma fórmula (não linear neste caso).

A análise das componentes principais é um método quantitativamente rigoroso que permite obter uma simplificação do número de variáveis. Este método gera um conjunto novo de variáveis, designadas *componentes principais*, cada uma sendo uma combinação linear das variáveis originais. Uma vez que todas as componentes principais são ortogonais entre si não existem componentes repetidas. Assim, um conjunto de componentes principais pode ser encarado como uma base ortogonal para o espaço dos dados.

Uma vez que existem infinitas maneiras de construir bases ortogonais para várias colunas de uma matriz de dados, no caso de uma base com as componentes principais tem que existir uma dada particularidade.

A primeira componente principal é um eixo único no espaço. Quando se projecta cada exemplo neste eixo gera-se uma nova variável. A variância desta nova variável tem o máximo de entre todas as escolhas possíveis para o primeiro eixo.

A segunda componente principal é um outro eixo no espaço, perpendicular ao primeiro. Projectando os exemplos neste eixo gera-se um nova variável cuja variância é também máxima de entre todas as possíveis escolhas de segundos eixos.

O conjunto completo de componentes principais é tão extenso como o conjunto original de variáveis. Mas é comum que a soma das variâncias das primeiras componentes principais exceda 80% da variância total dos dados originais. Examinado os gráficos destas novas variáveis pode obter-se uma melhor compreensão das forças motrizes que geraram os dados originais.

Para se ter uma ideia da matriz dos dados em bruto (Tabelas 4.1 e 4.2) gerou-se um gráfico em caixa (figura 4.3). Como as escalas das variáveis são muito diferentes algumas não estão representadas (as alturas AMA e os pesos PMA das mães).

Como os dados em bruto têm escalas diferentes convém normalizar os dados dividindo cada coluna pelo respectivo desvio padrão. Usou-se em seguida o comando do MATLAB *princomp(dados)*

A tabela 4.8 mostra as três primeiras componentes principais. As variáveis ecográficas com mais peso na primeira componente principal são DBP, PCE, PAB

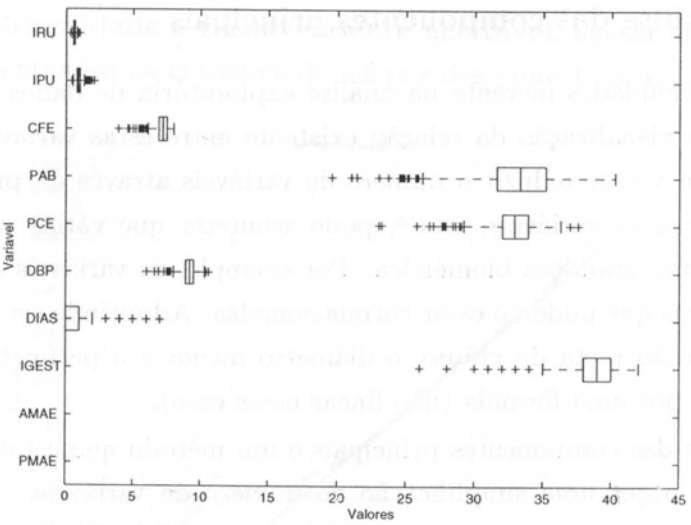


Figura 4.3: Gráficos de caixa das variáveis

	1^aC.Prin	2^aC.Prin	3^aC.Prin
PMA	0.1612	−0.1808	−0.5631
AMA	0.1006	−0.1318	−0.6432
IGE	0.3654	0.2606	0.1667
DIF	−0.0699	−0.0133	−0.4150
DBP	**0.3957**	0.2476	−0.0638
PCE	**0.4122**	0.2163	−0.0444
PAB	**0.3925**	0.1787	−0.0189
CFE	**0.4051**	0.1260	0.0297
IPU	−0.3065	0.5933	−0.1499
IRU	−0.2947	0.6115	−0.1988

Tabela 4.8: As primeiras 3 componentes principais

e CFE.

A Figura 4.4 representa os dados segundo as duas primeiras componentes principais. A percentagem da variabilidade total que é explicada por cada componente principal, indicada nas Tabelas 4.9 e 4.10, está representada na figura 4.5.

1^a C.P.	2^a C.P.	3^a C.P.	4^a C.P.	5^a C.P.
43.5886	12.0922	11.6813	9.5725	7.7839

Tabela 4.9: Variabilidade explicada (I)

6^a C.P.	7^a C.P.	8^a C.P.	9^a C.P.	10^aC.P.
4.7158	3.8300	2.7605	2.3521	1.6232

Tabela 4.10: Variabilidade explicada (II)

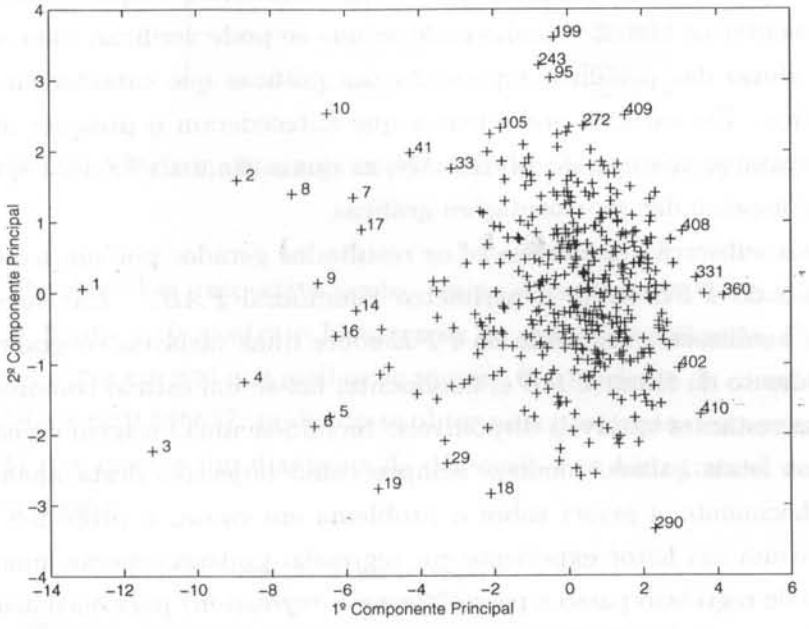


Figura 4.4: Gráfico dos dados segundos as duas primeiras novas variáveis (ou componentes principais). O números indicam os exemplos correspondentes.

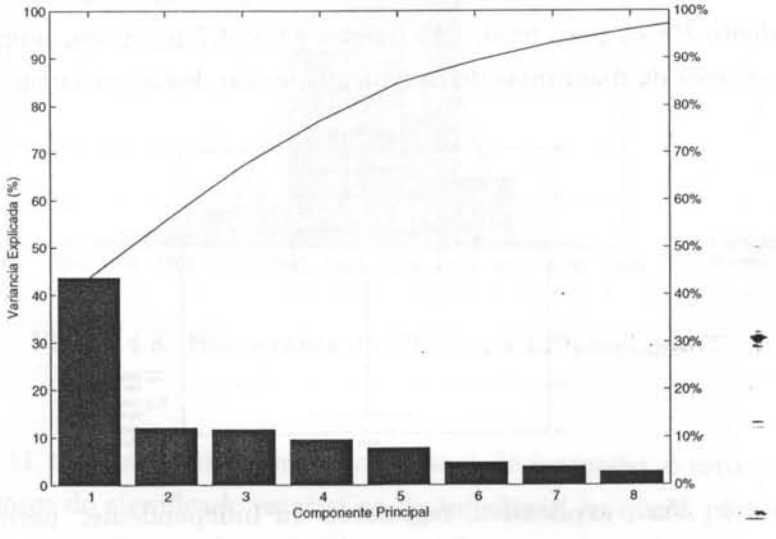


Figura 4.5: Gráfico de Pareto ("*scree*"). Percentagem de variabilidade explicada.

4.2.5 Regressão do PFE sobre as principais variáveis

Nesta subsecção mostra-se o processo que foi utilizado para obter uma boa familiarização com as especificidades do problema do peso fetal. Procedeu-se a uma análise com modelos lineares simples, recorrendo-se, na presente dissertação, às ferramentas do *STATISTICA*. Este package estatístico é aqui usado como um meio de apresentar as ideias, reconhecendo-se que se pode verificar, num ou noutro caso, algum abuso das prolíficas representações gráficas que caracterizam este package estatístico. Em estudos preliminares que antecederam o presente utilizaram-se as ferramentas estatísticas do *MATLAB*, as quais são mais flexíveis quer no número, quer no aspecto das representações gráficas.

Nesta subsecção descrevem-se os resultados gerados por um modelo simples de regressão do PFE sobre o perímetro abdominal PAB. Nas subsecções que se seguem analisa-se a regressão do PFE sobre duas variáveis ecográficas o PAB e o comprimento do fémur CFE e, finalmente, faz-se um estudo compreendendo quase todas as restantes variáveis disponíveis. Incluiu-se uma focagem específica nos casos de pesos fetais baixos, tendo-se sempre como objectivo desta análise a aquisição de conhecimento *a priori* sobre o problema em causa, o problema do peso fetal, e não, como um leitor experiente em regressão poderia esperar, uma utilização do método de regressão passo a passo (*stepwise regression*) para encontrar uma solução do problema. Nesta dissertação, como se sabe, a solução é investigada em modelos de rede neuronal e máquina de vectores de suporte.

Como já foi referido a amostra de trabalho é constituída por uma matriz com 414 casos e 16 variáveis. As tabelas 4.1 e 4.2 permitem ter-se uma percepção de alguns valores da variável independente PAB, perímetro abdominal, e da variável dependente PFE, peso fetal. As figuras 4.6 e 4.7 mostram, num abrir e fechar de olhos, através de diagramas de caixa a dispersão destas variáveis.

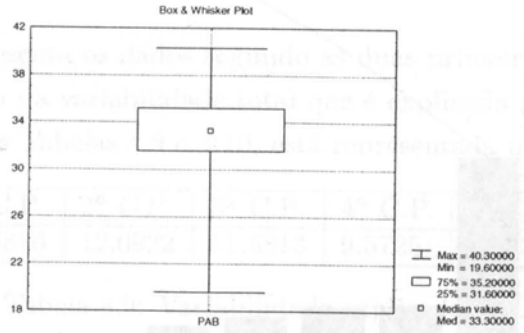


Figura 4.6: Var. explicativa, regressora ou independente: perimetro abdominal, PAB.

As figuras 4.8 e 4.9 mostram os histogramas dos valores observados de PFE,

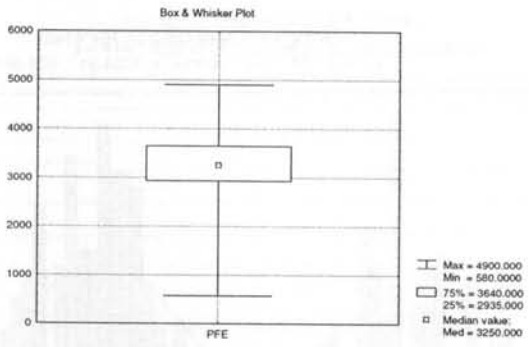


Figura 4.7: Var. resposta, regressanda ou dependente: peso fetal, PFE.

com 10 e 70 intervalos, respectivamente, assim como a curva normal que melhor se ajusta. A figura 4.10 mostra o histograma da variável explicativa PAB com 70 intervalos e a curva normal que melhor se ajusta. O coeficiente de correlação entre o PFE e o PAB é de 0.846657, podendo-se obter uma percepção desta estatística com a figura 4.11 que mostra um diagrama de dispersão e os histogramas das variáveis, em cima e à direita.

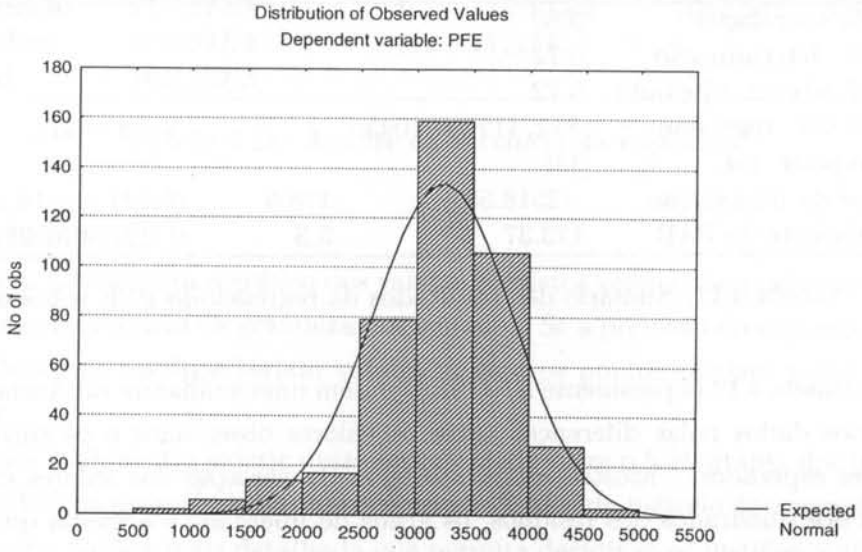


Figura 4.8: Histograma do PFE com 10 intervalos

A tabela 4.11 mostra os coeficientes (ou pesos) de regressão, o erro padrão e os respectivos valores do significado estatístico da inferência, os quais, para uma dado intervalo de confiança ($1 - alfa = 0.95$), se traduzem pelos valores$-p$ obtidos na tabela $t-Student$ correspondente, com 23 graus de liberdade e uma aba direita cuja área é igual a $alfa/2$.

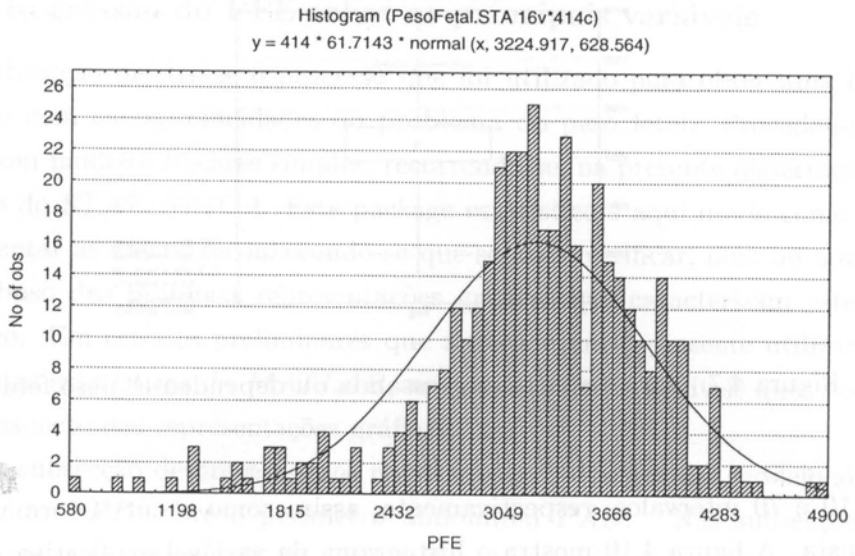


Figura 4.9: Histograma do peso fetal com 70 intervalos e a curva Gaussiana ajustada $\mathcal{N}(3225, 628)$

Regressão Linear		*Err Pad*	
Coef. correlação	0.85		
Coef. determinação	0.72		
Coef.determ. ajustado	0.72		
Anál var. regressão	$F(1, 412) = 1043$		$Num = 414, \quad p < 0.000000$
Erro padr. est.	335		
Valor da intersecção	-2518.50	178.6	$t(412) = -14.10\ p < .000000$
Coeficiente do PAB	173.37	5.3	$t(412) = 32.29\ p < .000000$

Tabela 4.11: Sumário dos resultados da regressãodo PFE sobre o PAB

A tabela 4.12 (tipicamente ANOVA) contém uma análise da variância dos resíduos (desvios dados pelas diferenças entre os valores observados e os correspondentes valores esperados). Mostra ainda a origem da variação dos valores explicados, a soma dos quadrados dos resíduos, os graus de liberdade e a média quadrática dos resíduos. Adicionalmente, apresenta-se, nesta tabela, os resultados de um teste de hipóteses, em que a hipótese nula, H_o : $coeficiente = 0$, versus a hipótese alternativa, $H_{alt.}$: $coeficiente \neq 0$. Uma vez que se obteve um valor para $F = 1042,,$ muito maior que $F(0.95; 1, 414) = 3.84$, logo, verifica-se a $H_{alt.}$, como é evidente. Esta análise da variância tem em geral mais importância na análise de regressão múltipla, do que na regressão simples.

Considera-se, agora, a análise dos resíduos com o objectivo de: (i) detecção de *outliers*, (ii) verificação de linearidade, (iii) testar a normalidade. Por vezes, há conveniência em interpretar este conjunto de valores em termos de uma variável

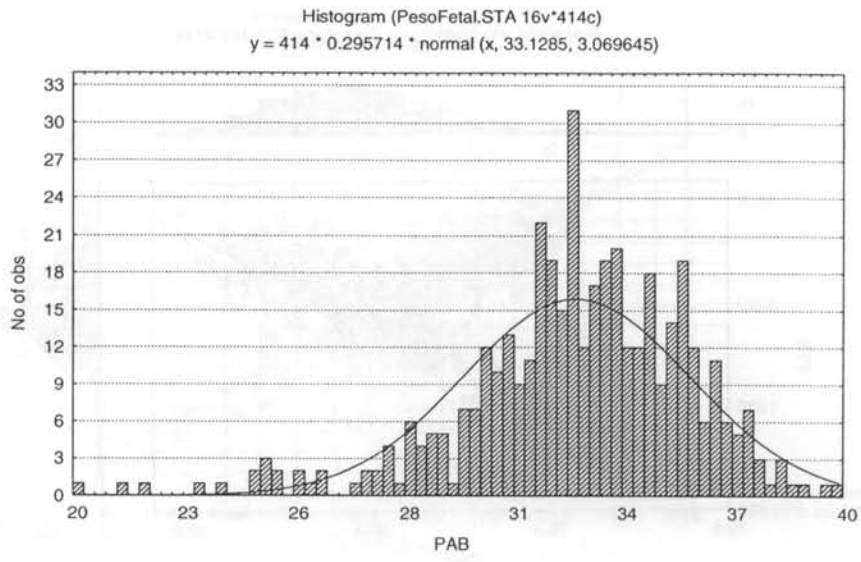


Figura 4.10: Histograma do perimetro abdominal com 70 intervalos e a curva Gaussiana ajustada $\mathcal{N}(33.1; 3.1)$

Efeito	Soma Quad. Res.	df	Média Quad.	F	valor$-p$
Regressão	$116967E3$	1	$116967E3$	1042.943	0.00
Residual	$462062E2$	412	112151.		
Total	$163173E3$				

Tabela 4.12: Análise da variância da regressão

aleatória residual.

A figura 4.12 mostra o gráfico dos valores previstos versus observados permitindo ver claramente a ordem de grandeza dos resíduos. Se a previsão do regressando fosse perfeita todos os resíduos teriam valores nulos e os pontos cairiam todos na recta bissectriz.

A figura 4.13 mostra existir uma conformidade entre o histograma dos resíduos e uma distribuição normal ajustada, pressuposto básico do método de regressão linear.

Para uma análise mais detalhada que permita decidir se se justifica a eliminação ou conservação de um caso usa-se uma tabela do dos casos cujos resíduos sejam superiores a $+2$ ou -2 vezes o desvio padrão dos resíduos, os quais se podem considerar *outliers*. A tabela 4.13 mostra os seis primeiros casos de um total de 24 casos de resíduos elevados. Para cada caso, tabelam-se, os valores observados e esperados, os resíduos, os valores esperados do PFE normalizados, os valores normalizados dos resíduos, os erros padrão do valor esperado do PFE não normalizado, a distância de Mahalanobis, o valor residual quando o caso é eliminado.

O valor normalizado do resíduo é dado pelo quociente entre o resíduo e a raiz

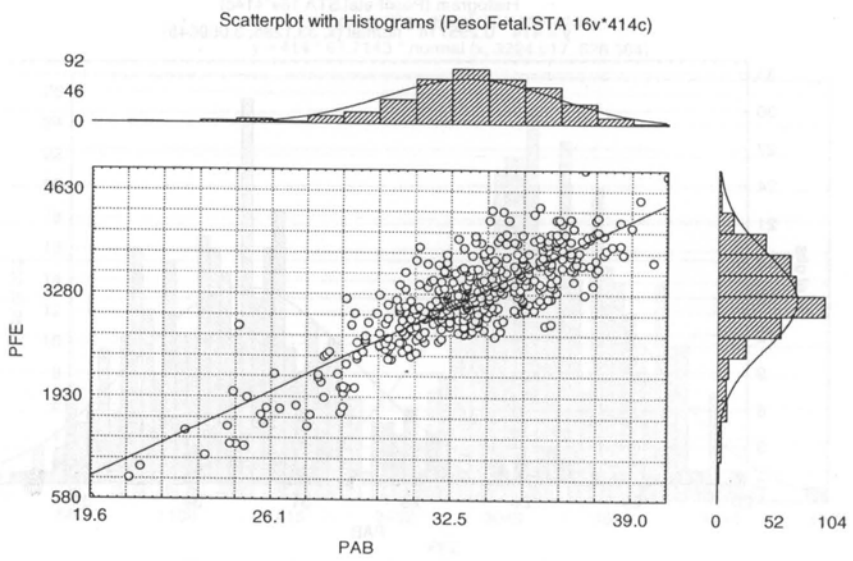


Figura 4.11: Diagrama de dispersão do PFE versus PAB, histogramas e recta de regressão, permite visualizar a correlação entre as variáveis.

quadrada da média dos quadrados dos resíduos. A distância de *Mahalanobis* é a distância do caso ao centróide do espaço multidimensional dado pelas variáveis independentes correlacionadas. Num estudo simples univariado não faz muito sentido falar-se correlação entre variáveis independentes, esta distância equivale neste caso à distância euclidiana. Esta medida dá uma indicação sobre se uma observação é, ou não é, um *outlier* com respeito aos valores da variável independente. Se o valor residual com o caso eliminado for inferior ao valor residual com o caso estava incluído, então, esse caso deve ser eliminado do conjunto de dados.

CAS	Obs	Prev	Res	PrevN	ResN	PrevErrP	MahaD	ResElim
10	1540	2231	−691	−1.8	−2.0	34.8	3.4	−699
14	1720	2439	−719	−1.4	−2.1	29.3	2.1	−725
86	2860	1798	1061	−2.6	3.1	47.1	7.1	1083
87	2870	3705	−835	.90	−2.4	22.1	.81	−839
91	2890	3740	−850	.96	−2.5	22.9	.93	−854
195	3210	2474	735	−1.4	2.1	28.4	1.9	740
Min	1540	1798	−850	−2.6	−2.5	16.4	.001	−854
Max	4900	4381	1061	2.1	3.1	47.1	7.1	1083
Med	3643	3220	423	−.007	1.2	22.9	1.1	424
Mdn	3835	3271	728	.088	2.1	19.7	.44	732

Tabela 4.13: Resíduos superiores a +sigma e inferiores a -sigma (n=414)

O interesse das estatísticas sobre os resíduos, que se mostram na tabela 4.13, reside na sua importância para a identificação dos possíveis *outliers*. Notar que

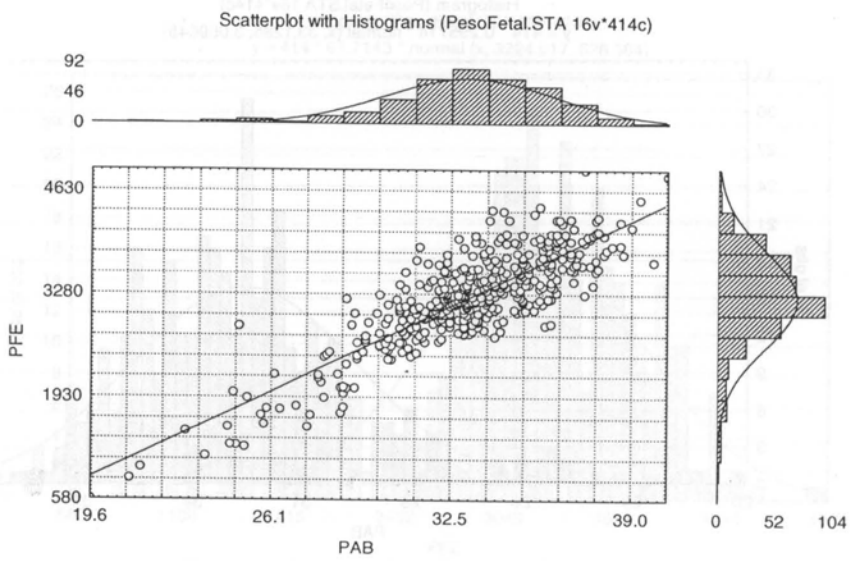


Figura 4.11: Diagrama de dispersão do PFE versus PAB, histogramas e recta de regressão, permite visualizar a correlação entre as variáveis.

quadrada da média dos quadrados dos resíduos. A distância de *Mahalanobis* é a distância do caso ao centróide do espaço multidimensional dado pelas variáveis independentes correlacionadas. Num estudo simples univariado não faz muito sentido falar-se correlação entre variáveis independentes, esta distância equivale neste caso à distância euclidiana. Esta medida dá uma indicação sobre se uma observação é, ou não é, um *outlier* com respeito aos valores da variável independente. Se o valor residual com o caso eliminado for inferior ao valor residual com o caso estava incluído, então, esse caso deve ser eliminado do conjunto de dados.

CAS	Obs	Prev	Res	PrevN	ResN	PrevErrP	MahaD	ResElim
10	1540	2231	−691	−1.8	−2.0	34.8	3.4	−699
14	1720	2439	−719	−1.4	−2.1	29.3	2.1	−725
86	2860	1798	1061	−2.6	3.1	47.1	7.1	1083
87	2870	3705	−835	.90	−2.4	22.1	.81	−839
91	2890	3740	−850	.96	−2.5	22.9	.93	−854
195	3210	2474	735	−1.4	2.1	28.4	1.9	740
Min	1540	1798	−850	−2.6	−2.5	16.4	.001	−854
Max	4900	4381	1061	2.1	3.1	47.1	7.1	1083
Med	3643	3220	423	−.007	1.2	22.9	1.1	424
Mdn	3835	3271	728	.088	2.1	19.7	.44	732

Tabela 4.13: Resíduos superiores a +sigma e inferiores a -sigma (n=414)

O interesse das estatísticas sobre os resíduos, que se mostram na tabela 4.13, reside na sua importância para a identificação dos possíveis *outliers*. Notar que

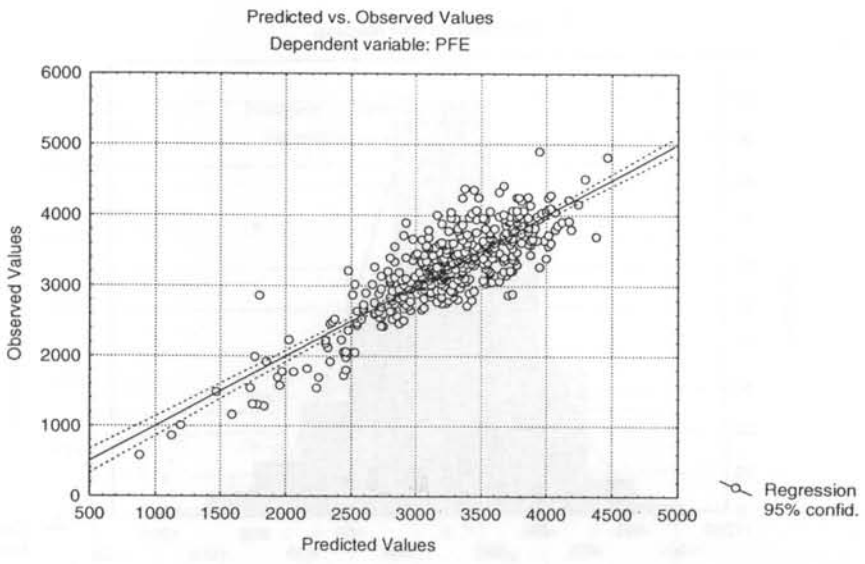


Figura 4.12: Valores previstos versus valores residuais para a variável peso fetal (PFE)

para amostras pequenas uma estimativa dos coeficientes da regressão (simples ou múltipla) pode ser muito instável. Por outras palavras, um caso com um valor extremo pode ter uma influência muito significativa nas estimativas finais.

Vai-se proceder, agora, a uma análise dos resíduos com vista ao estudo da *linearidade*. A figura 4.14 mostra o gráfico dos valores esperados *versus* resíduos, o qual permite testar a suposição de linearidade da relação entre as variáveis. Se esta suposição se verificar, então, os valores esperados tendem a formar uma "*núvem*" homogénea à volta da linha recta paralela ao eixo horizontal que passa no ponto residual nulo. Se aparecer um conjunto de pontos que esboçam um "U" invertido à volta desta recta isso pode significar que a relação entre estas variáveis poderá ter uma melhor representação através de uma fórmula polinomial.

Finalmente, vai-se proceder a uma análise dos resíduos com vista ao estudo da *normalidade*. O método de regressão linear pressupõe que a distribuição dos resíduos seja normal. A figura 4.15 mostra um gráfico que permite observar o grau desta normalidade. Depois de se ordenarem os resíduos determinam-se os valores de z, isto é, os valores normalizados da distribuição normal baseados no pressuposto de que eles são provenientes de uma distribuição normal. Estes valores são representados no eixo vertical. Se os valores dos resíduos observados forem distribuídos normalmente então os pontos devem pertencer a uma linha recta. Os *outliers* podem também aparecer evidenciados neste gráfico.

Parece que se verificam falhas de ajustamento à curva normal apenas nos valores elevados dos resíduos positivos e negativos. Os pontos parecem formar um padrão

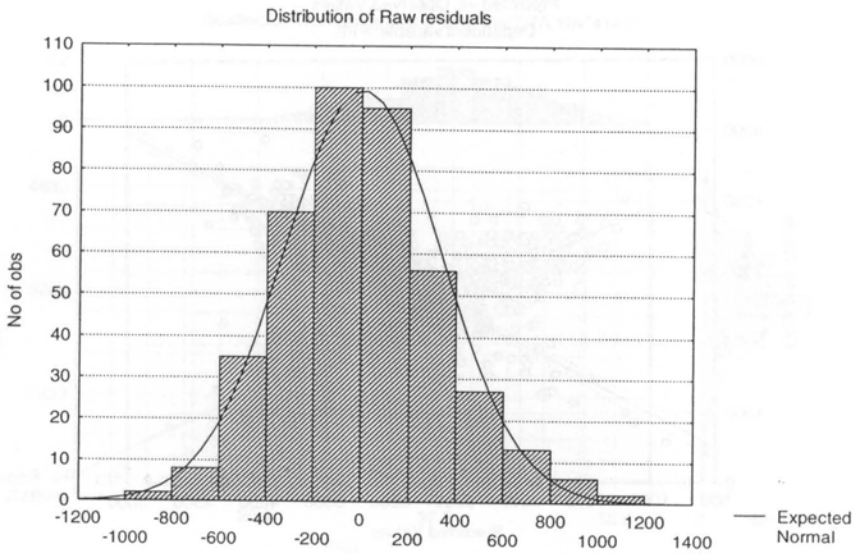


Figura 4.13: Histograma dos resíduos

evidente sobre a linha recta, que não parece ter a forma de um "S", mas antes em forma de um "$\cap$". Então, poder-se-ia verificar que se melhoraria esta situação se a variável dependente pudesse ser sujeitas a uma transformação que permitisse evitar esta situação.

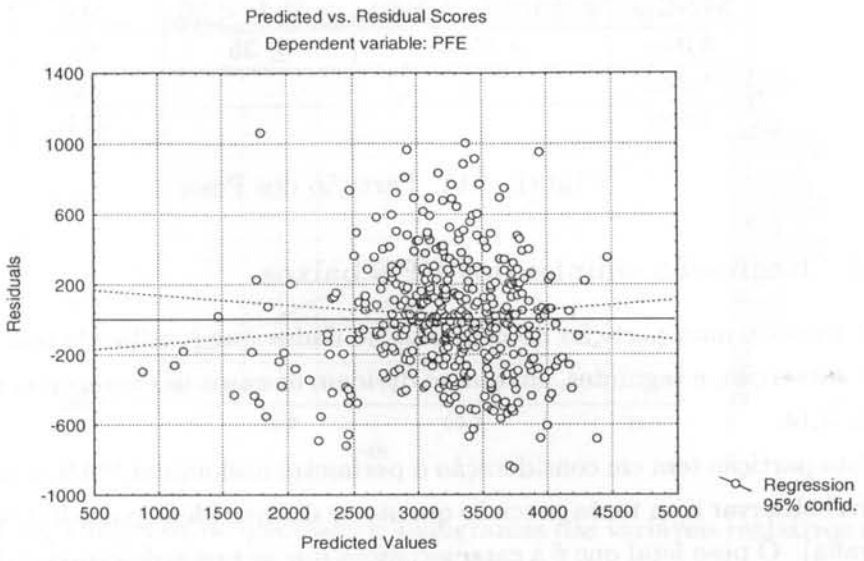


Figura 4.14: Valores previstos versus valores dos resíduos da variável PFE

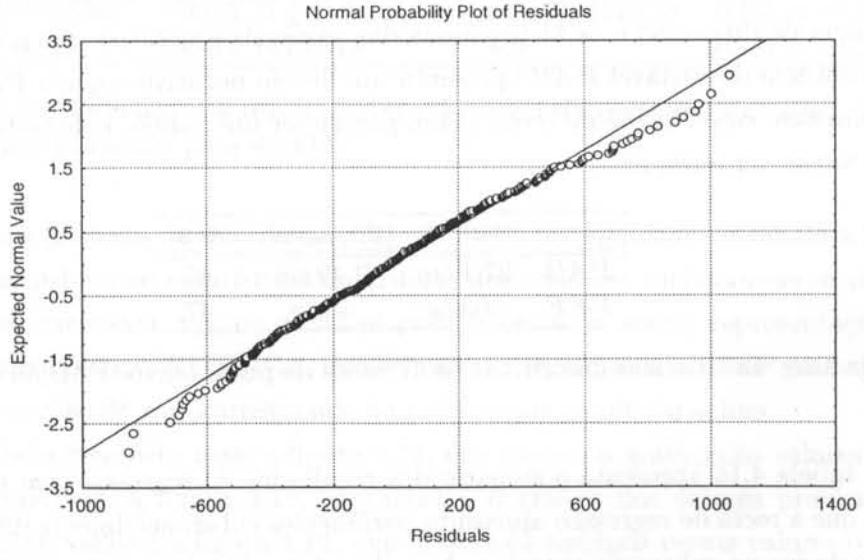


Figura 4.15: Diagrama de normalidade dos resíduos

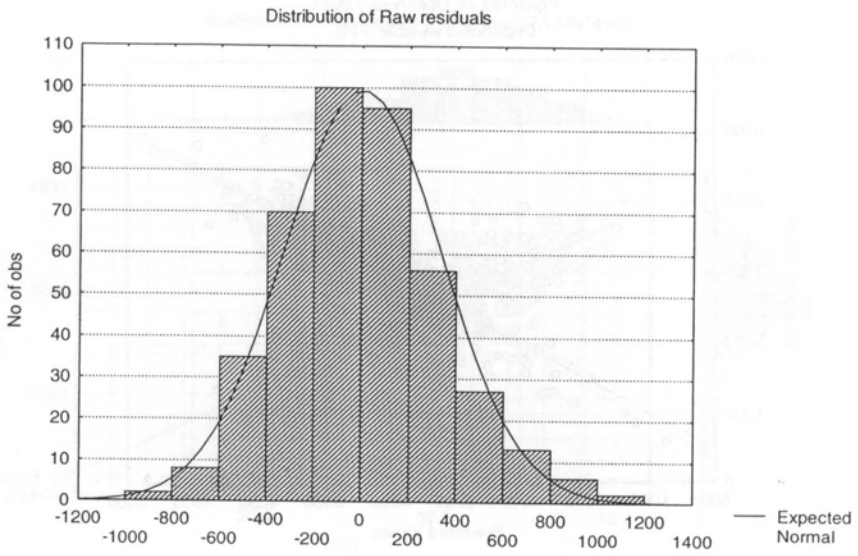


Figura 4.13: Histograma dos resíduos

evidente sobre a linha recta, que não parece ter a forma de um "S", mas antes em forma de um "$\cap$". Então, poder-se-ia verificar que se melhoraria esta situação se a variável dependente pudesse ser sujeitas a uma transformação que permitisse evitar esta situação.

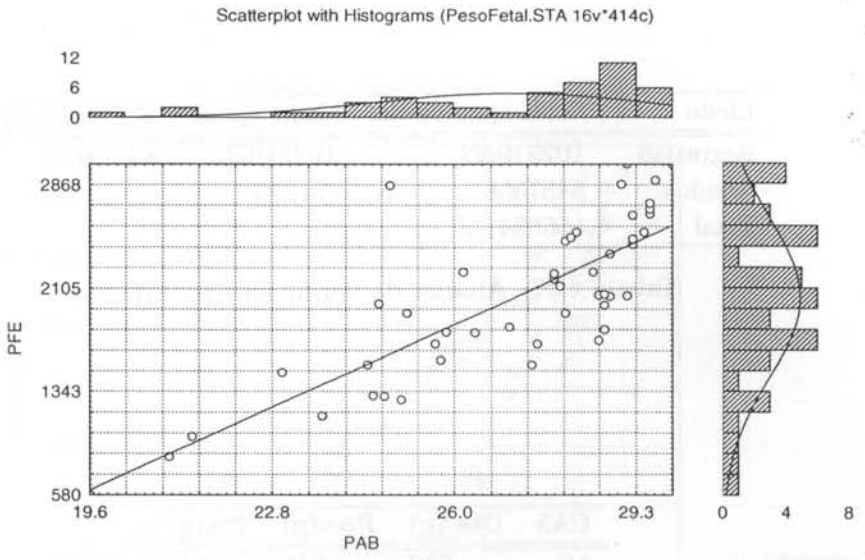


Figura 4.16: Diagrama de dispersão e histogramas das variáveis realativos a 47 casos de pesos baixos, $PAB < 30$cm e $PFE < 3210$ gr.

Regressão Linear		*Erro Pad.*	
Coef. correlação	0.81 (0.85)		
Coef. determin.	0.65 (0.72)		
Coef. det. ajust.	0.64 (0.72)		
Anál. variância	$F(1,45) = 84$		$Num = 47,\ p < 0.000000$
Erro padr. estim.	348 (335)		
Valor intersecção	-3101.21 g (-2518.50)	561.2 g (178.6)	$t(45) = -5.526p < .000002$
Coef. PAB	189.71 (173.37)	5.3	$t(45) = 32.29\ p < .000000$

Tabela 4.16: Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414)

de *outliers* (2) teste da *linearidade*, (3) e teste da *normalidade*. A tabela 4.18 mostra apenas as principais estatísticas descritivas dos resíduos, omitiram-se os valores dos casos que excediam $\pm\sigma$, os quais se podem observar numa representação gráfica icónica dada pelo STATISTICA, que se omitiu para simplificar. Este gráfico permitiu detectar o caso 86 que corresponde ao *outlier* que se referiu acima.

A tabela completa mais a figura 4.17, que mostra o gráfico dos valores previstos versus residuais, a figura 4.18, que mostra o gráfico dos valores previstos versus valores observados e a figura 4.19, que mostra os resíduos versus valores observados permitem concluir que existe um caso que constitui um *outlier* e que pode ter uma influência indesejada os resultados.

As figuras do teste de normalidade e histograma dos resíduos 4.20 e 4.21 permitem pôr em causa que a suposição de normalidade dos resíduos.

Efeito	Soma Quad.	df	Média Quad.	F	valor$-p$
Regressão	$102310E2$	1	$102310E2$	84	0.00
Residual	5457356	45	121275		
Total	$156884E2$				

Tabela 4.17: Análise da variância da regressão

CAS	Obs (g)	Prev(g)	Res(g)
Min	580	617	−604
Max	3020	2571	1237
Med	2032	2032	0
Mdn	2050	2210	−8

Tabela 4.18: Tabela (n=47) dos valores máximos e mínimos dos resíduos

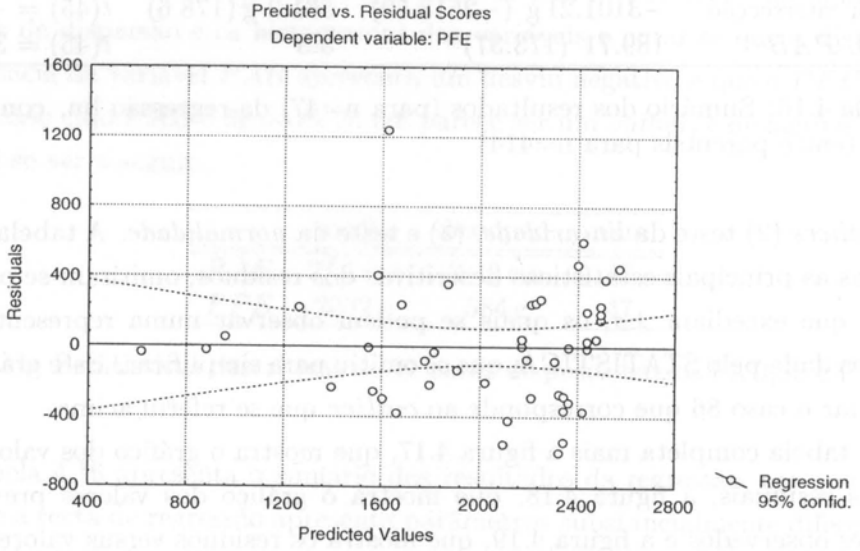


Figura 4.17: Valores previstos vs resíduos para a var PFE (pesos baixos, 47 casos)

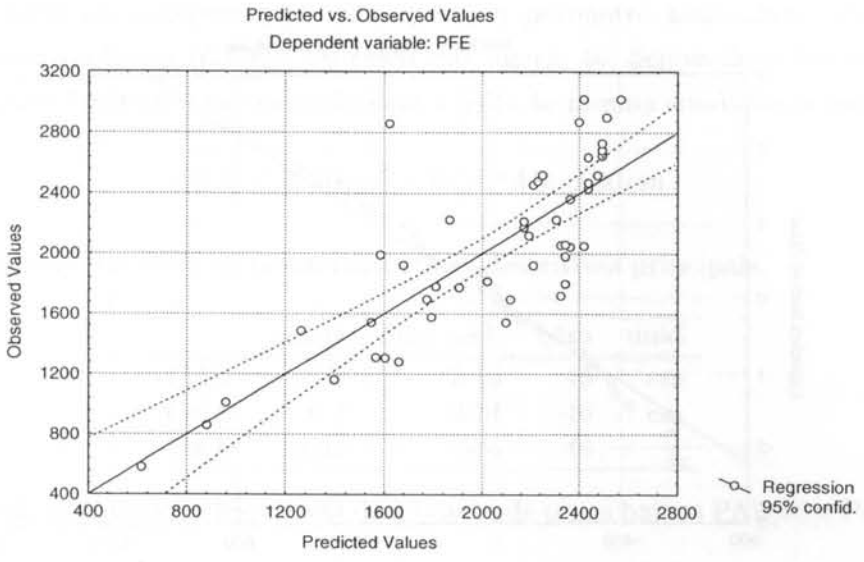


Figura 4.18: Valores previstos vs. observados para a var. PFE ($n = 47$, pesos fetais baixos)

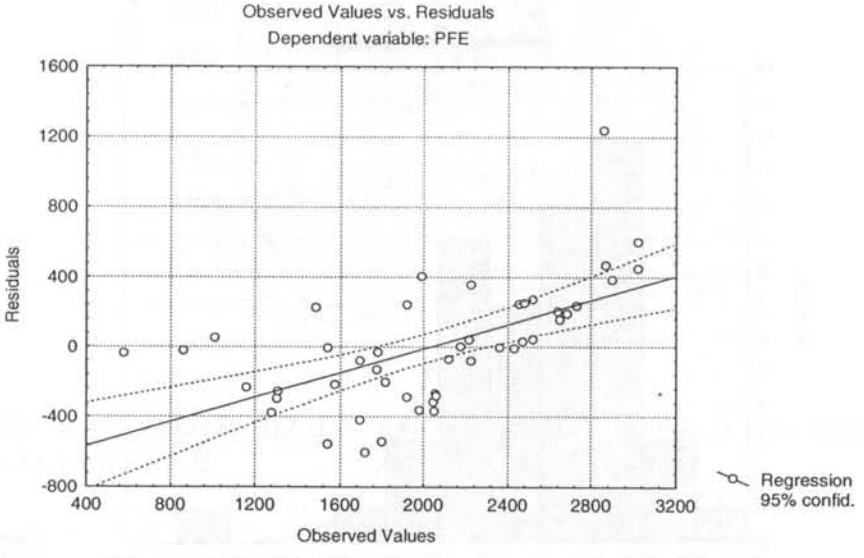


Figura 4.19: Valores observados vs. residuais para a var. PFE ($n = 47$, pesos baixos)

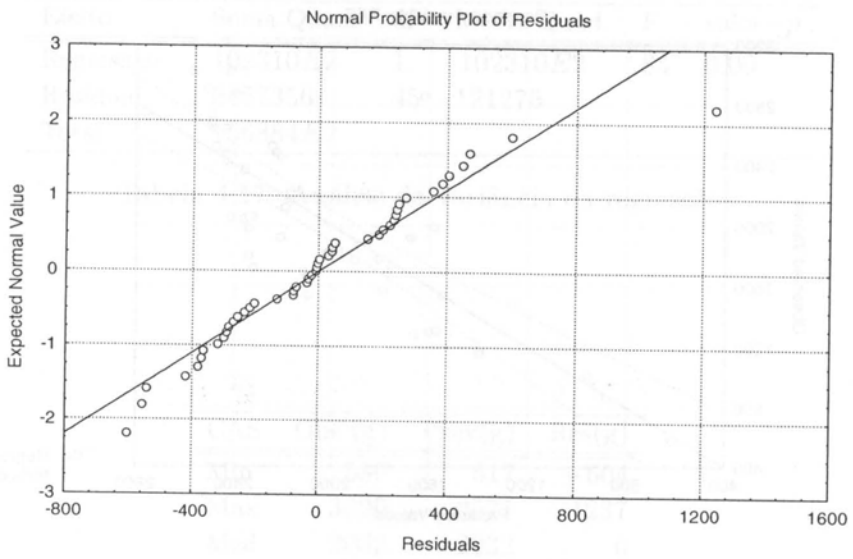


Figura 4.20: Gráfico de normalidad dos resíduos ($n = 47$)

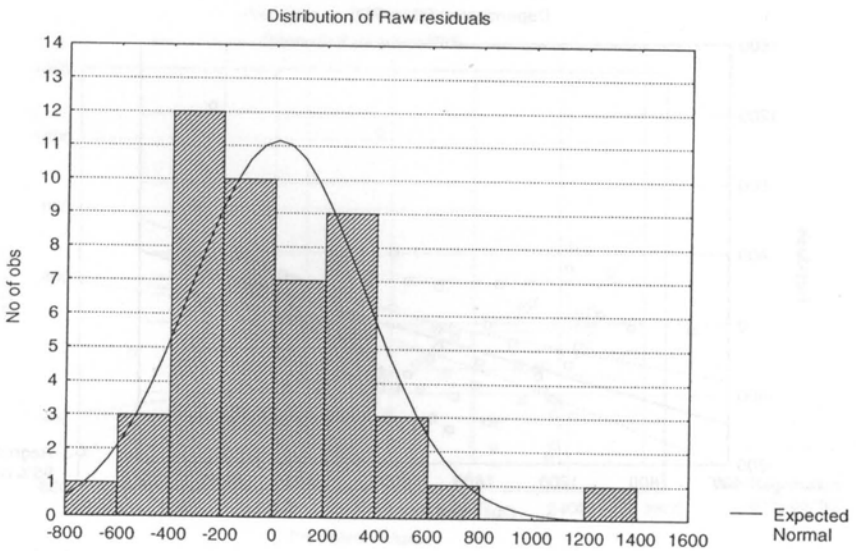


Figura 4.21: Histograma da distribuição dos resíduos

4.2.7 Regressão de PFE baixos sobre duas variáveis ecográficas

As duas variáveis independentes são, agora, o perímetro abdominal (PAB) e o comprimento do fémur (CFE). Os casos são, agora, 46, depois de se ter eliminado o caso *outlier* (#86) e a sua selecção está sujeita às mesma condições anteriores:

$$PFE < 3210\,\text{cm} \;\; AND \;\; PAB < 30\,\text{cm}\,.$$

A tabela 4.19 mostra as novas estatísticas descritivas principais.

	média	desv.pad.	num	unid
PAB	27.1	2.49	46	cm
CFE	6.2	0.74	46	cm
PFE	2032	584	46	g

Tabela 4.19: Estatisticas descritivas de 46 casos de pesos baixos PAB¡30 e PFE¡3210

A figura 4.22 mostra um diagrama de dispersão tridimensional.

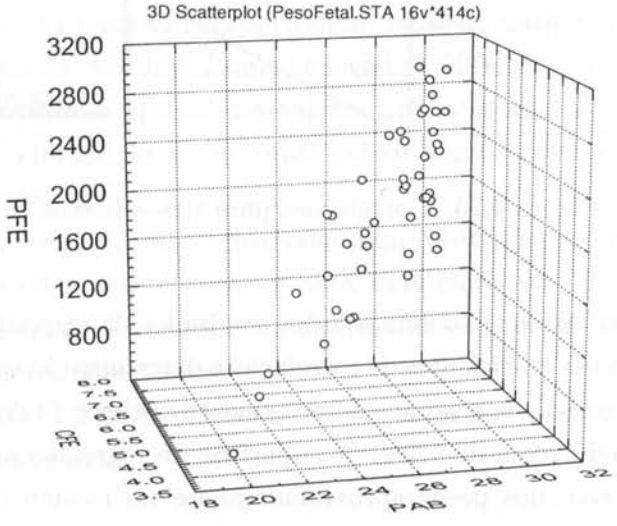


Figura 4.22: Representação $3D$ dos casos ($n = 46$). Segundo o eixo vertical está representada a var. PFE

A tabela 4.20 mostra os coeficientes de correlação podendo observar-se que os valores da correlação entre o PFE e as variáveis independentes PAB e CFE são de 0.86 e 0.90, respectivamente, ambos sendo valores elevados tal como se esperava. Notar que em consequência da eliminação do *outlier* o coeficiente de correlação do PAB sofreu um aumento em relação ao valor que se obteve anteriormente (0.81)

	PAB	*CFE-*	*PFE*
CFE	.83		
PFE	.86	.90	

Tabela 4.20: Matriz das Correlações de pesos baixos PAB¡30 e PFE¡3210

Os histogramas das variáveis podem agora observar-se no gráfico da figura 4.23 que mostra visualmente a correlação entre as variáveis.

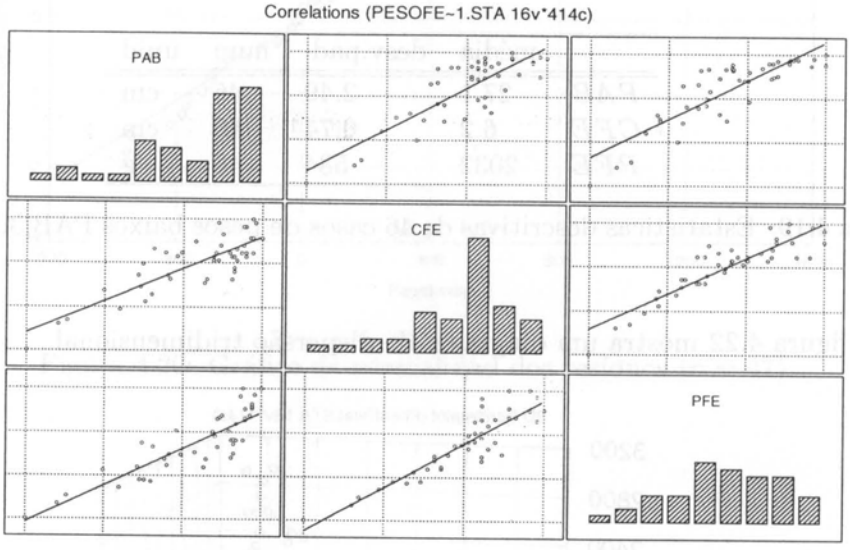


Figura 4.23: Correlações para pesos baixos (n=46)

A tabela 4.21 apresenta o sumário dos resultados da regressão podendo salientar-se que os valores dos coeficientes de correlação e determinação sofreram um acréscimo substancial em relação aos valores anteriormente obtidos. O erro padrão dos valores previstos diminuiu cerca de 120 g. A superfície de regressão apresenta, agora, uma intersecção do eixo dos pesos aproximadamente no mesmo ponto deste eixo que apresentava no caso anterior. Estes valores da intersecção e dos coeficientes têm significado estatístico, uma vez que os valores da estimativa de $t(43)$ são todos inferiores aos valores dados pela tabela para $t(0.975, 43)$ e os níveis de significância observados, isto é os $valores - p$ são apropriados.

A análise da variância que se mostra na tabela 4.22 apresenta valores próximos dos obtidos no caso anterior com uma única variável explicativa.

Tem-se agora uma nova tabela 4.23 das Correlações parciais.

As principais estatísticas descritivas dos valores dos resíduos que excediam $\pm\sigma$ mostram-se na tabela 4.24. Incluíram-se apenas os valores de dois casos porque eles apareceram também anteriormente na tabela 4.13 estando os seus valores indicados

Regressão Linear		*Erro Pad.*	
Coef. correlação	0.92 (0.81)		
Coef.determinação	0.85 (0.65)		
Coef. determ. ajust.	0.84 (0.64)		
Anál. var.	$F(2,43) = 123$		$Num.c. = 46,\ p < 0.000000$
Erro padr. estim.	228 (348)		
Valor intersecção	-3153.65 g (-3101)	373 g	$t(43) = -8.4\ p < .000002$
Coef. do PAB	83.79 (189.71)	24.7	$t(43) = 3.3\ p < .0015$
Coef. do CFE	464.27	82.8	$t(43) = 5.6\ p < .000001$

Tabela 4.21: Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414)

Efeito	Soma Quad.	df	Média Quad.	F	valor$-p$
Regressão	$127548E2$	2	6377411	122	0.00
Residual	2233532	43	51943		
Total	$149884E2$				

Tabela 4.22: Análise da variância da regressão pesos baixos (n=46)

entre-parênteses. Pode, assim, comparar-se o efeito nos resíduos da contracção do número de casos de 414 para 46 (já com a eliminação do outlier #86). Por exemplo, obteve-se no resíduo do caso n^o 14 uma redução de 70%. .

A figura 4.24 mostra o gráfico dos valores previstos *versus* valores observados, notando-se a possibilidade de poder ainda existir alguns casos *outliers* cujos PFE sejam superiores a 1800 g.

A figura 4.25 mostra o gráfico dos valores previstos *versus* residuais, notando-se uma possível tendência dos resíduos para uma curva em "U". Uma tendência análoga pode também ser observada na figura 4.26 em que se mostra o gráfico dos resíduos *versus* valores observados.

Os gráficos 4.27 e 4.28 evidenciam a tendência da distribuição dos resíduos para se afastar da normalidade.

Var	Co. Norm	Corr.Par	Cor.S.Par.	Toleranc	R^2	$t(43)$	valor$-p$
PAB	.361843	.459028	.199452	.303836	.696164	3.388085	.001516
CFE	.598758	.649840	.330044	.303836	.696164	5.606429	.000001

Tabela 4.23: Correlações Parciais pesos baixos (n=46)

CAS	Obs	Prev	Res	PrevN
10	1540	1788 (2231)	−248 (−691)	−0.423 (−1.8)
14	1720	1935 (2439)	−215 (−719)	−0.147 (−1.4)
Min	580	299	−448	−3.2
Max	3020	2753	575	1.3
Med	2014	2014	0.000	0.000
Mdn	2047	2132	−11.4	0.22

Tabela 4.24: Resíduos superiores a +sigma e inferiores a -sigma (I)

CAS	Obs	ResN	PrevErrP	MahaD	ResElim
10	1540	−1.0 (−2.0)	60.9 (34.8)	2.2 (3.4)	−267 (−699)
14	1720	−.94 (−2.1)	77.8 (29.3)	4.2 (2.1)	−244 (−725)
Min	580	−1.9	34.9	0.07	−481
Max	3020	2.5	114	10.3	600
Med	2014	0.000	55.9	1.9	2
Mdn	2047	−0.05	53.8	1.5	−11

Tabela 4.25: Resíduos superiores a +sigma e inferiores a -sigma (II)

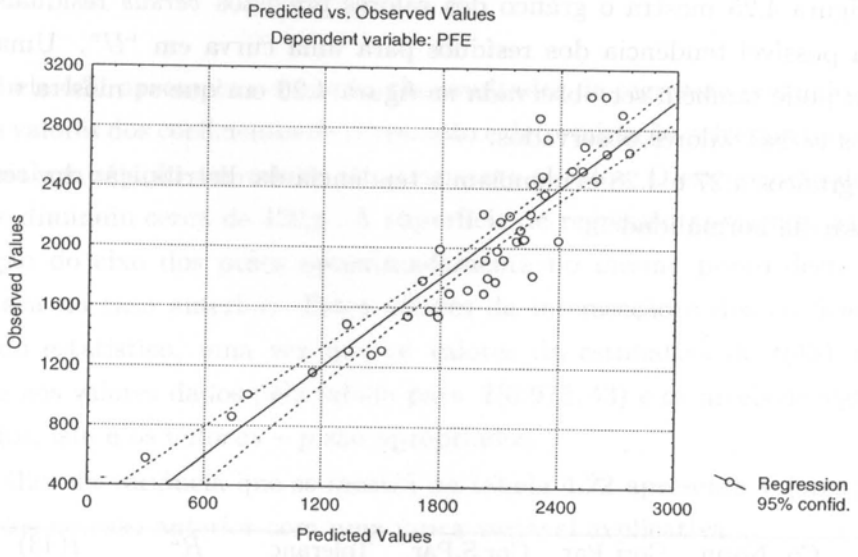


Figura 4.24: Valores previstos vs. observados, duas var. PAB e CFE ($m = 46$)

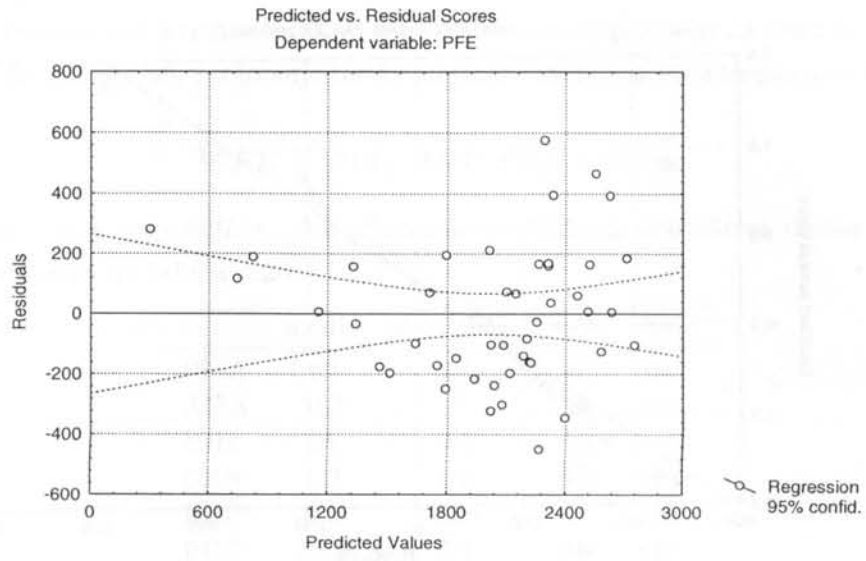


Figura 4.25: Valores dos previstos vs. resíduos, duas var. PAB e CFE, $(m = 46)$

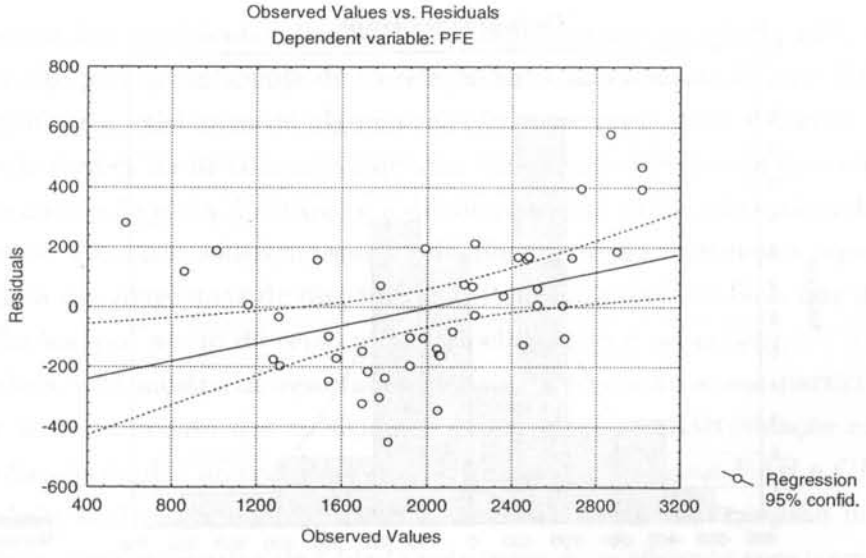


Figura 4.26: Valores observados vs. resíduos, duas var. PAB e CFE, $(m = 46)$

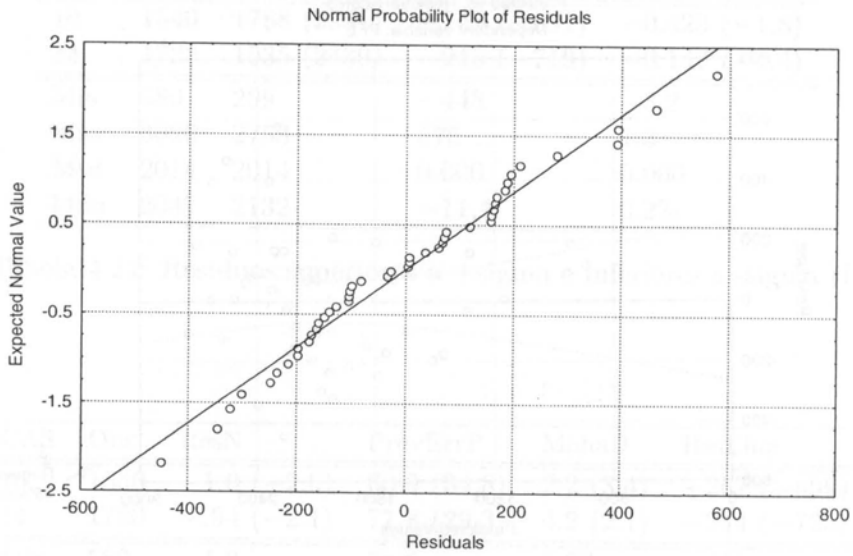


Figura 4.27: Gráfico de normalidade, duas var. PAB e CFE ($m = 46$)

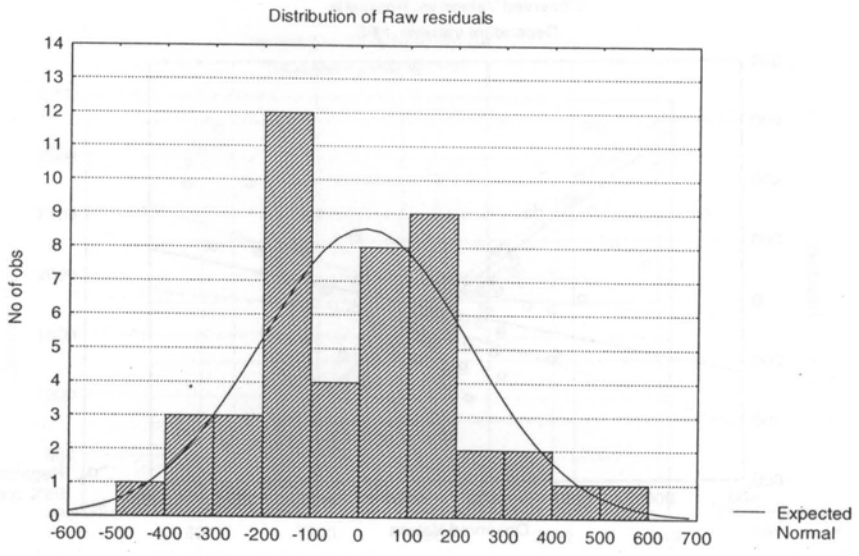


Figura 4.28: Histograma dos resíduos, duas var. PAB e CFE ($m = 46$)

4.2.8 Regressão dos PFE baixos sobre múltiplas variáveis

As variáveis independentes utilizadas na análise de regressão do PFE que se descreve nesta subsecção são a quase totalidade das variáveis disponíveis na base de dados. O número de casos é 46, estão sujeitos às mesmas condições de selecção anteriormente referidas

$$PFE < 3210\,\mathrm{g}\ \ AND\ PAB < 30\,\mathrm{cm}$$

e excluiu-se um caso *outlier*. As principais estatísticas descritivas destas variáveis estão indicadas na tabela 4.26.

	média	desv.pad.	num	unid
PMA	69	12	46	kg
AMA	157	7	46	cm
IGE	35	3.5	46	*sem.*
DEN	1.3	1.9	46	*dias*
DBP	8.2	0.66	46	cm
PCE	29.5	2.1	46	cm
PAB	27.1	2.49	46	cm
CFE	6.2	0.74	46	cm
IPU	1.172	0.368	46	–
IRU	0.685	0.123	46	–
PFE	2032	584	46	g

Tabela 4.26: Estatisticas descritivas de 11 var. e 46 casos de pesos baixos PAB¡30 e PFE¡3210

A matriz dos coeficientes de correlação, representada na tabela 4.27, apresenta um valor alto para o coeficiente de correlação entre as variáveis PCE e DBP o qual pode significar a existência de alguma relação entre estas duas variáveis. Uma tal relação não deverá ser de colinearidade uma vez que o perímetro de uma elipse interpoladora da secção recta do crânio e o diâmetro menor deste não estão relacionados linearmente. Para não sobrecarregar a presente exposição omitiu-se a representação dos gráficos dos diagramas de dispersão e histogramas das variáveis que ilustravam as correlações, por serem de certo modo semelhantes aos anteriores.

A tabela 4.29 mostra os resultados globais da regressão e, em particular, pode verificar-se um aumento dos coeficientes de correlação e determinação em relação aos que foram obtidos na regressão do PFE sobre as variáveis PAB e CFE.

A tabela 4.30 mostra as estimativas dos coeficientes da regressão múltipla, o erro padrão, uma estimativa de t (dada pela razão do coeficiente com o erro padrão) e os $valores - p$. Pode observar-se que são estatisticamente significativas apenas a intersecção e as variáveis PAB e CFE, uma vez que apresentam os valores de $t(35) > 2$ e os $valores - p$ compatíveis com um nível de significado pelo menos de 90%.

	PMA	AMA	IGE	DEN
AMA	.254			
IGE	.104	−.207		
DEN	.031	.121	−.326	
DBP	.031	.298	.598	−.158
PCE	.075	.263	.655	−.199
PAB	−.007	−.007	.639	−.154
CFE	.173	.104	.733	−.243
IPU	−.120	.033	−.436	−.089
IRU	−.054	−.104	−.412	−.048
PFE	.124	.098	.722	−.196

Tabela 4.27: Matriz (I) das Correlações de pesos baixos PAB¡30 e PFE¡3210

	DBP	PCE	PAB	CFE	IPU	IRU
AMA						
IGE						
DEN						
DBP						
PCE	.903					
PAB	.756	.752				
CFE	.750	.801	.834			
IPU	−.480	−.484	−.624	−.589		
IRU	−.485	−.485	−.571	−.646	.684	
PFE	.772	.768	.861	.900	−.562	−.611

Tabela 4.28: Matriz (II) das Correlações de pesos baixos PAB¡30 e PFE¡3210

A análise da variância que se mostra na tabela 4.31 apresenta valores próximos dos obtidos no caso visto anteriormente de uma única variável independente (PAB).

A tabela 4.32 mostra as correlações parciais.

Regressão Linear		
Numero de Casos (C)	46	
Número de variáveis(V)	11	
Coeficiente de correlação	0.93 (0.92)	
Coeficiente de determinação	0.87 (0.85)	
Coeficiente de determinação ajustado	0.83 (0.84)	
Análise da variância da regressão	$F(10, 35) = 23$	$p < 0.0000$
Erro padrão dos valores previstos	234 (228)	

Tabela 4.29: Sumário dos resultados da regressão lin. com todas as var. explic.

Regr.Lin.	Coeficiente	Erro Padrão	$t(35)$	$p <$
intersecção	−4176 (−3153)	1216 (373)	−3.4	0.0016
PMA	1.21	3.274	0.3	0.713
AMA	4.46	6.7	0.6	0.511
IGE	27.2	17.8	1.5	0.13
DEN	5.06	20.3	0.2	0.80
DBP	166.5	134.2	1.2	0.22
PCE	−43.7	43.2	−1.0	0.31
PAB	82.6 (83)	30.8 (24)	2.6 (3)	0.0110 (.000)
CFE	354.5 (464)	122.6 (82)	2.8 (5)	0.0065 (.000)
IPU	4.92	48.2	0.6	0.52
IRU	290.9	446.6	−0.6	0.51

Tabela 4.30: Sumário dos resultados (para n=47) da regressão lin. com var. explic. PAB (entre parentsis para n=414)

Efeito	Soma Quad.	df	Média Quad.	F	valor$-p$
Regr.	1.0×10^7	10	1.3×10^6	23.7	.000
Resid.	1.9×10^6	35	5.5×10^4		
Total	1.5×10^7				

Tabela 4.31: Análise da variância da regressão pesos baixos (n=46)

Var	CoefNorm	CorrPar	CorrSemiPar	Toleranc
DBP	.1921	.2052	.0751	.1531
PCE	−.1639	−.1687	.1626	.2076
PAB	.3570 (.3618)	.4132 (.4590)	.1626 (.1994)	.2076 (.3038)
CFE	.4572 (.5987)	.4390 (.6498)	.1751 (.3300)	.1468 (.3038)
IRU	−.0621	−.1094	−.0394	.4030

Tabela 4.32: Correlações Parciais (I) pesos baixos (n=46)

Var	CoefNorm	R^2	$t(43)$	valor$-p$
DBP	.1921	.8468	1.240	.223
PCE	$-.1639$	.7923	2.684	.318
PAB	.3570 (.3618)	.7923 (.6961)	2.684 (3.388)	.0110 (.0015)
CFE	.4572 (.5987)	.8531 (.6961)	2.890 (5.606)	.0065 (.000)
IRU	$-.0621$	.5969	$-.6512$	.519

Tabela 4.33: Correlações Parciais (II) pesos baixos (n=46)

4.2.9 Efeito na regressão dos erros nas medidas

Nesta subsecção analisa-se o efeito da presença de erros nas medidas, sejam na variável explicada Y, sejam nas variáveis explicativas X_i, seguindo muito de perto Neter et al. (1996) (pp.164-166). Quando nalguns casos existem erros aleatórios nas medidas da variável explicada Y, tal não deve em geral constituir um problema sério se esses erros forem pequenos, não originarem *outliers*, não estiverem nem correlacionados nem enviesados, uma vez que os erros de medida positivos e negativos tendem a cancelar-se. Por exemplo, se o peso de um recém nascido tiver sido medido ocasionalmente com um pano cujo peso não foi descontado, o erro de medida resultante, não sendo sistemático, é absorvido nos termos de erro do modelo ε_i :

$$Y_i = w_0 + w_1 X_i + \varepsilon_i.$$

Este erro aditivo e positivo pode ser compensado por um outro erro negativo, tal como, por exemplo, o erro que se comete quando se segura um recém nascido para não cair da balança e se diminui ocasionalmente o peso. O erro ε reflecte os efeitos compostos de um grande número de factores não considerados no modelo, os quais serão neste caso imprecisões ocasionais durante o processo de medida.

A situação é muito diferente quando os erros nas medidas ocorrem nas variáveis independentes (ou,explicativas) X. O valor de uma variável independente pode nunca ser susceptível de erro, como, por exemplo, quando se exprime o número de dias de intervalo entre o momento do exame ecográfico e o nascimento da criança, que, na investigação aqui descrita é um número sempre inferior a oito dias, não sendo muito provável que possa vir afectado de erros. Porém a situação pode ser muito diferente quando se trata do valor de uma medição de um PAB em ecografia. Esta medida pode ser susceptível de alguns erros aleatórios devido a: (i) Má visibilidade provocada pelo líquido aminoótico ser insuficiente; (ii) Obliquidade na secção obtida no abdomen do feto por este estar numa posição inapropriada para a ecografia; (iii) Imprecisão na extrapolação de pontos do círculo abdominal que não foram captados pela ecografia porque o feto tem dimensões superiores ao alcance do varrimento ecográfico. Supondo que A_i designa o valor verdadeiro do PAB do $i-ésimo$ caso, e que A_i^* representa um valor que foi medido com um erro desconhecido $\delta_i = A_i^* - A_i$, este tem a seguinte repercussão no modelo de regressão:

$$\begin{aligned} P_i &= w_0 + w_1 A_i + \varepsilon_i \\ &= w_0 + w_1 (A_i^* - \delta_i) + \varepsilon_i \\ &= w_0 + w_1 A_i^* + (\varepsilon_i - w_1 \delta_i) \end{aligned} \tag{4.9}$$

Este modelo é muito diferente do que se obtém de uma regressão trivial, porque

a variável independente A_i^* é uma variável aleatória que está correlacionada com o termo de erro $(\varepsilon_i - w_1\delta_i)$, enquanto que no modelo de regressão trivial se sabe que o termo de erro deve ser independente da variável independente. Por consequência os resultados do modelo podem aparecer falseados. Para evitar que isto aconteça pode-se tentar determinar a dependência formalmente começando por fazer as seguintes três suposições simples:

$$\mathbf{E}[\delta_i] = 0 \tag{4.10}$$

$$\mathbf{E}[\varepsilon_i] = 0 \tag{4.11}$$

$$\mathbf{E}[\delta_i \varepsilon_i] = 0 \tag{4.12}$$

Repare-se que a condição (4.10) implica que

$$\mathbf{E}[A_i^*] = \mathbf{E}[A_i + \delta_i] = A_i,$$

de tal forma que no exemplo dos pesos fetais os $PABs$ medidos seriam estimativas não enviesadas das verdadeiros $PABs$. A condição (4.11) constitui um requisito usual ao qual as parcelas de erro do modelo, ε_i, têm obrigatoriamente que satisfazer. Finalmente, a condição (4.12) requer que a medida do erro δ_i não esteja correlacionada com o erro do modelo ε_i, uma vez que a covariância para ser nula tem que se verificar:

$$\sigma[\delta_i, \varepsilon_i] = \mathbf{E}[\delta_i \varepsilon_i] - \mathbf{E}[\delta_i]\,\mathbf{E}[\varepsilon_i] = 0$$

Determinando-se, agora, a covariância entre as observações A_i^* e os termos aleatórios $(\varepsilon_i - w_1\delta_i)$ do modelo sob as condições (4.10), (4.11) e (4.12), as quais implicam que $\mathbf{E}[A_i^*] = A_i$, e que $\mathbf{E}[\varepsilon_i - w_1\delta_i] = 0$, obtém-se:

$$\begin{aligned}
\sigma[A_i^*, (\varepsilon_i - w_1\delta_i)] &= \mathbf{E}[(A_i^* - \mathbf{E}[A_i^*])\,((\varepsilon_i - w_1\delta_i) - \mathbf{E}[\varepsilon_i - w_1\delta_i])] \\
&= \mathbf{E}[(A_i^* - A_i)\,(\varepsilon_i - w_1\delta_i)] \\
&= \mathbf{E}[\delta_i\,(\varepsilon_i - w_1\delta_i)] \\
&= \mathbf{E}\left[\delta_i\varepsilon_i - w_1\delta_i^2\right]
\end{aligned}$$

Uma vez que $\mathbf{E}[\delta_i\varepsilon_i] = 0$ em virtude de (4.12), e $\mathbf{E}\left[\delta_i^2\right] = \sigma^2[\delta_i]$, obtém-se:

$$\sigma[A_i^*, (\varepsilon_i - w_1\delta_i)] = -w_1\sigma^2[\delta_i] \tag{4.13}$$

Esta covariância não se anula sempre que existir uma relação de regressão linear entre P e A.

Quando se aplicam os procedimentos habituais da regressão ao modelo (4.9), os estimadores w_0 e w_1 irão ficar enviesados e perdem a propriedade da consistência.

Poderão surgir grandes dificuldades para desenvolver estimadores não enviesados quando existem erros nas medidas de A. Uma das soluções sugeridas por Neter et al. (1996) (p.165) consiste em impor severas restrições ao problema, por exemplo, definir muito bem as propriedades da distribuição de δ_i e ε_i. Uma outra solução sugerida consiste em utilizar outras variáveis adicionais que se sabe que estão relacionadas com o verdadeiro valor de A mas não com os erros de medida δ. Estas variáveis são designadas *variáveis instrumentais* porque se utilizam para estudar a relação entre P e A.

Esta análise poderia aplicar-se, por exemplo, para fazer uma avaliação dos erros das variáveis diâmetro biparietal DBP e perímetro cefálico PCE, uma vez que estão funcionalmente relacionados. Uma secção cefálica é aproximadamente elíptica e alguns ecógrafos calculam o perímetro por ajustamento de uma elipse a alguns pontos. Por sua vez, o valor elevado para a correlação entre as variáveis DBP e PCE pode explicar-se atendendo a este mesmo facto. Também se poderia aplicar para fazer um estudo nos possíveis erros de medida detectados nalguns alguns casos, que apresentam diferenças significativas entre os perímetros cefálicos medidos antes e depois do nascimento. A correlação entre a variável perímetro cefálico PCE medido pela ecografia e a variável perímetro cefálico PCR medido no recém-nascido (passados 1 a 7 dias) apresenta um valor 0.78, inferior ao que seria de esperar. Uma explicação possível supondo que não existem eventuais erros introduzidos na medida da variável PCE, segundo Pereira-Leite (2000), consiste em que as diferenças podem ser devidas a um aumento temporário do volume cefálico causado durante o trabalho de parto, assim como a uma descompressão pós-parto (ver Figura 4.29).

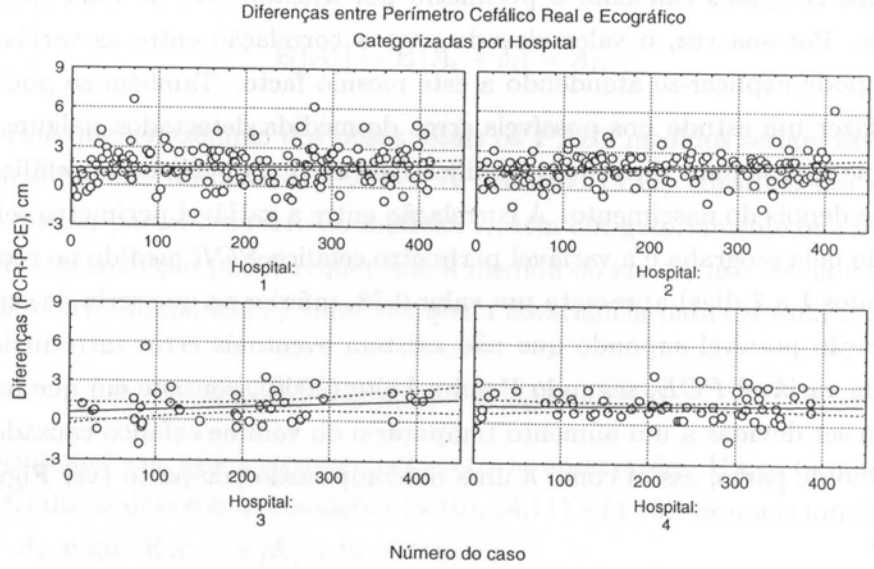


Figura 4.29: Perímetros cefálicos medidos no recém-nascido (PCR) e por ecografia fetal (PCE).

4.3 Tópicos sobre o modelo de regressão linear

4.3.1 Esperança condicional

Apresenta-se, nesta subsecção, uma revisão das razões pelas quais se sabe ser linear uma função de regressão de uma variável sobre outra variável, quando as distribuições de probabilidade destas variáveis são gaussianas. Este facto tem importância uma vez que sob certas condicionantes a soma de variáveis aleatórias independentes, não tendo que ter necessariamente distribuições normais, tende para uma distribuição normal quando o número de variáveis tende para infinito (ver *Teorema Limite Central,* Murteira et al. (2001)). A definição teórica de curva de regressão é o lugar geométrico dos pontos médios ou, mais precisamente, a esperança matemática da função densidade de probabilidade da variável aleatória Y condicionada à variável aleatória X, isto é, $P(Y \mid X)$. Portanto, pode escrever-se a função de regressão $f\ (X\ )$ pela seguinte expressão:

$$f\ (X\ ) = \mathbf{E}\,(Y \mid X) = \int_{-\infty}^{+\infty} Y\, P(Y \mid X) dY = \int_{-\infty}^{+\infty} Y\, \frac{P(X\ ,\ Y)}{P\,(X)} dY =$$
$$= \frac{\int_{-\infty}^{+\infty} Y\, P(Y\ ,\ X) dY}{P\,(X)},$$

em que:

$$P\,(X) = \int_{-\infty}^{+\infty} P(Y\ ,X) dY$$

é uma das funções de densidade de probabilidade marginal que se pode determinar a partir de$P(Y\ ,X)$.

Pode considerar-se um modelo de distribuição de probabilidade gaussiana, sem ruído aditivo, definindo por duas variáveis aleatórias, que se vão designar por p, peso fetal (PFE) e se supõe satisfazer uma distribuição $\mathcal{N}(3225, 628)$, e por a, perímetro abdominal (PAB), que se supõe satisfazer uma distribuição $\mathcal{N}(33, 3)$. Supõe-se que a dependência entre as variáveis PFE e PAB é caracterizada pelo coeficiente de correlação $\rho_{p,a}$, e que este tem o valor 0.85 (que se retirou da tabela 4.7). A função de distribuição de probabilidade conjunta $P_{a,p}$ é dada pela expressão(4.14). Substituindo os valores dos parâmetros

$$\mu_a = 33,\ \sigma_a = 3,\ \mu_p = 3225,\ \sigma_p = 628,\ \rho_{p,a} = 0.85,$$

obtém-se:

$$P_{a,p} = 1.603\,6 \times 10^{-4} \exp\left(-1.801\,8 \times \left(\left(\frac{p-3225}{628}\right)^2 - 2 \times 0.85 \times \frac{(p-3225)}{628} \times \frac{(a-33)}{3} + \left(\frac{a-33}{3}\right)^2\right)\right)$$

cuja representação gráfica tridimensional pode ver-se na Figura 4.30 :

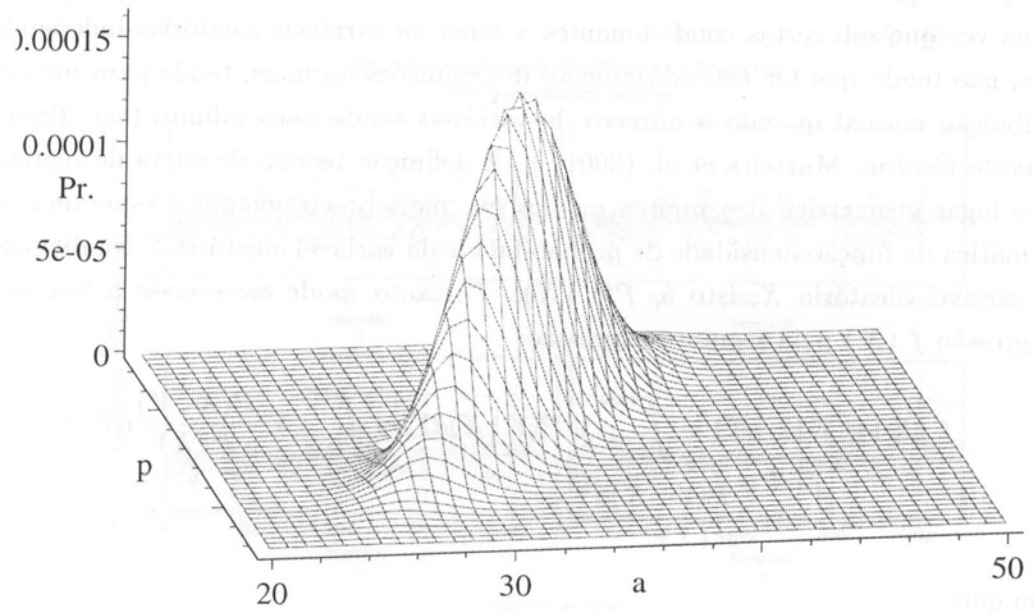


Figura 4.30: Representação gráfica tridimensional de $P(a,p)$.

A figura 4.31 representa as curvas de nível (P = *constante*) desta distribuição projectadas no plano a,p, e que ilustram as experiências históricas de Galton.

A densidade marginal $P(a)$, uma vez que se sabe os parâmetros $N\left(\mu_a, \sigma_a^2\right)$, não precisa de ser calculada pelo integral $P(a) = \int_{-\infty}^{+\infty} P(p\,, a)dp$, podendo obter-se directamente da expressão da densidade normal univariável:

$$P_a = \frac{1}{\sigma_a\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_a)^2}{2\sigma_a^2}\right) = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{a-33}{2\times 3}\right)$$

Logo a densidade de probabilidade do peso fetal, p, condicionada ao perímetro abdominal, a,exprime-se por:

$P(p|a) = \frac{P(p,a)}{P(a)}$

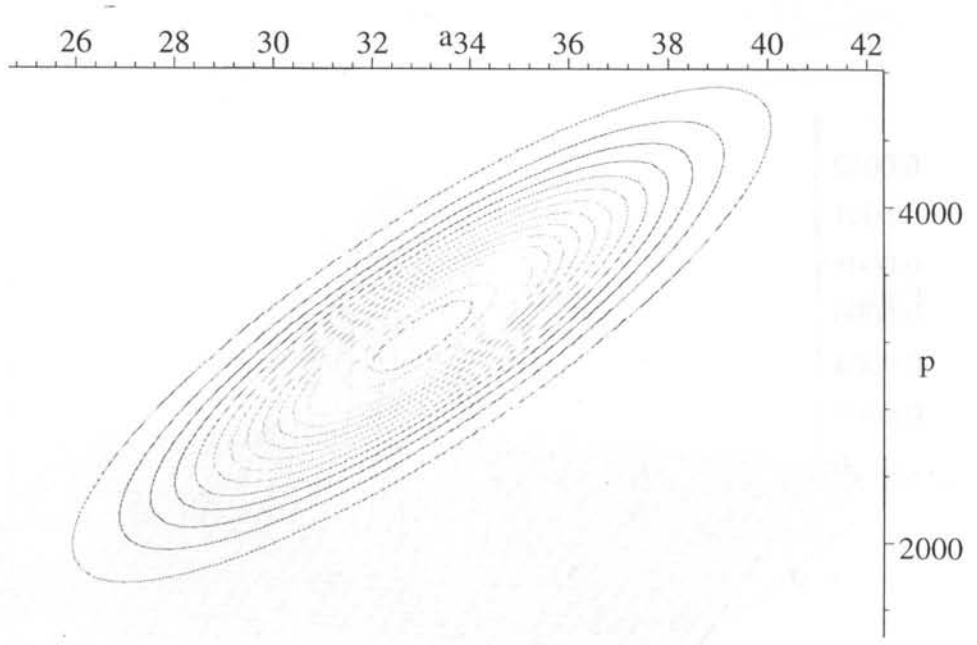


Figura 4.31: Representação das curvas de nível (P = *constante*) da densidade $P(a,p)$ projectadas no plano a,p,

$$= \left[1.60 \times 10^{-4} \exp\left(-1.8 \times \left(\left(\tfrac{p-3225}{628}\right)^2 - 2 \times 0.85 \times \tfrac{(p-3225)}{628} \times \tfrac{(a-33)}{3} + \left(\tfrac{a-33}{3}\right)^2\right)\right)\right] \div$$

$$\div \left[\tfrac{1}{3\sqrt{2\pi}} \exp\left(-\tfrac{a-33}{2\times 3}\right)\right]$$

$$= 1.2\times 10^{-3} \exp\left(-1.8 \times \left(\left(\tfrac{p-3225}{628}\right)^2 - 1.7 \times \tfrac{(p-3225)}{628} \times \tfrac{(a-33)}{3} + \left(\tfrac{a-33}{3}\right)^2\right) - \left(-\tfrac{a-33}{2\times 3}\right)\right)$$

$$= 1.2 \times 10^{-3} \exp\left(-1.8 \left(\tfrac{1}{628}p - \tfrac{3225}{628}\right)^2 + 3.063 \left(\tfrac{1}{628}p - \tfrac{3225}{628}\right)\left(\tfrac{1}{3}a - 11\right) -\right.$$

$$\left. -1.801\,8 \left(\tfrac{1}{3}a - 11\right)^2 + \tfrac{1}{6}a - \tfrac{11}{2}\right)$$

Dois pontos de vista da representação tridimensional estão representados nas figuras 4.32 e 4.33, as quais procuram sugerir visualmente que os valores esperados do peso fetal p se encontram sobre a projecção no plano (p,a) da recta que une os pontos máximos, isto é, os valores médios, das gaussianas, para os vários valores de a constantes, e que estes ponto deverão ser colineares.

$$E\left(p|a\right) = \int_{-\infty}^{+\infty} p \times 1.205\,9 \times 10^{-3} \times \exp\left(-1.801\,8\left(\frac{1}{628}p - \frac{3225}{628}\right)^2 +\right.$$

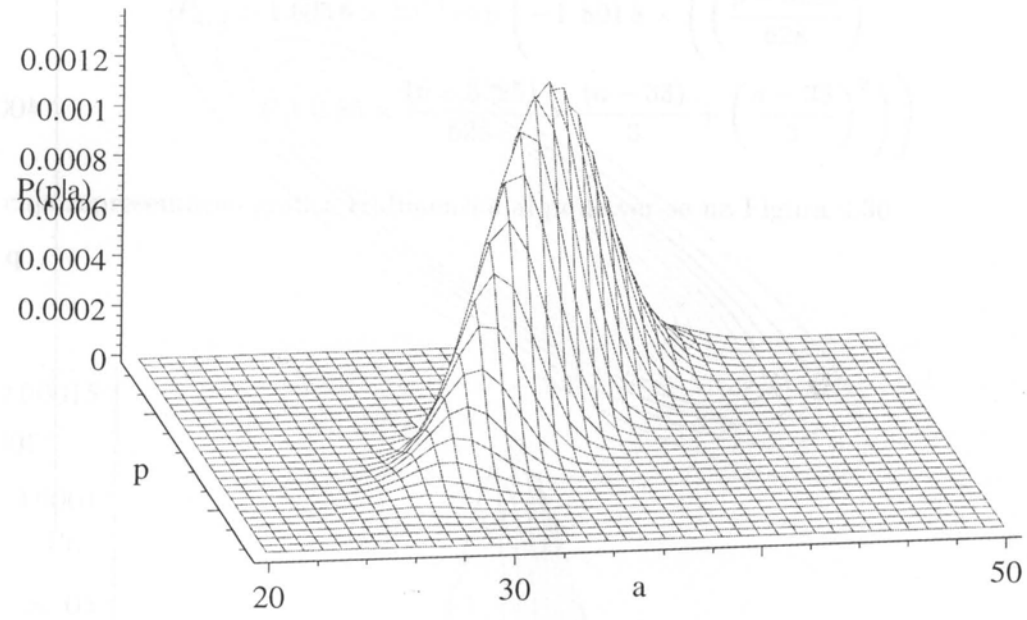


Figura 4.32: Representação I da densidade de probabilidade condicional $P(p|a)$

$$+3.063\,1\left(\frac{1}{628}p-\frac{3225}{628}\right)\left(\frac{1}{3}a-11\right)-1.801\,8\left(\frac{1}{3}a-11\right)^2+\frac{1}{6}a-\frac{11}{2}\Bigg)dp$$
$$=2.0\times10^{-7}\left(-1.3\times10^{10}+8.9\times10^{8}a\right)$$

4.3.2 Variância condicional

Proporção da variância explicada. Pode sempre calcular-se, independentemente do modelo, a variância total da variável dependente (soma dos quadrados total, SQT), a proporção da variância devida aos resíduos (soma dos quadrados dos erros, SQE) e a proporção da variância devida ao modelo de regressão (soma dos quadrados da regressão, $SQR = SQT - SQE$). A razão entre a soma dos quadrados da regressão e a soma dos quadrados total (SQR/SQT) explica a proporção da variância devida à variável dependente do modelo, portanto, esta razão é equivalente ao coeficiente de determinação. Mesmo quando a variável dependente não está normalmente distribuída ao longo dos casos esta medida pode permitir avaliar quando o modelo se ajusta bem aos dados Murteira et al. (2001) (p.519).

Aproximação condicional. Sendo dadas duas variáveis aleatórias Y e X

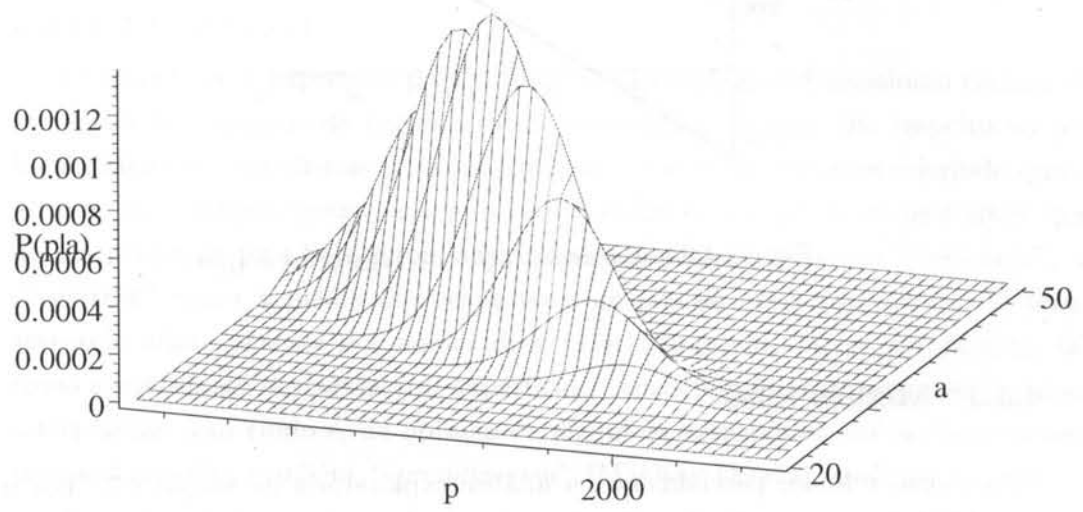


Figura 4.33: Representação II da densidade de probabilidade condicional $P(p|a)$

pretende-se obter uma função $f(X)$ que seja tão próxima de Y quanto for possível em média quadrática. Sabe-se que $f(X)$ realiza o mínimo de $\mathbf{E}\left[(Y - f(X))^2\right]$ porque $\mathbf{E}[Y|X]$ é a projecção ortogonal de Y sobre o sub-espaço de Hilbert fechado e convexo, L^2_X, constituído por todas as variáveis aleatórias funções apenas de X e que sejam do tipo $f(X)$, espaço que contém a recta Δ das constantes ($\mathbf{E}[Y]$). A qualidade da aproximação de Y por pode medir-se em relação com respeito à correlação

$$R^2 = \frac{Var[\mathbf{E}[Y|X]]}{Var[Y]} = \frac{\text{Var explicada}}{\text{Var total}} = \cos^2\theta.$$

A função que para uma dado valor x da variável aleatória X associa $\mathbf{E}[Y|X] = x$ chama-se *função de regressão* de Y sobre X e o seu gráfico é a curva de regressão de Y sobre X. Pode, portanto, escrever-se que a expressão, $Y = \mathbf{E}[Y|X] + \varepsilon$, sendo ε um resíduo aleatório (ou ruído) em geral não desprezível, e que satisfaz necessariamente a propriedade de ser $\mathbf{E}[\varepsilon] = 0$. Mais ainda, ε está correlacionado linearmente com X e com $\mathbf{E}[Y|X]$ porque ε é perpendicular a L^2_X. A variância de ε, ou variância residual, é então $Var[\varepsilon] = Var[Y] - Var[\mathbf{E}[Y|X]]$, logo, verifica-se:

$$Var[\varepsilon] = \left(1 - R^2\right) Var[Y].$$

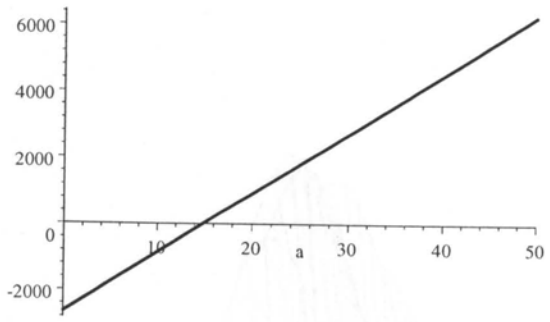


Figura 4.34: Representação da regressão $E(p|a)$

4.3.3 Modelo linear

Esta secção retoma, parcialmente, a análise exploratória da secção 4.2. Por mais paradoxal que possa parecer a *teoria matemática das probabilidades* e a *estatística* são campos das ciências matemáticas que podem apresentar diferenças significativas Moore and McCabe (1999) (p.xviii). Os dados são números num contexto definido e, para estes se integrarem nos resultados relevantes da teoria matemática das probabilidades, particularmente em teoremas sobre a convergência e consistência, é preciso realizar-se alguma experimentação. A teoria das probabilidades, como se sabe, tem uma fundamentação axiomática, Kolmogorov (1950) (p.v), o que lhe permite ter uma grande riqueza e generalidade, mas, ao mesmo tempo, torna problemático ter-se uma intuição clara de que os raciocínios e as deduções são verdadeiras (por exemplo, sente-se isto nas demonstrações dos teoremas em Vapnik (1998), tal como se refere na discussão epistemológica do último capítulo).

A ideia da análise de regressão surge quando se estuda a relação existente entre um conjunto de pares de dados $\{(x_i, y_i)\}$ e se verifica que os gráficos de dispersão dos (x_i, y_i) que têm a mesma frequência formam séries de elipses concêntricos (Figura). Esta foi a descoberta de Galton (1822 – 1911) que sugeriu a expressão da distribuição normal bidimensional:

$$P(X,Y \mid \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{\left(1-\rho^2\right)^{1/2}}{2\pi\sigma_X\ \sigma_Y} \times \tag{4.14}$$

$$\times \exp\left(-\frac{\left(1-\rho^2\right)^{-1}}{2} \times \left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)}{\sigma_X} \times \frac{(y-\mu_Y)}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

em que:

$$P(X,Y \mid \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$$

representa a função densidade da variável aleatória bidimensional (X, Y), dados:

$$E(X) = \mu_X,\ E(Y) = \mu_Y,\ Var(X) = \sigma_X^2,\ Var(Y) = \sigma_Y^2,$$

e em que ρ que designa o coeficiente de correlação entre X e Y (por outras palavras, $Cov(X, Y) = \rho\,\sigma_X\,\sigma_Y$).

Utilizando esta expressão para a distribuição normal bidimensional Galton desenvolveu os conceitos de regressão e de correlação. No que diz respeito ao problema histórico das alturas crianças exprimiu as suas descobertas referindo que as alturas das crianças, quer elas sejam provenientes de um grupo de pais altos, quer de um grupo de pais baixos, mostravam uma tendência para um "*retrocesso*", ou "*regressão*", para a média dos respectivos grupos de onde provinham. O termo *regressão* ainda persiste nos nossos dias, existindo porém outras designações, tais como ajustamento de curvas (ou superfícies), aproximação, estimação de um sinal, interpolação com ruído e, na opinião de alguns investigadores em modelos de rede neuronal significa também "*aprendizagem*" (MacKay (1992a), p.15).

A análise de regressão pode dizer-se que, citando Neter et al. (1996), tem três aplicações importantes que frequentemente se sobrepõem: (i) *descrição*, (ii) *controlo* e (iii) *previsão*. Na investigação descrita nesta dissertação já se explicou que se está mais interessado na terceira aplicação, uma vez que o objectivo é a utilização na prática clínica obstétrica de medidas biométricas obtidas por ecografia do feto na previsão do peso fetal.

Um modelo de regressão exprime formalmente a existência duma dada relação estatística entre um "*regressando*", isto é, um conjunto de m observações da variável dependente, ou explicanda, y, e um "*regressor*", isto é, um conjunto de m observações das variáveis independentes ou regressoras $\mathbf{x}$. A terminologia "regressando / regressor" aparece, por exemplo, em Murteira et al. (2001), p.501. Uma variável regressanda pode ter um conjunto de m valores medidos, por exemplo pesos fetais, e uma variável regressora, pode ter um conjunto valores de igual cardinal, por exemplo medidas do perímetro biparietal ou do comprimento do fémur.

A relação que se define através de um modelo de regressão não tem que ser necessariamente de causalidade, como acontece frequentemente na relação entre variáveis definidas por fenómenos físicos. Por exemplo, a relação de causalidade existente entre um pequeno incremento Δy de distensão duma mola e o incremento do peso Δp por ela suportado, coincide com uma relação funcional (linear neste caso) entre estas duas variáveis Δy e Δp.

Mas, se se tratasse de um incremento do perímetro abdominal (PAB) fetal facilmente se compreende que um maior ou menor volume fetal pode ser causado por um "mecanismo" genético programado, ou pela quantidade de energia rece-

bida através da placenta, ou pelo espaço máximo possível num determinado útero. Pode observar-se que a taxa de variação do PAB é constante até à 30^a semana, podendo estabelecer-se uma dependência linear entre o peso e o número de semanas de gestação. Mas, a partir de certa altura a taxa de variação do peso fetal passa a ser muito menor originado uma lei curvilínea com um máximo por volta da 42^a semana (mais rigorosamente no momento do nascimento, ver Figura 4.35).

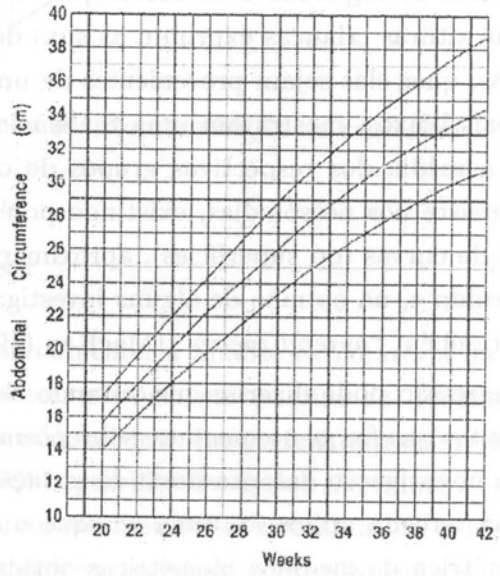


Figura 4.35: Evolução do perímetro abdominal em função do número de semanas de gestação

No caso de existir apenas uma única variável independente X, pode escrever-se a seguinte expressão que nos dá os vários valores da variável dependente:

$$Y_i = w_0 + w_1 X_i + \varepsilon_i.$$

A análise de regressão para um modelo de erro gaussiano é aplicável mesmo que X seja uma variável aleatória desde que as distribuições condicionais de Y obedeçam a determinadas condições e as distribuições marginais de X não envolvam os parâmetros do modelo de regressão w_0, w_1e σ^2. Portanto, no caso de X ser uma variável aleatória, só é preciso especificar as distribuições condicionais de Y e fazer uma restrição à distribuição marginal de X. A parcela de erro ε_i deverá ter uma esperança matemática nula, a sua variância deverá ser constante e não estar correlacionada com outras variáveis (caso existirem).

Um modelo muito simples de regressão linear múltipla, pode ser dado pela expressão:

$$\mathbf{y} = \mathbf{w}\,\mathbf{x} + \varepsilon$$

em que

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad x = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{m1} & \cdots & x_{m,p-1} \end{bmatrix},$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{p-1} \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}.$$

Pressupõe-se que as variáveis independentes $X_1, \ldots, X_{n-1}$são conhecidas, e que o interesse primário consiste em fazerem-se inferências acerca da variável explicanda Y com base naquelas variáveis preditivas Neter et al. (1996).

4.3.4 Modelos linear generalizado e não-linear

Quando empírica, ou teoricamente, se verifica a existência de razões para considerar inadequados os modelos de regressão lineares, então, podem utilizar-se os modelos de regressão linear generalizados e os modelos de regressão não lineares. Os primeiros modelos são lineares nos parâmetros, a designação " modelos lineares" também abrange modelos de regressão polinomiais numa ou mais variáveis independentes, desde que sejam lineares nos parâmetros. Por exemplo, o seguinte modelo com duas variáveis independentes tem termos lineares, quadráticos e de interacção:

$$Y_i = w_0 + w_1 X_{i1} + w_2 X_{i1}^2 + w_3 X_{i2} + w_4 X_{i2}^2 + w_5 X_{i1} X_{i2} + \varepsilon_i$$

Igualmente, pertencem à classe dos modelos de regressão (alguns autores designam-nos por " modelos lineares generalizados", por exemplo McCullagh and Nelder (1997)) os modelos com variáveis transformadas que são lineares nos parâmetros, como por exemplo:

$$\log_{10} Y_i = w_0 + w_1 \sqrt{X_{i1}} + w_2 \exp\left(X_{i2}\right) + \varepsilon_i$$

Em geral, pode representar-se um modelo de regressão não linear na forma Neter et al. (1996):

$$Y_i = f\left(\mathbf{X}_i, \mathbf{w}\right) + \varepsilon_i$$

em que $\mathbf{X}_i$ é o vector das observações nas variáveis preditivas par $i-ésimo$ caso:

$$\mathbf{X}_i = \left[\begin{array}{cccc} 1 & X_{i1} & \cdots & X_{i,p-1} \end{array} \right]^T,$$

em que $\mathbf{w}$ é o vector dos coeficientes de regressão, a função resposta $f(\mathbf{X}_i, \mathbf{w})$ representa o valor esperado $E\{Y_i\}$ e os ruídos ε_i são independentes e satisfazem $N(0, \sigma^2)$. No caso dos modelos de regressão lineares verifica-se ainda que

$$f(\mathbf{X}_i, \mathbf{w}) = \mathbf{X}_i^T . \mathbf{w}.$$

Os modelos de regressão *não-lineares* mais usados são o modelo *exponencial*, o qual tem como função resposta a expressão:

$$f(\mathbf{X}, \mathbf{w}) = w_0 + w_1 \exp(w_2 \mathbf{X}),$$

e o modelo *logístico*, cuja função resposta tem a expressão:

$$f(\mathbf{X}, \mathbf{w}) = w_0 / (1 + w_1 \exp(w_2 \mathbf{X})).$$

4.4 Máxima verosimilhança e estimação paramétrica Bayesiana

Faz-se nesta secção uma introdução ao ponto de vista bayesiano da regressão, para uma abordagem mais detalhada pode consultar-se, por exemplo, MacKay (1992b) (pp.15-24), e Murteira (1988). Este ponto de vista surgiu da teoria da decisão, Berger (1985), e a partir dos anos 1980 passou a ter uma importância relevante na interpretação dos modelos de rede neuronal.

Supõe-se que se tem um conjunto $\mathbf{Z}$ de m de exemplos a serem interpolados, $\mathbf{Z} = \{\mathbf{x}_i, \mathbf{y}_i\}$, $i = 1, 2, \ldots, m$, e admite-se, por simplicidade, que $\mathbf{x}$ *e* $\mathbf{y}$ são escalares. O método pode ser generalizado ao caso multidimensional. Para se definir um modelo de *interpolação* (ou, regressão) linear define-se um conjunto de funções $\mathbb{H} = \{\phi_h(x)\}$ de k funções de base, a partir das quais se consideram funções interpoladoras dadas pelo desenvolvimento em série, ou combinação linear:

$$h(x) = \sum_{j=1}^{k} w_j \phi_j(x)$$

em que os parâmetros w_j se inferem a partir dos dados.

Modeliza-se o conjunto de dados $\mathbf{Z}$ considerando os desvios entre os dados e os valores obtidos através desta correspondência funcional, $h(x)$, que poderão ser

causados por algum processo de ruído aditivo desconhecido:

$$y_i = h(x_i) + \varepsilon_i$$

Se o ruído ε_i for modelizado por um ruído gaussiano de média nula $N\left(0, \sigma^2\right)$, então, a probabilidade de se verificarem os y_i, dados os parâmetros $\mathbf{w}$, isto é,

$$P\left(\{y_i\} \mid \{x_i\}, \mathbf{w}, \beta, \mathbb{H}, \mathbb{N}\right)$$

chama-se *verosimilhança.*

Representa-se simplesmente por $P(\mathbf{Z}|\mathbf{w}, \beta, \mathbb{H}, \mathbb{N})$, é dada por:

$$P(\mathbf{Z}|\mathbf{w}, \beta, \mathbb{H}, \mathbb{N}) = \frac{\exp\left(-\beta E_{\mathbf{Z}}(\mathbf{Z}|w, \mathbb{H})\right)}{k_{\mathbf{Z}}(\beta)}$$

em que

$$\beta = \frac{1}{\sigma^2}; \quad E_{\mathbf{Z}} = \sum_{i=1}^{m} \frac{1}{2}\left(h(x_i) - y_i\right)^2; \quad e \quad k_{\mathbf{Z}}(\beta) = \left(\frac{2\pi}{\beta}\right)^{m/2}$$

O valor de $\mathbf{w}$ que minimiza $E_{\mathbf{Z}}$ pode ser indeterminado, ou, então, pode apresentar uma dependência com grande sensibilidade em relação aos pormenores do ruído subjacente aos dados. A máxima verosimilhança vai, neste caso, ter grandes oscilações procurando sempre um melhor ajustamento aos dados. Para evitar este efeito nocivo introduz-se no modelo interpolante uma quantidade R, definida *a priori*, para exprimir o tipo de suavização que se espera que tenha o interpolante $h(x)$. Esta quantidade R , sugerida em MacKay (1992b) (p.15), tem a forma:

$$P(h|R, \alpha) = \frac{\exp\left(-\alpha E_h(h|R)\right)}{Z_h(\alpha)}$$

em que E_h pode ser um funcional, $E_h = \int h''\left(x^2\right) dx$ (o qual é um regularizador para uma "*spline*" cúbica.) O parâmetro α é uma medida da suavidade que se espera $h(x)$ ter. Esta quantidade *a priori* pode também escrever-se como função dos parâmetros $\mathbf{w}$:

$$P(\mathbf{w}|\alpha, \mathbb{H}, \mathbb{R}) = \frac{\exp\left(-\alpha E_W(\mathbf{w}|\mathbb{H}, \mathbb{R})\right)}{k_W(\alpha)},$$

em que:

$$k_W = \int d^k\mathbf{w} \exp\left(-\alpha E_W\right),$$

e E_W se designa frequentemente por *função regularizadora.* O modelo de interpolação consiste, pois, na escolha do conjunto de funções de base $\mathbb{H}$, num modelo gaussiano N com um parâmetro β para o ruído, e um regularizador *a priori,* R, com

uma constante de regularização α.

Se α e β forem conhecidos, então a probabilidade a posteriori dos parâmetros $\mathbf{w}$ é:

$$P(\mathbf{w}|\mathbf{Z}, \alpha, \beta, \mathbb{H}, \mathbb{R}, \mathbb{N}) = \frac{P(\mathbf{Z}|\mathbf{w}, \beta, \mathbb{H}, \mathbb{N})\ P(\mathbf{w}|\mathbb{H}, \mathbb{R})}{P(\mathbf{Z}|\alpha, \beta, \mathbb{H}, \mathbb{R}, \mathbb{N})}$$
$$= \frac{\exp(-(\alpha E_W + \beta E_\mathbf{Z}))}{\int d^k\mathbf{w} \exp(-(\alpha E_W + \beta E_\mathbf{Z}))}$$

A minimização da função objectivo combinada:

$$M = \alpha E_W + \beta E_\mathbf{Z}$$

corresponde a determinar-se o interpolante mais provável, $\mathbf{w}_{MP}$. As curvas limites do erro (*error bars*) podem obter-se pelo Hessiano de M,

$$\mathbf{A} = \bigtriangledown \bigtriangledown M,$$

calculado para $\mathbf{w}_{MP}$.

Esta perspectiva bayesiana da regularização também é conhecida por "*máxima verosimilhança penalizada*" ou "*regressão de crista*" (*ridge regression*, Poggio and Girosi (1990)). Este conceito de regularização é o mesmo que foi referido na secção em que se abordam elementos da teoria da regularização do Capítulo 3.

A aprendizagem realizada com modelos de rede neuronal podem ser do tipo supervisado ou do tipo não-supervisado. A distinção entre estes dois tipos de aprendizagem tem vindo a esbater-se muito recentemente. Por exemplo, são híbridos os modelos de redes de FBR que se utilizam nos ensaios que se descrevem no capítulo 5. A aprendizagem da camada escondida é *não-supervisada* e a da camada de saída é *supervisada*. Em qualquer dos casos surge sempre o problema de como se devem interpretar as saídas. Esta opção não existe quando os dados têm etiquetas dicotómicas, por exemplo, quando o peso fetal ou é baixo ou alto. Ou, então, têm etiquetas com múltiplas classes, por exemplo, se se acrescentarem ao exemplo anterior uma terceira classe, de pesos médios. Porém, quando os dados de treino têm etiquetas que são *números reais*, são portanto representáveis por funções contínuas reais, de variáveis reais, a aprendizagem pode fazer-se usando uma função de custo erro médio quadrático, mas, depois de as redes estarem treinadas as suas saídas para casos novos podem ser interpretadas ou como valores reais, ou como classes às quais se podem afectar margens de confiança. Este tema que faz parte da concepção ("*design*") de uma rede neuronal é muito vasto pelo que se faz aqui uma revisão de alguns trabalhos relevantes nesta área.

4.4.1 Aprendizagem Bayesiana

Aborda-se seguidamente o problema da determinação de um modelo de uma função densidade de probabilidade $p(\mathbb{X})$, para uma dado conjunto de entradas com n variáveis, e um conjunto finito de exemplos, $\mathbf{x}_i \in \mathbb{X}$, $i = 1, \ldots, m$, $(\mathbb{X} \subset \mathbb{R}^n)$, extraídos num conjunto de amostragem com esta distribuição de probabilidade.

Um dos métodos que se pode utilizar nesta modelização é o da máxima verosimilhança Bishop (1995) (p.40). Neste método começa-se por tomar uma decisão sobre qual a forma paramétrica da função de densidade $p(\mathbb{X})$. Depois, determinam-se os valores óptimos dos parâmetros $\theta = (\theta_1, \ldots, \theta_M)^T$ da distribuição de onde provêm os dados através da maximização da função de verosimilhança, a qual se pode determinar a partir dos dados de treino.

Este método tem uma larga aplicação em problemas de classificação de padrões nos quais se consideraram as distribuições de probabilidade condicionadas às classes, $p(\mathbb{X}|C_k)$, em que $\mathbb{X}$ é um conjunto de m vectores, que são os exemplos extraídos independentemente na distribuição $p(\mathbb{X}|\theta, C_k)$. Então, a densidade de probabilidade conjunta de todos os exemplos é dada por:

$$p(\mathbb{X}|\theta, C_k) = \prod_{i=1}^{m} p(\mathbf{x}_i|\theta, C_k) \equiv L(\theta, C_k) \tag{4.15}$$

A função $L(\theta, C_k)$ pode ser vista como uma função do parâmetro θ em que $\mathbb{X}$ e C_k são fixos, neste caso chamando-se *verosimilhança* de θ dada por $\mathbb{X}$.

O método da máxima verosimilhança determina um valor para θ que maximiza $L(\theta, C_k)$. Isto corresponde à ideia intuitiva racional de escolher o valor de θ que tenha a maior probabilidade de originar os exemplos.

Na inferência (aprendizagem) bayesiana a incerteza que se tem acerca dos valores dos parâmetros é representada pela função densidade de probabilidade. Antes de se observarem os exemplos os parâmetros são descritos por uma densidade de probabilidade *a priori*, a qual é tipicamente muito ampla visto reflectir o facto de se ter uma ideia muito pobre sobre os valores que devem ter os parâmetros. Depois de se conhecerem os exemplos pode, então, usar-se o teorema de Bayes para encontrar a densidade de probabilidade *a posteriori* correspondente. Uma vez que certos valores dos parâmetros poderão ser mais consistentes com os dados do que outros verifica-se que a distribuição *à posteriori* é mais apertada que a distribuição *à priori*, fenómeno conhecido por *aprendizagem Bayesiana.*

Começa-se por escrever a função densidade de probabilidade desejada para o vector $\mathbf{x}_i$ dado o conjunto de treino $\mathbb{X}$ como o integral de uma distribuição composta

$$p(\mathbf{x}_i|\mathbb{X}, C_k) = \int p(\mathbf{x}_i, \theta \,|\, \mathbb{X}, C_k)\, d\theta$$

Recorrendo agora à definição de probabilidade condicional:

$$p\left(\mathbf{x}_i, \theta \,|\, \mathbb{X}, C_k\right) = p\left(\mathbf{x}_i \,|\, \theta, \mathbb{X}, \mathbb{C}_{\neg}\right) \cdot p\left(\theta | \mathbb{X}, C_k\right)$$

O primeiro factor é independente de $\mathbb{X}$,uma vez que é justamente a forma que se assumiu para a densidade parametrizada, a qual fica completamente especificada logo que os valores dos parâmetros θ ficarem estabelecidos, logo, obtém-se:

$$p\left(\mathbf{x}_i | \, \mathbb{X}, C_k\right) = \int p\left(\mathbf{x}_i | \, \theta, C_k\right) \cdot p\left(\theta | \, \mathbb{X}, C_k\right) \, d\theta$$

Portanto, em lugar de se escolher para θ um valor específico faz-se na abordagem Bayesiana uma média pesada de todos os valores de θ. O factor de peso $p\left(\mathbf{x}_i | \theta, C_k\right)$ é a distribuição *a posteriori* de θ que se determina considerando uma distribuição inicial assumida como *a priori,* e em seguida para se ter em consideração a influência dos exemplos, procede-se a uma actualização desta distribuição inicial usando o teorema de Bayes. Uma vez que se assume que os exemplos foram obtidos independentemente da mesma distribuição subjacente pode usar-se a expressão (4.15) e pelo teorema de Bayes pode escrever-se a distribuição à posteriori para θ:

$$p\left(\theta | \mathbb{X}, C_k\right) = \frac{p\left(\mathbb{X} | \theta, C_k\right) \cdot p\left(\theta | C_k\right)}{p\left(\mathbb{X} | C_k\right)} = \frac{p\left(\theta | C_k\right)}{p\left(\mathbb{X} | C_k\right)} \prod_{i=1}^{m} p\left(\mathbf{x}_i | \theta, C_k\right)$$

em que o factor de normalização do denominador é dado por

$$p\left(\mathbb{X} | C_k\right) = \int p\left(\zeta | C_k\right) \prod_{i=1}^{m} p\left(\mathbf{x}_i | \zeta, C_k\right) d\zeta$$

o qual assegura que $\int p\left(\zeta | \mathbf{X}, C_k\right) d\zeta = 1$.

4.4.2 Aprendizagem não paramétrica

Um dos problema da aprendizagem por aproximação paramétrica dos modelos de regressão reside no requisito de se conhecer a forma da distribuição de probabilidades subjacente aos exemplos e condicionada às classes às quais eles pertencem. Na prática não é possível, ou não é fácil, determinar-se a forma específica das distribuições condicionais com respeito às classes. Subsiste a dúvida sobre se estas serão de facto gaussianas, como habitualmente se supõe, ou se serão de outras formas mais ou menos parametrizáveis.

Por exemplo, considerando-se três classes de peso fetal, baixo, C_1 tal que $PFE \leq 3000\,\mathrm{g}$, médio, C_2 tal que $3000\,\mathrm{g} < PFE < 3500\,\mathrm{g}$ ou alto, C_3 tal que $PFE \geq 3500\,\mathrm{g}$. Os histogramas do perímetro abdominal, PAB para cada uma das classes de pesos C_1, C_2, C_3, representados na figura 4.36, não são muito sugestivos de uma suposição

de distribuição gaussiana. Um outro aspecto também importante, e que se mostra claramente nesta figura (4.36) tem a ver com a apreciável sobreposição dos polígonos que contornam superiormente os três histogramas do PAB, condicionados à classe C_1, à classe C_2 e à classe C_3.

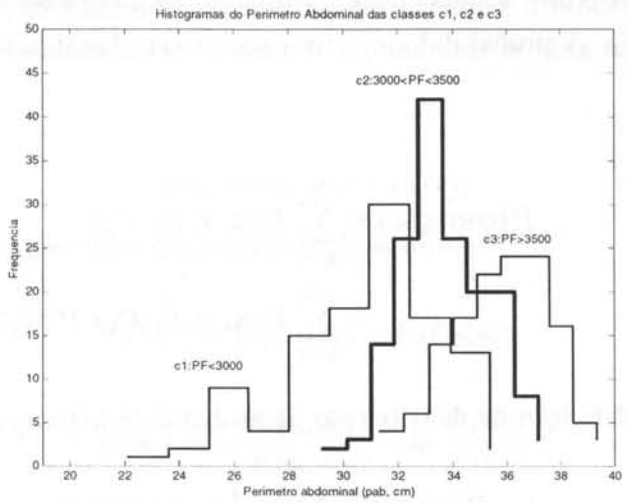


Figura 4.36: Histogramas do perímetro abdominal para 3 classes de PFE

4.4.3 Fronteira de decisão

A probabilidade *a posteriori* $p(C_k|\mathbf{x}_i)$ dá-nos a probabilidade do exemplo $\mathbf{x}_i$ pertencer à classe C_k condicionada a que esse exemplo tenha sido observado. A probabilidade de se errar numa dada classificação pode ser minimizada seleccionando a classe C_k cujo valor da probabilidade *à posteriori* seja mais elevado. Sendo assim, atribui-se o exemplo $\mathbf{x}_i$ à classe C_k se

$$p(C_k|\mathbf{x}_i) > p(C_j|\mathbf{x}_i), \ \forall j \neq k.$$

Utilizando a fórmula de Bayes e desprezando o seu denominador, uma vez que a densidade não condicional $p(\mathbf{x}_i)$ é independente da classe, pode definir-se o critério de comparação das probabilidades *a posteriori* :

$$p(\mathbf{x}_i|C_k) \cdot p(C_k) > p(\mathbf{x}_i|C_j) \cdot p(C_j), \forall j \neq k \tag{4.16}$$

Um classificador dispõe sempre de uma regra de atribuição de cada ponto do espaço das variáveis de entrada $n-$dimensional a uma dada classe $C_1, C_2, \ldots, C_c$. Pode considerar-se que neste espaço existe uma partição em regiões de decisão $R_1, R_2, \ldots, R_c$,uma para cada classe, as quais podem ser contíguas, ou disjuntas.

As hiper-superfícies que delimitam estas regiões são conhecidas por fronteiras ou *superfícies de decisão.*

Existem critérios óptimos para a colocação das fronteiras de decisão de forma a minimizar a probabilidade do acontecimento a classificação estar errada, isto é., seja atribuído a um vector a uma classe C_k quando de facto esse vector pertence a uma outra classe C_j. A probabilidade de um vector ser classificado correctamente pode exprimir-se:

$$\begin{aligned} P\left(\text{correção}\right) &= \sum_{k=1}^{c} P\left(\mathbf{x}_i \in R_k, C_k\right) \\ &= \sum_{k=1}^{c} P\left(\mathbf{x}_i \in R_k | C_k\right) P\left(C_k\right) \end{aligned}$$

Usando a definição de distribuição de probabilidade (cumulativa):

$$P\left(\mathbf{x}_i \in R_k | C_k\right) = \int_{R_k} p\left(\mathbf{x}_i | C_k\right) d\mathbf{x}$$

Logo,

$$P\left(\text{correção}\right) = \sum_{k=1}^{c} \int_{R_k} p\left(\mathbf{x}_i | C_k\right) P\left(C_k\right) d\mathbf{x}$$

Esta probabilidade é maximizada pela escolha da região R_k tal que a cada $\mathbf{x}_i$ seja atribuída a classe para a qual o valor do integral é máximo, o qual é equivalente à expressão (4.16).

Pode reformular-se o processo de classificação em termos de um conjunto de funções discriminantes:

$$y_k\left(\mathbf{x}_i\right) = P\left(C_k | \mathbf{x}_i\right), k = 1, \ldots, c$$

tal que ao vector $\mathbf{x}_i$ seja atribuído uma classe C_k se:

$$y_k\left(\mathbf{x}_i\right) > y_j\left(\mathbf{x}_i\right), \forall j \neq k.$$

Usando o teorema de Bayes e aplicando logaritmos obtém-se:

$$y_k\left(\mathbf{x}_i\right) = \ln p\left(\mathbf{x}_i | C_k\right) + \ln P\left(C_k\right), k = 1, \ldots, c$$

De um modo geral as fronteiras de decisão são dadas pelo conjunto de pontos nos quais se verifica a igualdade entre duas funções discriminantes, de tal forma que se as regiões R_k e R_j são contíguas, então, a fronteira de decisão que as separa é

dada por:

$$y_k(\mathbf{x}_i) = y_j(\mathbf{x}_i), \forall k \neq j$$

A localização das fronteiras de decisão não é afectada por transformações monotónicas das funções discriminantes como, por exemplo, a função logarítmica.

No caso de a classificação ser dicotómica pode usar-se uma única função discriminante:

$$y(\mathbf{x}_i) = y_1(\mathbf{x}_i) - y_2(\mathbf{x}_i)$$

escrita em qualquer uma das seguintes duas formas:

$$y(\mathbf{x}_i) = P(C_1|\mathbf{x}_i) - P(C_2|\mathbf{x}_i)$$

$$y(\mathbf{x}_i) = \ln \frac{P(\mathbf{x}_i|C_1)}{P(\mathbf{x}_i|C_2)} - \ln \frac{P(C_1)}{P(C_2)}$$

Pode parecer que as funções discriminantes não introduzem nenhuma novidade em relação à regra de Bayes mas, tal como se viu no estudo da teoria estatística da aprendizagem aplicada ao modelos de MVS, FBR e PMC, é possível determinar as funções discriminantes com base nos dados de treino sem ter primeiro que determinar a função densidade de probabilidade e, ao mesmo tempo, cumprir os critérios óptimos da teoria da decisão acima referidos.

A título de exemplo pode ver-se que o diagrama de dispersão representado na figura 4.37 mostra a pobreza da separação entre as duas classes extremas de pesos fetais, a classe C_1 de pesos baixos, e a classe C_3 de pesos altos, obtida na representação de um conjunto de exemplos no espaço bidimensional das variáveis PAB e CFE.

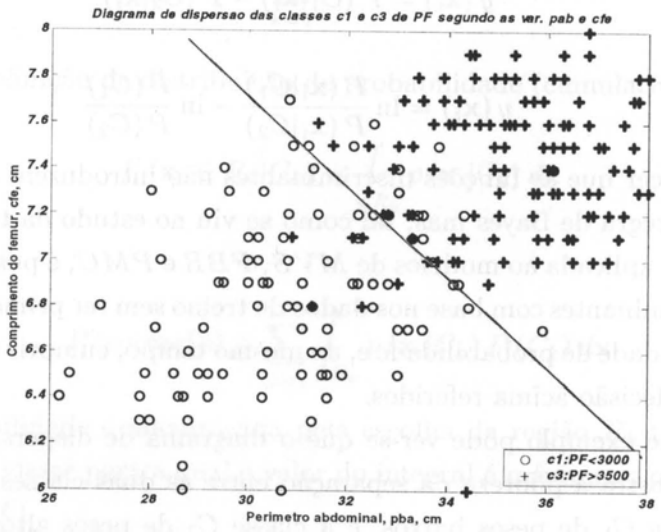


Figura 4.37: Digramas de dispersão CFE vs. PAB para 2 classes de PFE

4.5 Sumário

Procedeu-se a uma análise exploratória de dados focada na estrutura do conjunto de indivíduos observados sem procurar necessariamente deduzir leis válidas para a população de onde estes eram provenientes.

Procurou-se ilustrar aspectos que permitem uma compreensão mínima de conceitos básicos de biometria fetal, obtidos pelo método ecográfico, de forma a permitir encontrar soluções para uma compensação dos erros aleatórios que as medidas das variáveis explicativas por vezes possam enfermar.

Fez-se uma revisão de trabalhos pioneiros relevantes sobre a previsão do peso fetal dando ênfase aos modelos lineares generalizados, que se baseiam em estudos rigorosos elaborados ao longo de décadas por especialistas da mais elevada competência. Esta revisão tem uma importância fundamental para o capítulo seguinte, uma vez que algumas fórmulas e resultados existentes podem transmitir conhecimento *a priori* a ser adequadamente " reciclado" e aproveitado nas experiências com modelos de rede neuronal e de máquina de vector de suporte.

Os principais objectivos de investigação que se procuraram tratar neste capítulo foram: (*i*) Experimentar modelos de regressão multilinear com vista à concepção de modelos de rede neuronal para o processamento das medidas ecográficas, que permitam a obtenção de estimativas do peso fetal mais precisas do que as dadas pelos métodos tradicionais; (*ii*) Revisão teórica sobre modelos de regressão e linear máxima verosimilhança e estimação paramétrica bayesiana; (*iii*) Enquadramento do estudo e das experiências na teoria da probabilidade e inferência estatística, aplicadas à modelização biométrica de dados observacionais.

4.6 Referências bibliográficas e notas históricas

As potencialidades, coerência e elegância da teoria bayesiana aplicada ao reconhecimento de padrões faz desta, na opinião de Duda et al. (2001), um dos mais belos formalismos da ciência. Tendo sido Bayes (1702-1761) que estabeleceu o teorema para o caso de os apriorismos serem uniformes, foi posteriormente redescoberta por Laplace (1749-1827) que a apresentou no caso discreto e mais geral.

A primeira utilização de teoria da decisão bayesiana ocorreu em 1957, devida a Chow, que posteriormente estabeleceu relações fundamentais entre o erro e a taxa de rejeição.

A estimação por máxima verosimilhança e bayesiana têm uma história longa no reconhecimento de padrões. A aproximação bayesiana começou com uma sugestão de que a maneira apropriada para se utilizarem os exemplos quando se desconhecem as densidades condicionais é o cálculo de $P(c_i|\mathbf{x}, \mathcal{D})$ tal como se descreve emBishop

(1995). O próprio Bayes apreciou o papel de apriorismos não informativos. Uma análise dos diferentes apriorismos em estatística pode ser encontrada em Jeffreys (1939).

A análise dos componentes principais é uma técnica estatística clássica, com uma extensa aplicação em engenharia, e que está desenvolvida na obra de Ian T. Jolliffe (1986).

O trabalho de Fisher sobre discriminantes lineares foi pela primeira vez publicado em 1936 e encontra-se muito bem explicado na obra de G. J. McLachlan (1992).

A ideia da análise de regressão surgiu pela primeira vez ao biólogo Francis Galton (1822-1911) da Universidade de Cambridge, quando estudava uma possível associação entre um conjunto de medidas das alturas de crianças e dos respectivos pais. Galton representou pela primeira vez graficamente a dispersão destas duas variáveis e observou, nas sua próprias palavras: "*that lines drawn through entries of the same value formed a series of concentric and similar ellipses* (1885)" (Spanos (1999),p.293). A sua primeira reacção foi utilizar estas elipses concêntricas para construir superfície em duas dimensões, da frequência conjunta que lhes está subjacente. Reconhecendo que as suas competências matemáticas não eram suficientes entregou esta tarefa ao seu amigo e matemático Dickson e, em 1986, apareceu publicado por Galton, num apêndice, a expressão da distribuição normal bidimensional.

Uma interpretação geométrica do método dos mínimos quadrados pode ser encontrada em Murteira et al. (2001) (p.519) e Saporta (1990).

Em Pestana and Velosa (2002) (pp.149-160) pode encontrar-se um estudo interessante da recta de regressão, ou dos mínimos quadrados, para modelar a dependência entre o comprimento do biparietal (DBP) e o perímetro cefálico (PCE). Esta referência e este estudo, lamentavelmente, eram desconhecidos durante a redacção desta dissertação.

Capítulo 5

Regressão por Rede Neuronal e Máquina de Vectores de Suporte

Introdução

Este capítulo tem como principal objectivo a apresentação de resultados de experiências que visam a obtenção de um modelo estimador do peso fetal baseado na rede neuronal de perceptrão multicamada(PMC), funções de base radial (FBR) e máquina de vectores de suporte (MVS). A pesquisa de métodos para se obter um modelo estimador de uma função real multivariável, representativa de um conjunto de vectores, baseada em dados experimentais é considerada tradicionalmente como um problema central em muitas aplicações no domínio da estatística. A principal técnica para a resolução deste problema de estimação ainda continua a ser o método dos mínimos quadrados, o qual tendo sido sugerido por Legendre (1752-1833) e Gauss (1777-1855), beneficiou de inúmeros enriquecimentos e importantes desenvolvimentos analíticos no século XX. Foi este método que se aplicou na análise exploratória de dados apresentada no capítulo anterior.

No presente capítulo consideram-se métodos de estimação de funções que recorrem às redes neuronais e máquinas de vectores de suporte que surgiram nos anos 1980 e 1990, e que foram referidas extensivamente nos capítulos segundo e terceiro. Um estudo comparativo entre os erros de previsão do peso fetal que se obtêm com os métodos clássicos, paramétricos, e os eros que se obtêm com estes métodos mais recentes, semi-paramétricos, de rede neuronal PMC, FBR e máquina MVS, está na génese do desenvolvimento das ideias que mais influenciam a abordagem do problema da estimação do peso fetal. Estudos comparativos desta natureza precederam e prepararam as condições que permitiram uma selecção de tópicos que se apresentaram nos capítulos dois e três, nos quais se trataram aspectos da *Teoria Estatística da Aprendizagem*, assim como no capítulo seis, em que se irá tratar aspectos dos

modelos baseados na *Teoria da Informação.*

Na secção 5.1 referem-se as experiências, a metodologia e os resultados obtidos com modelos de perceptrão multicamada (PMC). Apresenta-se uma descrição detalhada do problema da previsão do peso fetal com vários modelos de rede neuronal dos tipos PMC e FBR, assim como as condições e resultados de experiências divulgadas em Sereno et al. (1999) e Sereno et al. (2000b) .

Na secção 5.2 referem-se experiências e resultados obtidos com modelos de redes de funções de base radial (FBR) e que foram divulgados em Sereno et al. (2000b) Sereno et al. (2000a). Estas duas primeiras secções contêm também algumas análises teorizantes e exploratórias das metodologias usadas na concepção e avaliação das redes. Referem-se sobretudo soluções inovadoras e resultados relevantes para os objectivos desta tese .

Na secção 5.3 aborda-se empiricamente a máquina de vectores de suporte (MVS) na estimação quer de funções indicatrizes (classificação dicotómica) Sereno et al. (2000a), quer de funções reais (regressão) estimadoras do peso fetal Sereno et al. (2001b). Os modelos baseados na MVS têm tido um assinalável sucesso em aplicações que envolvem não só problemas gerais de classificação Schölkopf (1997) e regressão Smola and Schölkopf (1998), mas também em problemas específicos de biomedecina e biologia Odone et al. (1998), Veropoulos et al. (1999), e muitos outros tipos de problemas, nomeadamente quando se requer a estimação de funções de distribuição. Os algoritmos das MVS baseiam-se, como se viu no capítulo três, na teoria estatística das amostras finitas, que estipula que a quantidade crítica dos algoritmos destas máquinas, em lugar de ser o número dos seus parâmetros a serem estimados, passa a ser uma medida de capacidade, a chamada dimensão VC da classe de funções impletementadas pela máquina para a geração de estimativas com exemplos novos.

Na secção 5.4 faz-se uma comparação entre os resultados obtidos com PMC, FBR e MVS, usando nos três modelos um mesmo conjunto de treino e de teste (este, obviamente, disjunto do de treino). A implementação do algoritmo de aprendizagem MVS usada neste estudo foi o *algoritmo da decomposição* ($SVMTorch$).

Na secção 5.5 faz-se uma análise numérica do algoritmo e aprendizagem das MVS, segundo três aspectos, aproximação, estimação e problemas de álgebra linear numérica.

Na secção 5.6 estuda-se empiricamente alguns limites da capacidade de generalização das MVS, no que diz respeito ao número de vectores de suporte, escolha óptima dos parâmetros C e ε, selecção de modelos, e estabelece-se uma distinção entre método de kernel, específico das MVS, e método de dicionário, específico das redes neuronais de PMC e FBR.

5.1 Rede Neuronal de Perceptrão de Multicamada

Efectuaram-se experiências com vários modelos de rede neuronal de perceptrão de multicamada (PMC), os quais como se sabe são antero-alimentadas ('*feedforward*'), utilizando-se unidades de activação de função *tan-sigmóidal* nas camadas escondidas e função *linear* pura na camada de saída. Utilizou-se como função de custo a ser minimizada o erro médio quadrático entre as saídas da rede e os valores alvo.

Estes modelos de regressão foram criados utilizando funções pertencentes ao pacote de ferramentas *Neural Network Toolbox*, incorporado no $MATLAB$ (ver. 5 e 6) como, por exemplo, a que se designa "$newff$" ('*new feedforward*') que cria uma arquitectura para uma nova rede e a que se designa "mse" que implementa a função de custo. O tipo de treino foi sempre em modo de pacotes (*batch*). Os métodos de treino clássicos baseiam-se na descida de gradiente, com ou sem *momentum,* usada no algoritmo da rectropropagação, porém, este torna-se por vezes muito lento. As ferramentas do $MATLAB$ recorrem a outros métodos com técnicas de aceleração da adaptação dos pesos que permitem aumentar a rapidez para cerca de dez a cem vezes em relação aos métodos clássicos. Uma primeira categoria de algoritmos rápidos emprega técnicas heurísticas que foram desenvolvidas a partir da análise da performance do algoritmo estandardizado da descida mais íngreme, por exemplo, com variação da taxa de aprendizagem [Marques_de_Sá (2001a), Nabney (2002).] Uma segunda categoria utiliza técnicas estandardizadas de optimização como, por exemplo, o algoritmo dos gradientes conjugados, o algoritmo de *quasi-Newton* e o algoritmo de *Levenberg-Marquardt.*

O método de optimização que se adoptou foi o algoritmo de Levenberg-Marquardt [Marquardt (1963)], cuja descrição aparece em muitos textos, como por exemplo, em Bishop (1995), Hagan et al. (1996), Haykin (1999), Marques_de_Sá (2001a). Referem-se aqui apenas alguns detalhes com importância prática para as aplicações. Este algoritmo tem semelhanças com o algoritmo de quasi-Newton, porém, está concebido para ter uma velocidade de treino típica de uma algoritmo de segunda ordem sem ter o trabalho de calcular a matriz Hessiana. Quando a função de custo tem a forma de uma soma de quadrados, então, a matriz Hessiana $\mathbf{H}$ pode ser aproximada de uma forma mais simples usando a matriz jacobiana $\mathbf{J}$,que contém as primeiras derivadas dos erros da rede em relação aos pesos e polarizações (*bias*):

$$\mathbf{H} = \langle \mathbf{J}, \mathbf{J} \rangle$$

E, designando por $\mathbf{e}$ o vector de erros da rede, o gradiente $\mathbf{g}$ pode calcular-se:

$$\mathbf{g} = \langle \mathbf{J}, \mathbf{e} \rangle$$

A função do MATLAB '*trainml*' determina a matriz jacobiana **J** por rectropropagação, Hagan and Menhaj (1994), obtendo a matriz Hessiana através da seguinte fórmula de recorrência do tipo método de Newton:

$$\mathbf{w}_n = \mathbf{w}_{n-1} - [\langle \mathbf{J}, \mathbf{J} \rangle + \lambda \mathbf{I}]^{-1} \langle \mathbf{J}, \mathbf{e} \rangle .$$

Se o escalar λ for nulo esta fórmula coincide com a do método de Newton, utilizando a aproximação pela matriz Hessiana. Se λ tiver um valor grande a fórmula transforma-se no método da descida de gradiente com um passo pequeno. Como o método de Newton é mais rápido e mais preciso perto de um mínimo da função de custo, então o objectivo é mudar para este método logo que seja possível. Assim, o valor de λ é decrementado em cada passo em que se verifica de facto uma redução da função de custo, garantindo assim que este diminui sempre em todas as iterações do algoritmo.

5.1.1 Resultados obtidos com um conjunto de treino de 140 casos

Apresenta-se a seguir os resultados reportados em Sereno et al. (1999). As variáveis de entrada utilizadas eram as medidas ecográficas do diâmetro biparietal (DBP), perímetro cefálico (PCE), perímetro abdominal (PAB), comprimento do fémur (CFE) e índice de resistência umbilical (URI). Todas estas características ecográficas tinham sido medidas com a mesma metodologia baseada no protocolo que tinha sido estabelecido. A variável de saída era o peso fetal do recém-nascido. Todas as entradas e a saída eram pré-processadas utilizando-se os respectivos valores normalizados ($p_{norm} = (p - mean\,(p))/stdp\,(p)$).

As redes neuronais eram constituídas por: (i) uma camada com as 5 entradas; (ii) uma primeira camada escondida com um número variável de unidades (entre 2 e 5) e funções de activação tan-sigmóidais; (iii) uma segunda camada escondida com um número variável de unidades (entre 2 e 3) e funções de activação tan-sigmóidais; (iv) uma unidade de saída usando uma função de activação linear simétrica. Estas redes foram treinadas com um conjunto de treino de 140 casos, sem paragem antecipada, ao longo de 10 épocas usando a função de custo '*mse*' e o algoritmo de Levenberg-Marquardt com um valor de $\lambda = 0.0010$ e os restantes parâmetros $\lambda_{inc} = 10$, $\lambda_{dec} = 0.1$, $\lambda_{\max} = 10^{10}$. A inicialização dos pesos utilizava a função '*initnw*', a qual, nas redes '*feedforward*', consiste numa atribuição aos pesos de valores aleatórios compreendidos entre -1 e 1.

Seguidamente, procedeu-se a uma simulação de cada uma das redes utilizando os valores para os pesos que se obtiveram no fim do treino de cada uma delas e, utilizando um conjunto disjunto com 140 casos, testaram-se as performances de previsão do peso fetal de cada modelo de rede neuronal, tendo-se obtido os valores para os erros relativos das previsões que se indicam na tabela 5.1.

Modelo	N_{unid}	N_{param}	Erro relat. (%)
PMC **5:5:3:1**	9	52	**9.07**
PMC **5:4:2:1**	7	37	**8.08**
PMC **5:3:2:1**	6	29	**8.73**
PMC **5:2:2:1**	5	21	**8.77**
PMC **5:5:1**	6	36	**8.42**
RLG Hadlock		4	**9.23**
RLG Shepard		4	**9.30**

Tabela 5.1: Resultados de experiências com PMC treinados com140 casos e testadas com 140 casos (disjuntos)

Supondo que N_{entr} é o número de entradas da rede, E_1 é o número de unidades na primeira camada escondida, E_2 é o número de unidades na segunda camada escondida, S é o número de unidades na camada de saída, sabendo que o número de polarizações (bias) é igual ao número de unidades total da rede N_{unid}, o número de parâmetros, ou pesos, N_{param}que a rede tinha que adaptar pode-se calcular pela fórmula:

$$N_{param} = N_{entr} * E_1 + E_1 * E_2 + E_2 * S + N_{unid}$$

Em conformidade com as fórmula sugeridas por Lippmann (1987), e Mirchandi and Cao (1989), o número máximo de unidades da camada escondida não deveria exceder $\log_2(140) \simeq 7$. Os erros relativos obtidos são cerca de 1% inferiores aos obtidos pelos modelos clássicos de Hadlock et al. (1985) e Shepard (1982). Estes resultados, na altura em que foram obtidos (em 1999-06-16), foram interpretados como sendo insuficientes, e que tal se podia dever ao facto de o número de casos de treino ser muito reduzido. Tornava-se, assim, difícil a determinação quer de um hiperplano de decisão capaz de separar correctamente os casos em classes de pesos, quer de uma função capaz de aproximar os casos de treino com uma boa capacidade de generalização.

Os histogramas das frequências de perímetros abdominais PAB para três classes de peso fetal PFE, representados na Figura 4.36, mostram a existência de sobreposições significativas. Nesta situação não se afigura muito fácil obter-se uma função que permita aproximar os dados com uma boa capacidade de generalização.

O diagrama de dispersão do comprimento do fémur CFE *versus* o perímetro abdominal PAB, que se mostra na Figura 4.37 para duas classes de pesos fetais distintos (extremos), mostra igualmente a existência de uma mistura significativa das classes. Como se sabe a variável PAB apresenta o mais elevado coeficiente de correlação com o peso. Outros diagramas de dispersão para pares de outras variáveis ecográficas não mostravam uma situação melhor. A utilização da variável de fluxo sanguíneo, obtida por ecografia em modo Doppler, o chamado índice de

resistência umbilical, *IRU*, o qual se pode ver na matriz de correlação 4.6 que tem nesta colheita um baixo valor absoluto do seu coeficiente de correlação linear com o peso fetal ($-.40$), poderia também ter tido uma contribuição no treino dos modelos desfavorável à aquisição de uma boa capacidade de generalização destes.

Modelos simples de *PMC*, como por exemplo o melhor que se obteve nestas experiências (Tabela 5.2), deveriam, em princípio, permitir estabelecer qualquer relação funcional entre os vectores de entrada e as etiquetas. Uma redução na dimensionalidade dos vectores de entrada iria conduzir a uma perda de informação pelo que se advogava a necessidade de utilizar um número de casos de treino substancialmente maior. .

Melhor Modelo	$\mathbf{N_{unid}}$	$\mathbf{N_{param}}$	**Média dos Erros (%)**
PMC **5:4:2:1**	7	37	**8.08**

Tabela 5.2: Melhor rede PMC-5-4-2-1 (140 casos treino, 140 casos teste)

Adicionalmente faziam-se estudos da redução do erro relativo através da incorporação quer de melhorias no pré-processamento, particularmente, um re-escalonamento das variáveis de entrada proporcionalmente ao seu grau de importância, quer de conhecimento a priori. Este último iria influenciar o pré-processamento tornando-o mais eficaz na filtragem de casos contraditórios ou, eventualmente poderia sugerir a utilização de uma estrutura de rede mais apropriada ao problema.

Vem a propósito apresentar-se ainda um esclarecimento relevante acerca do número de épocas de treino de facto usadas nas experiências relatadas nesta subsecção. Este número não deveria ter sido muito superior a 10 épocas, tal como se referiu acima, pelo que importa esclarecer que deve ter sido um lapso a indicação do valor de 1000 épocas em Sereno et al. (1999).

Como conclusão final assinala-se que os resultados obtidos eram positivos, uma vez que as redes neuronais pareciam ser capazes de fazer a previsão do peso fetal com mais precisão comparativamente com as fórmulas que os obstetras estavam a utilizar na prática clínica. Este facto justificou o esforço quer da recolha dos dados efectuada pelos médicos, quer do desenvolvimento dos algoritmos de aprendizagem pela equipa do INEB motivando assim optimismo no prosseguimento da investigação.

Seguidamente, relata-se um novo conjunto de experiências realizadas com um conjunto de treino com mais 73 casos, introduzia-se uma técnica de regularização do treino e tirava-se mais partido do conjunto de dados usando o método de avaliação da performance por validação cruzada com $k-$conjuntos (*k-fold cross validation*).

5.1.2 Resultados obtidos com um conjunto de treino de 213 casos

Apresenta-se a seguir os resultados reportados em Sereno et al. (2000a) sobre experiências realizadas com modelos de redes treinadas para uma previsão do peso fetal utilizando como variáveis de entrada as medidas ecográficas diâmetro biparietal (DBP), perímetro cefálico (PCE), perímetro abdominal (PAB), comprimento do fémur (CFE) e índice de resistência umbilical (URI). As medidas destas características ecográficas obedeciam à mesma metodologia de colheita de dados, nos quatro centros hospitalares, que constava no protocolo que se refere na subsecção 4.1.1.

As variáveis de saída, ou etiquetas, eram o peso e o comprimento fetal, isto é, do recém-nascido uma vez que estas características tinham sido medidas após o nascimento. As entradas e saídas foram todas pré-processadas obtendo-se os respectivos valores normalizados ($p_{norm} = (p - mean\,(p))/stdp\,(p)$).

A arquitectura das redes neuronais utlizada que se designará por *PMC-5:5:2*, compreendia: (i) uma camada com as 5 entradas; (ii) uma camada escondida com 5 unidades cujas funções de activação eram tan-sigmóidais; (iii) uma camada de saída com 2 unidades, uma para a saída dos valores estimados dos pesos e a outra para os comprimentos, ambas usando uma função de activação linear pura.

O conjunto de dados utilizados nesta experiência compreendia 354 casos que tinham sido recolhidos até à data da publicação dos resultados das experiências (2000-03-17). Considerou-se uma partição deste conjunto em cinco subconjuntos, $\{\mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_4\}$, contendo cada um 71 casos, e um conjunto $\mathbb{Z}_5$ contendo 70 casos. A técnica usada nesta partição é relevante. Para garantir que os conjuntos fossem aleatórios e originários de uma distribuição o mais similar possível, procedeu-se a uma ordenação de todos os casos por ordem crescente do peso fetal. Seguidamente, formaram-se cinco sucessões aritméticas de índices $\{i_n\}$, em que o primeiro elemento de cada sucessão era 1, 2, 3, 4, ou 5 e os restantes, de cada sucessão, eram obtidos pela fórmula de recorrência, $i_n = i_{n-1} + 5$. Cada subconjunto era formado pelos vectores do conjunto de 354 casos originais cujos índices de cada caso eram os elemento de cada uma das sucessões aritméticas que se referiu.

O método de treino usava a função de custo e o algoritmo de treino do MATLAB, '*mse*' e '*trainlm*', respectivamente. O algoritmo de treino Levenberg-Marquardt, '*trainlm*', requer dois argumentos que devem ser conjuntos disjuntos do de treino, '*v.tr*' e '*t.tr*', designados por validação e teste, respectivamente, para fazer uma regularização por paragem antecipada (*early stopping*).

As redes neuronais foram treinadas com um conjunto de treino de 213 casos, com paragem antecipada, ao longo de $10 - 20$ épocas usando a função de custo '*mse*', o algoritmo de Levenberg-Marquardt com um valor de $\lambda = 0.0010$ e os restantes

parâmetros $\lambda_{inc} = 10$, $\lambda_{dec} = 0.1$, $\lambda_{\max} = 10^{10}$. A inicialização dos pesos utilizava a função '*initnw*', a qual nestas redes '*feedforward*' consiste numa inicialização dos pesos com valores aleatórios compreendidos entre -1 e 1.

A tabela 5.3 mostra os subconjuntos utilizados para treino, validação (coluna *v.tr*) e teste (coluna *t.tr*), o número de épocas de treino (coluna Ne), o erro mínimo quadrado em unidades estandardizadas (coluna MSE), o conjunto usado para teste (coluna *tst*), os erros relativos da previsão do peso fetal cometidos pela rede (coluna E_{tst}), pela fórmula de Hadlock (coluna E_{had}) e pela fórmula de Shepard (coluna E_{she}).

	Conj. Treino 213 casos	**v.tr** 70c.	**t.tr** 71c.	**Ne**	**MSE**	**tst** 71c.	**E_{tst}** %	**E_{had}** %	**E_{she}** %
1	$\{\mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_3\}$	$\mathbb{Z}_5$	$\mathbb{Z}_4$	16	0.265	$\mathbb{Z}_4$	7.41	7.56	8.36
2	$\{\mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_4\}$	$\mathbb{Z}_5$	$\mathbb{Z}_1$	11	0.294	$\mathbb{Z}_1$	7.05	7.76	7.86
3	$\{\mathbb{Z}_3, \mathbb{Z}_4, \mathbb{Z}_1\}$	$\mathbb{Z}_5$	$\mathbb{Z}_2$	10	0.294	$\mathbb{Z}_2$	7.20	8.36	8.58
4	$\{\mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_2\}$	$\mathbb{Z}_5$	$\mathbb{Z}_3$	10	0.267	$\mathbb{Z}_3$	7.03	8.95	7.83
					Média	=	**7.17**	**8.16**	**8.16**

Tabela 5.3: Resultados PMC-5:5:1 (213 casos treino, 71 casos teste)

Como se pode ver na tabela 5.3 o subconjunto usado no treino para fazer a paragem antecipada (teste '*t.tr*') é o mesmo que se usou para determinar um valor aproximado do erro verdadeiro. No Capítulo 7 procede-se a uma discussão sobre algumas dúvidas, aparentemente sem justificação, que se podem levantar a esta escolha do mesmo conjunto, usado quer no treino quer na avaliação da performance da rede.

A figura 5.1 representa as curvas de evolução dos três tipos de erros amostrais, treino, validação e teste, ao longo de 14 épocas de adaptação dos pesos. A figura 5.2 mostra 71 casos, enquanto que a figura 5.3 mostra apenas 24 casos, dos valores estimados pela rede, os quais se podem comparar com os valores verdadeiros do peso fetal, estes últimos tendo servido como critério de ordenação (crescente) dos casos.

Modelos	**N_{unid}**	**N_{param}**	**Média dos Erros (%)**
PMC **5:5:2**	7	42	**7.17**

Tabela 5.4: Resultados PMC-5:5:2 (213 casos treino, 71 casos teste)

Este conjunto de experiências confirmou os resultados positivos que já tinham sido obtidos. As redes neuronais apresentam erros relativos médios de **7.17%**, os quais são inferiores aos apresentados pelas fórmulas clássicas usadas pelos obstetras, **8.16%** (Tabela 5.4). Estas experiências permitem sugerir que as redes PMC poderão ser soluções sérios à aplicação na prática clínica para efeitos de previsão do peso fetal.

As experiências com PMC não se resumem apenas a estes dois estudos que estão

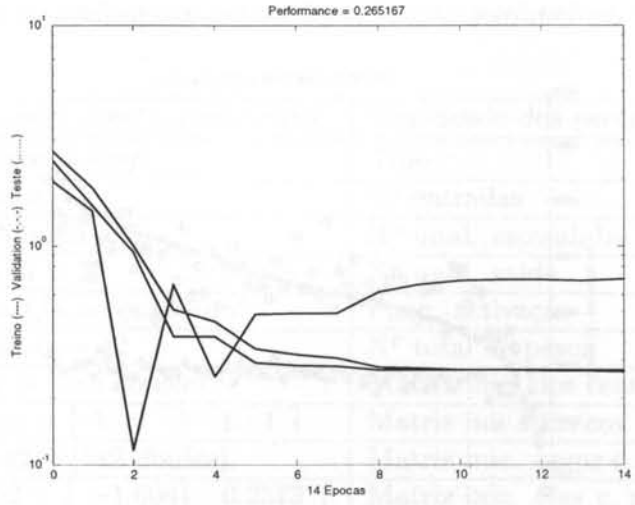


Figura 5.1: Gráficos dos erros amostrais de um *PMC-5:5:2* em função do número de épocas, para os conjuntos de: (1) treino (traço contínuo), (2) validação (traço misto) e (3) teste (traço interrompido).

publicados, prosseguiram, em muitas direcções, mas, agora, na secção seguinte, vai-se apresentar um conjunto de resultados que visam um melhoramento dos resultados da presente secção, e que foram obtidos com um dos mais sofisticados *packages* de algoritmos neuronais de aprendizagem desenvolvido por Chris Bishop e Ian Nabney, na Universidade de Aston, Cambridge, UK.

5.2 Rede Neuronal de Funções de Base Radiais

Os resultados reportados nesta secção foram apresentados em duas comunicações apresentadas em Sereno et al. (2000b) e em Sereno et al. (2000a), em que se referem experiências realizadas com modelos usando o mesmo conjunto de 315 dados utilizado nas experiências que se acaba de relatar, subsecção 5.1.2.

Na presente investigação foram criadas redes de Funções de Base Radiais, *FBR*, utilizando um pacote *Netlab* contendo 150 funções escritas em código *MATLAB*, produzidas por Christopher Bishop e Ian Nabney, disponibilizadas já em 1999 na página da web do *NCRG- Neural Computing research Group*, Cambridge, UK, e, (lamentavelmente) só em 2002 documentadas em livro Nabney (2002).

As funções mais usadas nestas experiências são as funções *'rbf'*, construtora da estrutura das redes, *'rbftrain'*, que faz o treino das redes pelo método EM, e *'rbffwd'*, que faz a simulação das redes e permite obter as previsões de pesos fetais.

A função *'rbf'* constrói a estrutura e inicializa os pesos de uma rede de funções de base radiais, *FBR*, restituindo uma estrutura de dados de rede, representada

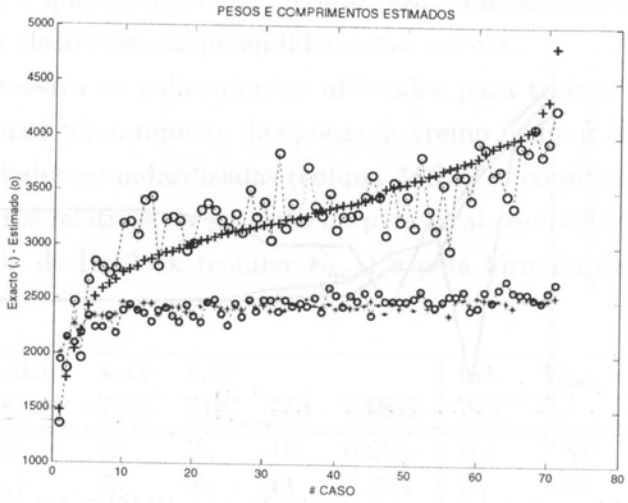


Figura 5.2: Gráfico (*PMC-5:5:2*), para um dos subconjuntos de teste, das previsões do: (a) Peso fetal, círculos na parte superior, (b) Comprimento fetal, na parte inferior. (As cruzes e asteriscos representam valores verdadeiros; em ordenadas apenas a escala dos pesos em gramas)

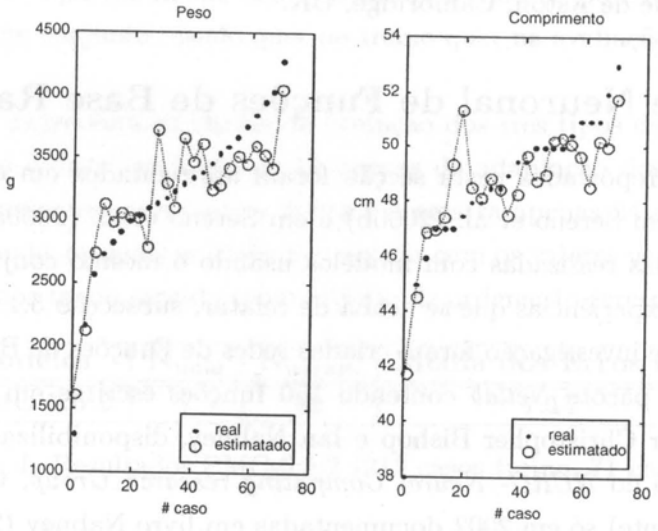


Figura 5.3: Gráfico ($PMC-5:5:2$) para um subconjuntos de casos das estimativas do: (a) Peso fetal (esquerda) (b) Comprimento fetal (direita). (Os círculos a cheio representam valores verdadeiros)

por um prefixo e os parâmetros respectivos, *<net>.parâmetros*, tal como mostra a tabela 5.5.

Estrutura *<net>.parâmetros*	Significado dos *parâmetros*
$net.type = 'rbf'$	Tipo
$net.nin = 5$	N^o entradas
$net.nhiden = 5$	N^o unid. escondidas
$net.out = 2$	N^o unid. saída
$net.actfn = 'gaussian'$	Func. activação
$net.nwts = 42$	N^o total de pesos
$net.c = [5x5\ double]$	Matriz inic. dos centróides
$net.wi = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}$	Matriz inic. das cov
$net.w2 = [5x2\ double]$	Matriz inic. pesos c. esc.
$net.b2 = \begin{bmatrix} -1.6041 & 0.2573 \end{bmatrix}$	Matriz inic. *bias* c. saída

Tabela 5.5: Exemplo de uma estrutura de rede FBR-5:5:2 (Netlab)

Os pesos iniciais são valores aleatórios obtidos duma distribuição normal de média zero e variância unitária e os parâmetros das variâncias iniciais têm valores unitários. As funções de activação utilizadas são gaussianas definidas em termos da distância entre um ponto de dados e o centro deste. Os dados estão normalizados. Assume-se que cada função tem um centro e um parâmetro único de largura, e que os pesos da camada escondida para a camada de saída incluem um *bias*. A largura requerida para a função de activação gaussiana é unitária, as quais sendo definidas em termos da distância entre os exemplos e os centróides são calculadas segundo uma dada escala, não necessitam de ser normalizadas, uma vez que as saídas são linearmente combinadas na camada seguinte.

A função *'rbftrain'* recebe a estrutura *<net>.parâmetros* e adapta os pesos usando um algoritmo de treino em duas fases. Os centros são calculados pelo ajuste de um modelo de mistura de gaussianas com covariâncias circulares usando *o algoritmo EM*. O modelo de mistura é inicializado usando um pequeno número de iterações do algoritmo *k-means*. Como as activações utilizadas são gaussianas as larguras das funções de base são fixadas na distância máxima quadrática inter-centrada. Os pesos da camada escondida que originam a solução segundo os mínimos quadráticos podem determinar-se usando a pseudo-inversa. Este procedimento é muito mais rápido mesmo que possa originar soluções com um erro um pouco superior ao dado por optimizadores não lineares.

O treino requer a criação de um modelo de mistura de gaussianas com uma dada arquitectura. Para isso necessita-se de saber qual é a dimensão do espaço de entradas, o número de centros do modelo de mistura e o tipo de modelo de mistura. Nas experiências aqui descritas usou-se componente gaussianas esféricas, e os parâmetros únicos da variância foram memorizados como vectores. As probabilida-

BIBLIOTECA F.E. PORTO

des a *priori* têm os mesmos valores iniciais, e de soma unitária, e as covariâncias são todas matrizes identidade. Os centros tomam valores iniciais aleatórios dados por uma distribuição também gaussiana de média nula e variância unitária .

Seguidamente, usa-se o conjunto de treino para inicializar os parâmetros do modelo de mistura de gaussianas. Os centros são determinados pelo algoritmo dos *k-means* . Os valores *a priori* são calculados a partir da proporção de exemplos que pertence a cada aglomerado. As matrizes de covariância são calculadas como a amostra da covariância dos pontos associados com os centros correspondentes, isto é, o mais próximo possível destes. Esta iniciação é o ponto de partida para o treino do modelo pelo *algoritmo EM*. Este algoritmo restitui uma nova estrutura da mistura de gaussianas com os novos *a priori*, os centros que passam a ser os da FBR e as covariâncias. Fixa ainda as variâncias de modo a que estas sejam a máxima distância quadrática aos centros, e com estes valores fixa a largura da função de base gaussiana. Calcula usando a função *rbffwd* a matriz de activações. Aplica a esta matriz a pseudo-inversa e obtém os pesos e as polarizações (*bias*) da rede.

A função *rbffwd* recebe a estrutura *<net>.parâmetros* treinada, que contém todos os pesos e o tipo de funções de activação utilizadas, e a matriz do conjunto de teste. Procede em seguida à propagação antero-posterior na rede dos valores de cada entrada dos casos de teste, obtendo-se, assim, á saída o valor estimado dos pesos fetais.

Algoritmo EM

A exposição que se vai fazer seguidamente poderá parecer demasiado teórica para estar inserida numa secção cujo interesse primordial é a apresentação de resultados. Considera-se no entanto que um nível mínimo de compreensão do algoritmo *EM* tem um papel relevante na afinação das redes de *FBR*, e é por esta razão que se inclui a presente descrição, uma boa abordagem completa podendo encontrar-se em McLachlan and Krishnan (1997).

O algoritmo EM significa algoritmo de *expectativa* e *maximização,* foi proposto por Dempster et al. (1976), para permitir estimar os parâmetros de um modelo de mistura de gaussianas, em particular quando existem dados omissos.

Suponha-se que se tem um aglomerado de pontos X,por exemplo as variáveis de medidas ecográficas, *pab* e cfe, classificadas em classes de pesos *baixos* e *altos.* Admita-se que pela sua observação gráfica, num histograma ou diagrama de dispersão, se considera que uma ou mais funções gaussianas podem descrever a função de distribuição destes pontos.

O centro, ou centróide, deste aglomerado de pontos pode obter-se pela média dos m valores das componentes dos exemplos $\bar{\mu} = [\bar{x}_1, \ldots, \bar{x}_2]^T$ em que $\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m}[x_i]_j$. Pode dizer-se que o centróide de um conjunto de pontos contém uma aproximação

BIBLIOTECA F.E. PORTO

de primeira ordem que em termos de um simples vector diz qualquer coisa acerca do conjunto. Para se dizer mais qualquer coisa pode utilizar-se os n^2 momentos de segunda ordem, as médias dos termos $(x_i - \bar{x}_i)(x_j - \bar{x}_j)$, que se podem representar pela matriz das covariâncias, em notação vectorial:

$$\mathbf{V} = \frac{1}{m-1} \sum (\mathbf{x} - \bar{\mu}) .$$

A matriz $\mathbf{V}$ é simétrica, positiva e semi-definida, pelo que pode ser diagonalizada por uma matriz $\mathbf{Q}$ de rotação, que seja matriz ortogonal de determinante unitário e uma matriz diagonal $\mathbf{\Lambda}$, tal que

$$\mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}^{-1} = \mathbf{V}.$$

Os elementos da diagonal da matriz $\mathbf{\Lambda}$ são os valores próprios de $\mathbf{V}$, a imagem da base canónica em $\mathbf{R}^n$ por $\mathbf{Q}$ chama-se base própria para $\mathbf{V}$, e a imagem dos vectores de base chama-se conjunto de vectores próprios. Quando se diz um vector próprio quer dizer-se um espaço unidimensional de vectores próprios constituído por todos os múltiplos de um dado vector próprio, os quais são também vectores próprios. Se os valores próprios forem todos diferentes então $\mathbf{Q}$ é única, pelo que os espaços próprios são todos unidimensionais. Cada valor próprio é não-negativo uma vez que a matiz $\mathbf{V}$ é positiva e não definida. Se os valores próprios forem todos-não nulos, então a matriz $\mathbf{V}$ é não-singular e positiva definida, logo $\mathbf{V}^{-1}$ é diagonalizável pela mesma matriz $\mathbf{Q}$ e tem como matriz diagonal $\mathbf{\Lambda}^{-1}$, cujos elementos são os inversos dos valores próprios correspondentes. Existe ainda a matriz raiz quadrada $\mathbf{V}^{\frac{1}{2}}$, isto é, $\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{\frac{1}{2}} = \mathbf{V}$, a qual pode ser diagonalizada pela mesma matriz $\mathbf{Q}$, em que os termos diagonais são as raizes quadradas dos valores próprios correspondentes.

Tendo-se obtido a partir dos exemplos o centróide $\bar{\mu}$ e a matriz $\mathbf{V}$ e, uma vez que na literatura clássica da estatística se usa o símbolo $\mathbf{\Sigma}$, pode definir-se a função $g(\bar{\mu}, \mathbf{\Sigma}) : \mathbf{R}^n \to \mathbf{R}$, pela expressão:

$$g(\bar{\mu}, \mathbf{\Sigma}) = \frac{1}{\left(\sqrt{2\pi}\right)^n \left(\sqrt{\det(\mathbf{\Sigma})}\right)} e^{-\frac{(\mathbf{x}-\bar{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\bar{\mu})}{2}} \tag{5.1}$$

O interesse pela modelização com gaussianas reside no facto de: (i) Entre todas as escolhas possíveis de $\bar{\mu}$ e $\mathbf{\Sigma}$ pode mostrar-se que os modelos gaussianos são os que dão a verosimilhança máxima para um dado conjunto de exemplos; (ii) Incluir informação sobre os exemplos dada pelos momentos de primeira ordem através de $\bar{\mu}$ e de segunda ordem através de $\mathbf{\Sigma}$, momentos estes que são extremamente fáceis de calcular; (iii) A grande generalidade das aplicações à análise de dados em biologia e medicina dá mais preferência às leis com misturas de distribuições gaussianas.

Na prática usam-se misturas de gaussianas que podem ser optimamente determinadas pelo algoritmo *EM*, que é reconhecido como sendo uma das técnicas estatísticas mais versáteis e potentes, uma vez que a lentidão da sua convergência está hoje em dia minimizada pela velocidade de processamento dos computadores actuais. Um dos problemas do algoritmo *EM* é a sua vulnerabilidade aos valores iniciais que podem por vezes originar um comportamento divergente.

Dado um conjunto de exemplos o algoritmo *EM* é usado para criar um número de gaussianas igual ao número N de unidades na camada escondida de rede FBR. O algoritmo tem de determinar o modelo de máxima verosimilhança para:

$$\sum_{i=1}^{N} w_i g_{[\mu_i,\Sigma_i]}(\mathbf{x})$$

em que w_i representa os pesos que têm de ser escolhidos de modo a que o integral ao longo de todo o espaço seja igual a 1. Obtêm-se assim N pesos, cada um sendo um número real, N centros, cada um vector em $\mathbf{R}^n$, e N matrizes definidas positivas, $n \times n$ cada uma, que se determinam de forma a maximizar o produto das verosimilhanças para cada exemplo, em que $g_{[\mu_i,\sigma_i]}(\mathbf{x})$ é dada pela expressão 5.1, isto é:

$$\max \prod_{x \in X} \sum_{i=1}^{N} w_i g_{[\mu_i,\Sigma_i]}(\mathbf{x})$$

Este problema é praticamente impossível de resolver pelos métodos de optimização com restrições, mas pode no entanto resolver-se facilmente pelo algoritmo EM, que se descreve a seguir Adler (1997).

1. Inicializam-se as N gaussianas de forma a que as matrizes $\mathbf{\Sigma}_i$ sejam matrizes identidade, e cada w_i inicia-se com o valor $\frac{1}{N}$, e os N centros são iniciados num ponto aleatório de $\mathbf{R}^n$ suficientemente afastado de um exemplo casual qualquer.

2. Para cada exemplo, e para cada i entre 1 e k, calcula-se a verosimilhança pesada $w_i g_{[\mu_i,\mathbf{\Sigma}_i]}(\mathbf{x})$ utilizando-se as últimas estimativas do parâmetros $w_i, \bar{\mu}_i, \mathbf{\Sigma}_i$. Para cada exemplo $\mathbf{x}$, seja $L_i(\mathbf{x}) = w_i g_{[\mu_i,\mathbf{\Sigma}_i]}(\mathbf{x})$ a máxima verosimilhança atribuída ao i-ésimo membro da mistura, seja a soma $S_{\mathbf{x}} = \sum_{i=1}^{N} L_i(\mathbf{x})$, e seja a fracção $P_i(\mathbf{x}) = \frac{L_i(\mathbf{x})}{S_{\mathbf{x}}}$. Pode pensar-se o exemplo $\mathbf{x}$ como pronto para ser separado entre as N gaussianas de acordo com as N fracções $P_i(\mathbf{x})$. Por consequência, a expressão:

$$\sum_{\mathbf{x} \in X} \sum_{i=1}^{N} P_i(\mathbf{x}) = |x|,$$

permite obter o cardinal de x.

3. Determinam-se agora valores novos para as estimativas de $w_i^{novo}, \mu_i^{novo}, \Sigma_i^{novo}$:

$$w_i^{nova} = \frac{\sum_{\mathbf{x}\in X} P_i(\mathbf{x})}{|x|},$$
$$\mu_i^{nova} = \frac{\sum_{\mathbf{x}\in x} P_i(\mathbf{x})\ \mathbf{x}}{\sum_{\mathbf{x}\in x} P_i(\mathbf{x})},$$
$$\Sigma_i^{nova} = \frac{1}{\left(\sum_{\mathbf{x}\in x} P_i(\mathbf{x}) - 1\right)} \sum_{\mathbf{x}\in x} P_i(\mathbf{x}) \left[(\mathbf{x} - \mu_i^{novo})\ (\mathbf{x} - \mu_i^{novo})^T\right].$$

4. Repetem-se os passos **2** e **3** até que não se verifiquem variações significativas entre os valores que estão a ser estimados.

O algoritmo EM de Bishop e Nabney (1997), também referido em Nabney (2002), segue aproximadamente este esquema, no entanto com algumas diferenças. Supondo que as funções de base radial são gaussianas, esquematicamente, pode descrever-se este algoritmo pelos seguintes passos:

Repete-se até que $|erro_{novo} - erro_{velho}| \leq \varepsilon$, ou até ser atingido o número máximo de iterações:

1. Calcula-se as probabilidades *a posteriori*,
2. Calcula-se o $erro = \sum \log(prob)$,
3. Ajusta-se as novas estimativas para os parâmetros,
4. Calcula-se as distancias de cada vector de entrada aos centros,
5. Calcula-se a nova matriz das covariancias,
6. Certifica-se de que as covariancias não são próximas do valor nulo e, neste caso, atribui-lhes o seu valor anterior, ou inicial,
7. Restitui a estrutura da mistura de gaussianas com os novas probabilidades a prior, centros e covariancias.

5.2.1 Resultados obtidos com um conjunto de treino de 283 casos

Os resultados reportados nesta secção foram apresentados pela primeira vez em Sereno et al. (2000b) em que se referem experiências realizadas com modelos usando o mesmo conjunto de 315 dados que tinha sido utilizado nas experiências relatadas na subsecção 5.1.2. O treino das redes de FBR não precisa de um terceiro conjunto

de validação, como foi preciso no treino das redes PMC, por consequência, o número de casos de treino passou a ser $213 + 70 = 283$ casos.

A tabela 5.6, que é análoga à tabela 5.3 (que se apresenta na subsecção 5.1.2), mostra os subconjuntos utilizados no treino, o conjunto usado para teste (coluna Tst), assim como os erros relativos da previsão do peso fetal cometidos pela rede *FBR-5:7:2*, (coluna E_{tst}), pela fórmula de Hadlock (coluna E_{had}) e pela fórmula de Shepard (coluna E_{she}).

	Conj. Treino 283 casos	**Tst** 71c.	**E_{tst}** %	**E_{had}** %	**E_{she}** %
1	$\{\mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5\}$	$\mathbb{Z}_4$	7.17	7.56	8.36
2	$\{\mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_4, \mathbb{Z}_5\}$	$\mathbb{Z}_1$	6.90	7.76	7.86
3	$\{\mathbb{Z}_3, \mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_5\}$	$\mathbb{Z}_2$	7.46	8.36	8.58
4	$\{\mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_5\}$	$\mathbb{Z}_3$	7.32	8.95	7.83
	Média	=	**7.20**	**8.16**	**8.16**

Tabela 5.6: Resultados FBR-5:7:2 (283 casos treino, 71 casos teste)

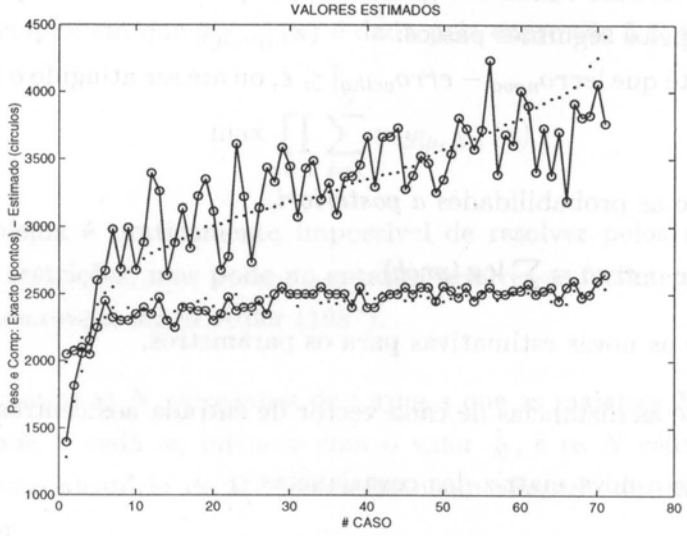


Figura 5.4: Gráfico das estimativas (círculos) obtidas com uma rede $FBR-5:7:2$ (71 casos): (a) Peso fetal, parte superior, (b) Comprimento fetal, parte inferior. Os valores verdadeiros estão representados por pontos.

A figura 5.4 mostra para 71 casos de um conjunto de teste separado, os valores estimados para o peso (escala em gramas) e para o comprimento (não tem a escala indicada), por uma rede *FBR-5:7:2*. Podem comparar-se estas estimativas com os valores verdadeiros do peso e comprimento fetal, estes últimos tendo servido como critério de ordenação (crescente) dos casos. Notar que esta rede apresentou um erro relativo médio na previsão do peso fetal de apenas 6.90% e uma percentagem de

pesos cujo erro de estimativa é inferior a 150 gramas de 42%.

Modelos	N_{unid}	N_{param}	Média do Erro (%)
PMC **5:5:2**	9	58	**7.20**

Tabela 5.7: Resultados PMC-5:5:2 (213 casos treino, 71 casos teste)

Este conjunto de experiências confirmou os resultados positivos que já tinham sido obtidos. As redes neuronais apresentam erros relativos médios de **7.17%**, os quais são inferiores aos obtidos pelas fórmulas clássicas usadas pelos obstetras, **8.16%** (Tabela 5.7). Estas experiências permitem sugerir que as redes PMC poderão ser soluções sérias à aplicação na prática clínica para efeitos de previsão do peso fetal.

5.2.2 Resultados obtidos com um conjunto de treino de 220 casos

Os resultados reportados nesta secção foram apresentados pela primeira vez em Sereno et al. (2000a) e dizem respeito a experiências realizadas com modelos em que se usou um conjunto de 220 casos para treino e um outro conjunto de 55 casos para teste. Estes 275 casos resultaram de um pré-processamento aplicado ao conjunto de 414 casos, ao qual se filtraram alguns casos menos apropriados para o treino de redes neuronais, porque estavam deteriorados (continham dados omissos, *outliers*, dados errados) ou eram *inconsistentes*. A filtragem destes últimos casos, considerados casos com dados muito inconsistentes, baseou-se num controlo da variabilidade de duas ou três medidas ecográficas, tendo-se em vista as recomendações que se encontram na literatura sobre técnicas estatísticas e aprendizagem neuronal, nomeadamente as sugestões apresentadas em Hudson and Cohern (2000) (pp.121-126) e Masters (1993) (pp.246-278) para a formação de conjuntos apropriados para o treino de redes neuronais.

Esta filtragem foi também motivada pelas experiências realizadas com modelos de regressão linear, que não se descrevem por serem muito semelhantes às que se referem na subsecção 4.2.8. Fundamentalmente procurava-se a redução do "*ruído*" e a detecção de casos com dados muito inconsistentes. Por exemplo, casos que apresentassem um valor do comprimento do fémur CFE que fosse típico de um feto de peso baixo, e um valor de diâmetro biparietal DBP que fosse típico de um feto de peso elevado. A não utilização destes casos no treino de um dado modelo não significa que esse modelo vai ficar enviesado para dar melhores resultados apenas neste conjunto filtrado de dados, nem que os casos não utilizados no treino devam ser desprezados.

O modelo não fica enviesado porque os casos foram filtrados segundo um critério bem definido, que se descreve mais abaixo, tendo originado um novo conjunto de

onde se extraíram aleatoriamente os subconjuntos de treino e de teste. Foram filtrados todos os casos que, para um determinado ponto central cujas coordenadas eram definidas pelos valores médios do PAB, DBP e CFE, apresentassem uma distância a este pontoque ultrapassasse um determinado limiar, o qual foi empiricamente ajustado de modo a não serem eliminados muito mais do que um terço dos casos e, em especial, fossem eliminados os casos pertencentes à categoria de pesos médios que, estes sim, enviesavam o treino (de facto foram eliminados 33.6% dos casos disponíveis).

Por outro lado, os casos que não foram utilizados no treino deveriam ser reutilizados para um outro modelo de rede, uma vez que se preconiza que uma boa solução do problema deveria ser concebida com vários modelos de redes trabalhando cooperativamente. Na presente fase importava obter-se um modelo de rede que previsse com fiabilidade os casos mais típicos e deixar para uma outra rede, especializada nos *casos raros* também importantes na prática clínica, para uma segunda fase (que se esboça parcialmente no capítulo seis).

Retiraram-se do espaço de entrada os dados ecográficos obtidos por modo Doppler na medida em que a sua validação suscitava dúvidas que eram suportadas quer por análise da variância comparativa entre centros hospitalares, quer pela literatura obstétrica, por exemplo, [Mandruzzato et al. (1994)]. Para a filtragem dos casos que não pertenciam à vizinhança dos centros de categorias específicas de pesos fetais estes foram analisados segundo três características PAB, DBP e CFE (previamente normalizadas). O intervalo $[1000\,\mathrm{g}, 4500\,\mathrm{g}]$, de PFE, foi dividido em sete categorias de pesos fetais cujos extremos estavam espaçados de $500\,\mathrm{g}$. Os centros de cada categoria foram definidos por vectores cujas componentes eram os valores médios das três características ecográficas dos casos pertencentes a essa categoria. Os casos eram excluídos de uma categoria se apresentavam uma distância aos centros dessa categoria superior a 1.5 (uma vez que se usavam valores normalizados). Para além desta filtragem ainda se excluíram quatro casos raros na base de dados cujos pesos tinham valores extremos, dois de pesos inferiores a $1000\,\mathrm{g}$ e os outros dois de pesos superiores a $4600\,\mathrm{g}$.

Os subconjuntos utilizados para conjuntos de treino constam da tabela 5.8: o número de épocas de treino (coluna Ne), um índice do erro de treino (coluna Err), o conjunto usado para teste (coluna Tst), os erros relativos da previsão do peso fetal cometidos pela rede *FBR-5:10:2*, (coluna E_{tst}), pela fórmula de Hadlock (coluna E_{had}) e pela fórmula de Shepard (coluna E_{she}).

A tabela 5.9 representa um exemplo de uma *matriz de confusão*. Considerou-se 7 categorias de pesos fetais tal como se indica pelos intervalos na primeira coluna (em gramas). Apresentam-se nas linhas as categorias de pesos verdadeiras (Cv) a que os casos pertencem, e nas colunas as categorias estimadas (Ce) pela rede. A

Rede	**Conj. Treino** 220 casos	**Ne**	Err	**Tst** 55c.	$\mathbf{E_{tst}}$ %	$\mathbf{E_{had}}$ %	$\mathbf{E_{she}}$ %
1	$\{\mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5\}$	39	1093	$\mathbb{Z}_4$	6.56	7.30	6.95
2	$\{\mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_4, \mathbb{Z}_5\}$	60	1080	$\mathbb{Z}_1$	7.29	8.76	7.95
3	$\{\mathbb{Z}_3, \mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_5\}$	67	1148	$\mathbb{Z}_2$	7.08	6.91	7.58
4	$\{\mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_5\}$	94	1129	$\mathbb{Z}_3$	6.87	7.56	7.31
			Média =		**6.95**	**7.63**	**7.45**

Tabela 5.8: Resultados FBR-5:10:2 (220 casos treino, 57 casos teste)

tabela 5.9 apresenta os resultados referentes à rede 4, treinada com os conjuntos $\{\mathbb{Z}_4, \mathbb{Z}_1, \mathbb{Z}_2, \mathbb{Z}_5\}$ e testada no conjunto $\mathbb{Z}_3$. Os números da diagonal são os casos em que a categoria estimada pela rede coincide com a categoria verdadeira a que o caso pertence.

Interv. Pesos	Cat.	Ce:1	Ce:2	Ce:3	Ce:4	Ce:5	Ce:6	Ce:7
[1000, 1500[	Cv:1	**0**	0	0	0	0	0	0
[1500, 2000[	Cv:2	0	**1**	2	0	0	0	0
[2000, 2500[	Cv:3	0	0	**2**	0	0	0	0
[2500, 3000[	Cv:4	0	0	0	**5**	7	0	0
[3000, 3500[	Cv:5	0	0	0	0	**17**	5	0
[3500, 4000[	Cv:6	0	0	0	0	2	**9**	0
[4000, 4500[	Cv:7	0	0	0	0	0	4	**1**

Tabela 5.9: Matriz de confusão da rede 4 (Erro de teste=6.87 %.)

Uma das medidas mais usadas em problemas de classificação é a *precisão global* (*accuracy*), dada pela razão entre o número de casos que estão correctamente classificados (pertencentes à diagonal da *matriz de confusão*) e o número total de casos (cardinal do conjunto de teste):

$$precisão = \frac{1+2+5+17+9+1}{1+4+5+26+18+1} = \frac{35}{55} = 0.636$$

A tabela 5.10 mostra os valores da precisão que se obtém com a rede FBR, a fórmula de Hadlock e a fórmula de Shepard, respectivamente, $\mathbf{P_{FBR}}$, $\mathbf{P_{HAD}}$, e $\mathbf{P_{SHE}}$. Pode-se observar que a precisão da rede é, em média, superior à precisão das fórmulas, apesar de se tratar de uma diferença pequena.

5.3 Máquina de vectores de suporte

Os modelos obtidos nas experiências que se acabaram de relatar, com PMC e FBR, baseiam-se, em última instância, no método dos mínimos quadrados, que se usa na optimização da função de custo. Ora, este método, que na Teoria da Aprendiza-

Rede	Tst	P_{FBR}	P_{HAD}	P_{SHE}
1	$\mathbb{Z}_4$	0.582	0.473	0.546
2	$\mathbb{Z}_1$	0.509	0.473	0.527
3	$\mathbb{Z}_2$	0.582	0.600	0.600
4	$\mathbb{Z}_3$	0.636	0.691	0.491
Média =		**0.577**	**0.559**	**0.541**

Tabela 5.10: Resultados FBR-5:10:2 (220 casos treino, 55 casos teste)

gem Estatística se designa por *minimização do funcional do risco empírico* Vapnik (1998) (p.34), tem sido sujeito a vários tipos de críticas e problematizações. Por exemplo, muitos resultados com uma grande importância na justificação deste método caracterizam-se por requererem condicionantes muito restritivas, e uma grande parte dos teoremas que dizem respeito a propriedades favoráveis ao método contêm restrições cujo cumprimento é imprescindível para que ele possa ser considerado óptimo. Adicionalmente, pode-se colocar dúvidas no que diz respeito a certos teoremas analíticos do método dos mínimos quadrados para o caso da estimação de funções lineares. Por exemplo, o teorema Markov-Gauss estipula que o método dos mínimos quadrados é o que apresenta a mais pequena variância de entre todos os métodos lineares e não enviesados de estimação de parâmetros de funções lineares, porém, pode perguntar-se porque é que este estimador terá obrigatoriamente que ser linear, e não pode ter outro tipo de enviesamento. Um outro exemplo, o teorema sobre a estimação de parâmetros de funções lineares corrompidas por ruído aditivo normal em que se estipula que de entre todos os estimadores não enviesados o dos mínimos quadrados é aquele que apresenta a mínima variância. Existe ainda a questão de se saber não só porque é que este estimador tem de ser não enviesado, como porque é que a função de distribuição de probabilidade do ruído tem de ser necessariamente gaussiana. Uma solução para estas e outras dúvidas do género pode residir na proposta de Vapnik da utilização do princípio da minimização estrutural do funcional de risco, que se cristaliza no algoritmo da máquina de vectores de suporte (MVS).

5.3.1 Resultados obtidos com um conjunto de treino de 72 casos

O algoritmo duma MVS baseia-se na teoria das estatísticas de amostras finitas para a qual a quantidade verdadeiramente crítica não é o número de parâmetros de um sistema mas a dimensão VC do conjunto de funções disponibilizadas por esse sistema. Uma explicação do funcionamento da MVS foi apresentada no Capítulo terceiro.

As experiências a que se refere a presente descrição foram apresentadas em Sereno et al. (2000a), e aparecem referidas nalguns aspectos teóricos em Marques_de_Sá

Entradas	Sensibilidade	Especificidade	Precisão
PAB, CFE	0.78	0.45	0.70
PAB, DBP	0.85	0.45	0.86
PAB, PCE	0.85	0.27	0.68
PAB, IRU	0.81	0.27	0.65

Tabela 5.11: Erros de Clasificação de MVS num conjunto de teste (37 casos)

(2001a) e Marques_de_Sá (2001b). Experimentou-se um algoritmo de MVS para classificação desenvolvido dentro da filosofia das "*MATLAB Neural Networks Tools*" por Gunn (1998). Uma MVS classificadora foi treinada com um conjunto de 72 exemplos. Cada exemplo consistia num vector de dois dados numéricos sobre duas características ecográficas e uma etiqueta ± 1. A primeira característica de entrada era sempre o perímetro abdominal (PAB) e a outra uma característica que poderia ser o comprimento do fémur (CFE), o diâmetro biparietal (DBP), o perímetro cefálico (PCE) ou o índice de resistência umbilical (IRU).

A etiqueta associada a cada exemplo era $+1$ se o peso fetal correspondente pertencesse à classe de peso considerada, e -1 no caso contrário. Testaram-se cinco classes de pesos, $[1500\,\mathrm{g}, 2000\,\mathrm{g}[$, $[2000\,\mathrm{g}, 2500\,\mathrm{g}[$, $[2500\,\mathrm{g}, 3000\,\mathrm{g}[$, $[3000\,\mathrm{g}, 3500\,\mathrm{g}[$, $[3500\,\mathrm{g}, 4000\,\mathrm{g}[$.

Fizeram-se experiências com vários *kernels,* opcionais no algoritmo, tendo-se optado pela utilização de um kernel de *spline* de ordem 20, por uma constante de regularização $C = 100000$. O modelo de classificação obtido foi testado num conjunto disjunto com 37 casos, tendo-se obtido os resultados indicados na tabela 5.11

Os resultados obtidos nestas experiências com redes classificadoras podem resumir-se:

- Sensibilidade média de 0.82. O valor mais alto da sensibilidade 0.85 obtido num classificador em que se usavam as medidas do PAB e DBP é ligeiramente superior aos obtidos com as fórmulas de Hadlock e Shepard, 0.60, e com rede de FBR, 0.75.

- Especificidade média de 0.36. O valor mais alto da especificidade 0.45 obtido num classificador em que se usavam as medidas de PAB e CFE é ligeiramente inferior ao obtido com as fórmulas de Hadlock e Shepard 0.55, e com rede de FBR, 0.74.

- Precisão média de 0.72. O valor mais alto da precisão 0.86 obtido num classificador em que se usavam as medidas de PAB e DBP é inferior ao que se obteve com as fórmulas de Hadlock e Shepard foi 0.55, e com rede de FBR, 0.69.

- Dado que estes valores foram calculados em cinco classes de pesos existe uma subestimação da validade clínica uma vez que se fossem testados apenas a pertença à classe mais numerosa [1500 g, 3500 g[dos pesos mais prováveis, se obteriam melhores valores para estas características.

- A previsão do peso fetal dada por PMC, FBR e MVS subestimam, dum modo geral, os pesos fetais elevados, enviesamento que pode ser devido ao escasso número de exemplos em face de um tão elevado número de pesos fetais da classe mais numerosa [1500 g, 3500 g[. Por consequência nesta classe os falsos negativos podem influenciar uma previsão errada do risco de cesariana de emergência.

5.3.2 Resultados obtidos com um conjunto de treino de 187 casos

Os resultados reportados nesta secção foram originalmente apresentados em Sereno et al. (2001a), e dizem respeito a experiências realizadas com MVS usando um conjunto de 187 casos. Treinou-se uma máquina MVS em dois subconjuntos, um de pesos baixos e o outro de pesos altos, das extremidades da fdp do peso fetal (PFE), tendo cada subconjunto aproximadamente o mesmo número de casos ($\sim$ 60 casos). Os pesos intermédios, os quais apresentam a mais elevada frequência, não foram considerados.

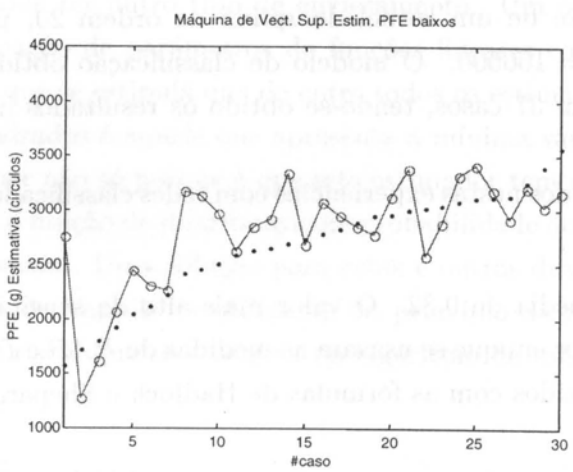


Figura 5.5: Estimação de pesos fetais **baixos** (PFE) por regressão com uma máquina de vectores de suporte (n *casos teste* = 30).

A experiência tinha por objectivo a determinação da performance da MVS em dois subconjuntos disjuntos de teste, o dos pesos baixos tinha 30 casos e o dos pesos altos 38 casos. A estimação do PFE utilizava como variáveis de entrada duas características ecográficas, o perímetro abdominal (PAB) e o comprimento do fémur

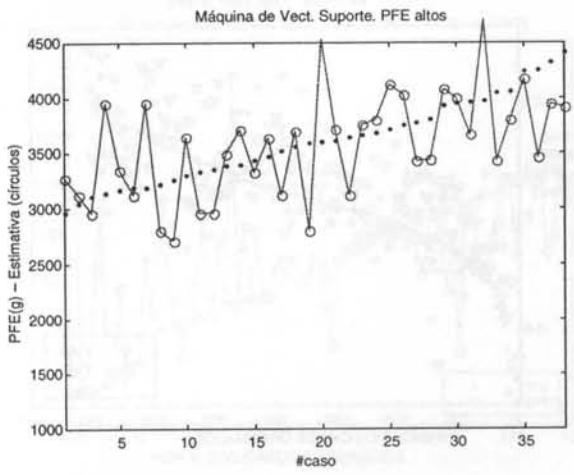


Figura 5.6: Estimação de pesos fetais **altos** (PFE) por regressão com uma máquina de vectores de suporte (n *casos teste* $= 38$)

(CFE), em duas regiões extremas da curva de distribuição. Utilizavam-se *kernels* polinomiais de grau 7, a função de custo ε-*insensitiva* com valores de $\varepsilon \simeq 0.10$ e um parâmetro de regularização $C = 1000$.

No subconjunto dos pesos *baixos* o número de vectores de suporte determinados pelo algoritmo foi 90.5% e a taxas de erro obtida 12.2%. A figura 5.5 mostra as estimativas obtidas em 30 casos representadas por círculos (as linhas que os unem são para facilitar a visualização), estando os pesos observados representados por pontos e ordenados por ordem crescente de PFE.

No subconjunto dos pesos *altos* o número de vectores de suporte determinados pelo algoritmo foi 96.5% e a taxa de erro obtida foi 10.0%, podendo ver-se na figura 5.6 uma representação análoga à que existe da figura que representa as estimativas dos pesos baixos. Estes dados permitem sugerir que as MVS permitiram melhorar o desempenho da previsão de PFE em casos com uma elevada sobreposição como nos pesos latos superiores a 4000 g.

5.3.3 Resultados de MVS treinadas com um conjunto de treino obtido por edição

Os resultados descritos nesta subsecção foram apresentados em Sereno et al. (2001b) e dizem respeito a experiências realizadas com modelos usando um conjunto de casos obtidos por um pré-processamento com uma MVS. Nestas experiências utilizou-se uma MVS,segundo a implementação de Gunn (1998), quer para fazer o pré-processamento e a selecção dos exemplos a serem utilizados para o treino, quer para fazer o treino do modelo MVS de regressão estimador do peso fetal.

O erro relativo médio conseguido foi cerca de 5% e a percentagem dos casos cuja

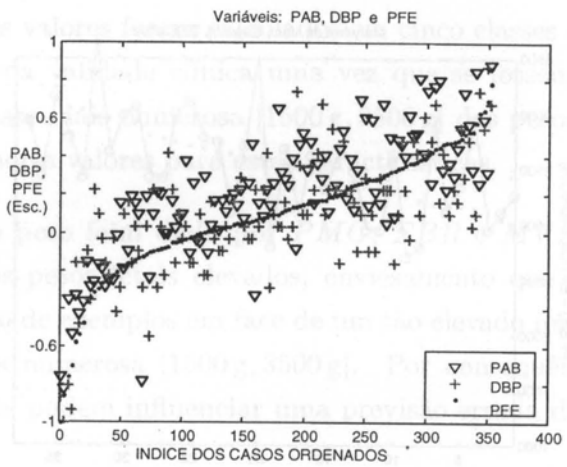


Figura 5.7: Variabilidade das medidas de perímetro abdominal (PAB), diâmetro biparietal (DBP) e peso fetal (PFE).

estimativa do PFE não ultrapassava os 5% (isto é, não excedia entre 100 g a 250 g) foi de 67%. Estes resultados são aplicáveis à previsão de novos casos cujo PFE seja superior a 2000 g e inferior a 4500 g, e constituem um melhoramento em relação aos que foram obtidos com o melhor modelo de rede de FBR, acima descrito, para o qual a taxa de erro foi de 6.9% (Tabela 5.6, subconjunto 2).

Vai a seguir descrever-se como se utilizou uma MVS numa abordagem em duas fases, a primeira, consistindo numa fase de edição na qual se realizou uma regressão para melhorar a estrutura e consistência entre os vectores de entrada e os alvos correspondentes e, seguidamente, numa fase de treino na qual uma nova máquina MVS foi treinada com base neste subconjunto "*purificado*" de exemplos, com o fim de obter um modelo estimador de PFE.

Apesar de a planificação da colheita de dados ter sido objecto de grandes cuidados que eram motivados pelo protocolo existente, este por seu lado também tinha em consideração as condições reais da prática obstétrica e o ambiente clínico em geral. Assim, com o protocolo utilizado verificou-se a existência de várias fontes de variabilidade nos dados. Destacam-se as três tipos de variabilidade:

- Intrínseca à qualidade da imagem ecográfica.
- Entre observadores na fase de extracção das medidas.
- Peso fetal para uma dada configuração corporal.

Estas fontes de variabilidade são certamente factores difíceis de considerar quando se está a ajustar um modelo de regressão a dados, tanto mais que muitas vezes existe um reduzido conhecimento a priori sobre estes factores. Tal como se pode visualizar

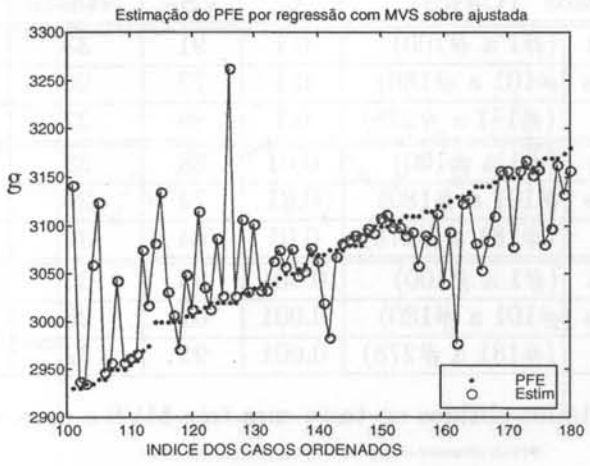


Figura 5.8: Representação do sobre-ajustamento de uma MVS regressora aos dados.

na Figura 5.7 existem bastantes "inconsistências lógicas" aparentes, isto é, valores das variáveis de entrada em conflito com pesos à nascença quase iguais.

Na primeira fase da abordagem considerada na resolução deste problema, a edição de dados por MVS, todos os casos foram normalizados e sujeitos a uma suavização utilizando uma máquina MVS com um *kernel* polinomial.

Fizeram-se diversas tentativas de combinações de parâmetros, do grau do polinómio g, dos valores da margem ε, da função de custo $\varepsilon-$insensitiva e do parâmetro de regularização C. Finalmente, adoptou-se uma MVS de regressão de *kernel* polinomial com os parâmetros $g = 7$, $\varepsilon = 0.1$ e $C = 10$. Nesta fase de edição procurou-se propositadamente um modelo mais complexo, uma vez que o objectivo era a rejeição de casos que exibissem erros devido aos factores de variabilidade mencionados e, por consequência, não ajudavam o processo de aprendizagem. Contudo, devido às limitações da implementação de MVS que se estava a utilizar, apenas foi possível um ajustamento do grau polinomial com um valor mais elevado, e um processamento de apenas duas características ecográficas dos 410 exemplos. Tentativas com parâmetros mais ajustados e mais variáveis de entrada conduziam sistematicamente a tempos computacionais indeterminados. O conjunto de treino e o conjunto de teste eram, pois, o mesmo, logo todos os casos em que se verificava um erro excessivo foram interpretados como tendo muito ruído nas variáveis de entrada ecográficas (Figura 5.8).

Nesta fase de edição os casos considerados como apropriados para o conjunto de treino/teste deviam ter resíduos inferiores a ± 50 gramas e, no caso contrário, seriam eliminados uma vez que hipoteticamente estes casos poderiam sofrer de inconsistências lógicas nos valores das variáveis de entrada. O conjunto inicial usado nesta experiência tinha 410 casos, as duas variáveis de entrada utilizadas neste pro-

Pesos (Casos)	C	VS%	Nteste	Erro%
Baixos (#1 a #100)	0.1	91	33	**6.3**
Médios (#101 a #180)	0.1	72	26	**2.2**
Altos (#181 a #278)	0.1	89	32	**4.8**
Baixos (#1 a #100)	0.01	88	33	**8.6**
Médios (#101 a #180)	0.01	74	26	**2.2**
Altos (#181 a #278)	0.01	84	32	**5.4**
Baixos (#1 a #100)	0.001	86	33	**10.9**
Médios (#101 a #180)	0.001	61	26	**2.3**
Altos (#181 a #278)	0.001	92	32	**7.4**

Tabela 5.12: Resultados obtidos no teste com três MVS e para vários valores de C

cesso de edição foram o PAB e o DBP. O processo de edição filtrou 138 casos, e deixou, portanto, 278 casos dos quais cerca de 2/3 eram usados como conjunto de treino e 1/3 como conjunto de teste.

Na fase de modelação do estimador de peso fetal utilizaram-se três MVS de regressão com *kernel* polinomial e parâmetros $g = 3$, e $\varepsilon = 0.08$. Cada uma das MVS foi treinada com casos de pesos baixos (#1 a #100), médios (#101 a #180), e elevados (#181 a #278) . Utilizando-se 3 variáveis ecográficas de entrada, PAB, DBP e CFE.

Os resultados obtidos para o erro relativo de estimação do peso fetal, num dado conjunto de teste, estão representados na Tabela 5.12, e coluna **Erro%**, para vários valores da constante de regularização C que se experimentou. No conjunto de teste foram utilizados 1/3 dos casos (coluna **Nteste**), disjuntos dos de treino. Como se pode observar o número de vectores de suporte é elevado o que sugere que o modelo está muito sobreajustado. O erro relativo médio tem o valor 4.43% para os melhores três resultados, que se observam quando $C = 0.1$.

A Figura 5.9 permite visualizar o desempenho de uma das MVS na estimação dos pesos fetais de 26 casos de teste, cujos PFE estão no intervalo $[3050\,\mathrm{g}, 3400\,\mathrm{g}]$. Os pesos estimados estão representados por círculos ligados por linhas para melhor visualização. Os pesos observados após nascimento estão representados por pontos sólidos e encontram-se ordenados por ordem crescente.

A busca de qualidade nos dados tipicamente inclui: (*i*) inspecção visual dos valores dos dados com vista detectar casos logicamente inconsistentes; (*ii*) identificação de pequenos grupos de observações discrepantes; (*iii*) verificação da metodologia de colheita de dados para descobrir fontes de enviesamentos nas medidas Cox (1992). Tal como se pode ver na Figura 5.7 existem discrepâncias óbvias, valores de PFE muito próximos, mas cujos correspondentes valores das variáveis ecográficas apresentam uma grande variação. Por exemplo, considerando os casos #50 e #100, que correspondem a PFE na ordem dos $2790 \pm 150\,\mathrm{g}$, pode observar-se que o valor do

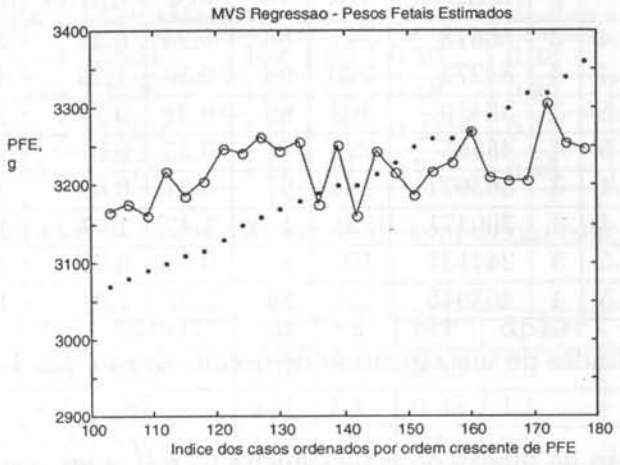


Figura 5.9: Peso fetal (PFE) estimado por uma MVS em 26 casos teste.

PAB oscila entre -80% a $+20\%$ da amplitude máximo-mínimo, e o DBP apresenta variações similares. A verificação se os dados são precisos e apropriados para o treino de redes neuronais aparece referido como de crucial importância em Hudson and Cohern (2000) e foi sempre tida em linha de conta como por exemplo se refere em Sereno et al. (2000a). Algumas ideias foram também obtidas em Marques_de_Sá (2001a) e Marques_de_Sá (2001b).

5.3.4 Experiências com uma MVS que utiliza um algoritmo de decomposição

Estes resultados não tendo ainda sido publicados dizem respeito a experiências realizadas com um conjunto de 280 casos e vários modelos de MVS segundo a implementação em $C++$ de Collobert et al. (2002) ($SVMTorchII$) do algoritmo de decomposição. Este algoritmo é muito semelhante ao de Joachims (1999), o qual por sua vez se baseia numa ideia de Osuna et al. (1997). Em Collobert and Bengio (2001) pode encontrar-se uma descrição pormenorizada deste algoritmo. Aqui apenas se refere que o algoritmo de regressão é subdividido nos 4 passos seguintes:

1. Seleccionam-se q variáveis (multiplicadores de Lagrange) α_i ou α_i^* as quais passam a constituir o novo conjunto de trabalho que se denota $\mathcal{S}$ (ver coluna "*var*" na Tabela 5.13)

2. Mantêm-se as outras variáveis com os valores correntes e resolve-se o problema

C	ε	g	numIte	var	vs	err	dpErr	eMax
100	0.4	3	58618	141	88	0.24	0.47	2.3
100	0.4	4	84272	220	93	0.36	1.26	4.0
100	0.5	3	33459	103	66	0.31	0.77	2.9
100	0.5	4	48565	228	71	0.35	0.81	2.7
1000	0.4	3	563971	184	97	0.31	0.65	2.5
1000	0.4	4	780471	231	105	1.42	13.4	13.9
1000	0.5	3	244131	107	81	0.34	0.79	3.0
1000	0.5	4	463045	117	89	1.07	7.9	10.6

Tabela 5.13: Resultados de um algoritmo de decomposição com kernel polinomial

de maximização da função objectivo sujeita às restrições (ver secção 3.1.5):

$$W(\alpha, \alpha^*) = \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \\ - \varepsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i)$$

com respeito ao conjunto $\mathcal{S}$.

3. Procuram-se as variáveis cujos valores estão compreendidos entre 0 e C durante muitas iterações seguidas (ver coluna "*NumIte*" na Tabela 5.13), e que portanto não irão mudar. Este passo chama-se a fase de *encolhimento* (*"shrinking"*) uma vez que estas variáveis são retiradas do problema (ver coluna "*var*" na Tabela 5.13).

4. Faz-se um teste para verificar se a optimização ficou concluída, senão, regressa-se ao primeiro passo.

Experimentou-se *kernels polinomiais*:

$$(a, b) \mapsto (sa \cdot b + r)^g ,$$

com $s = 0.05$ e $r = 1.0$, e para valores de $C = 100$ e $C = 1000$, valores $\varepsilon = 0.4$ e $\varepsilon = 0.5$, polinómios de grau $g = 3$ e $g = 4$, tal como se pode ver nas três primeiras colunas da Tabela 5.13. Ainda nesta mesma tabela podem-se ver alguns resultados em termos do número de iterações necessárias (coluna "*numIte*"), numero de variáveis após o encolhimento (coluna "*var*"), número de vectores de suporte que estão no modelo (coluna "*vs*"), o erro absoluto médio (coluna "*errumIte*"), desvio padrão do erro absoluto (coluna "*dpErr*"), e erro máximo (coluna "*eMax*").

C	ε	numIte	var	vs	err	dpErr	eMax
100	0.4	6864	102	80	0.19	0.28	1.4
100	0.5	3035	121	55	0.19	0.30	1.5
1000	0.4	67646	103	81	0.2259	0.45	2.3
1000	0.5	37761	104	60	0.21	0.43	2.3

Tabela 5.14: Resultados de um algoritmo de decomposição com kernel FBR gaussiano

C	ε	numIte	var	vs	err	dpErr	eMax
100	0.4	2538	107	83	0.35	1.3	3.9
100	0.5	787	151	63	0.33	1.1	3.6
1000	0.4	8521	107	118	32	244	53
1000	0.5	9526	113	105	31	240	53

Tabela 5.15: Resultados de um algoritmo de decomposição com kernelsigmoidal

Experimentou-se *kernels de função de base radial gaussiana*:

$$(a,b) \mapsto \exp\left(\frac{-\|a-b\|^2}{\sigma^2}\right)$$

com $\sigma = 1.5$, e para valores de $C = 100$ e $C = 1000$, e valores $\varepsilon = 0.4$ e $\varepsilon = 0.5$, tal como se pode ver nas duas primeiras colunas da Tabela 5.14.

Experimentou-se *kernels sigmoidais*:

$$(a,b) \mapsto \tanh\left(sa \cdot b + r\right),$$

com o factor de escala $s = 0.005$ e $r = 0.0$, e para valores de $C = 100$ e $C = 1000$, e valores $\varepsilon = 0.4$ e $\varepsilon = 0.5$, tal como se pode ver nas duas primeiras colunas da Tabela 5.15.

Comparando as tabelas 5.13, 5.14 e 5.15 reconhece-se que os melhores resultados foram obtidos com MVS cujos *kernels* são função de base radial gaussiana.

5.4 Comparação de Resultados obtidos com PMC, FBR e MVS (algoritmo de decomposição)

Nesta secção apresentam-se alguns resultados que permitem comparar que o desempenho obtido na previsão do peso fetal com modelos em rede PMC e FBR, e máquina MVS usando o algoritmo decomposição.

Utilizou-se um conjunto de dados de 280 casos com etiquetas peso à nascença PFE e comprimento à nascença (COM). Os ensaios foram realizados segundo um método de validação cruzada de 4−dobras, isto é, o conjunto de dados foi subdividido

	MVSpolinomial	MVSradial	MVSsigmoidal
C	1000	1000	1000
ε	0.85	0.60	0.50
$kernel$	$(1.0ab+0)^3$	$\frac{exp(-\|\|a-b\|\|^2}{1.8^2}$	$tanh(0.0005ab+0)$

Tabela 5.16: Valores dos parâmetros usados no treino das MVS (com Algoritmo Decmposição)

em 4 subconjuntos (D_1, D_2, D_3 e D_4) cada um com 70 casos. Para cada modelo, escolheu-se um destes subconjuntos para teste (por exemplo D_4, 70 casos) e treinou-se a rede com os outros 3 subconjuntos (por exemplo, D_1, D_2, D_3, com 210 casos).

Todos os modelos foram treinados com todas as 5 variáveis ecográficas como entradas: PAB - perímetro abdominal; CFE - comprimento do fémur; DBP - diâmetro biparietal; PCE - perímetro cefálico; IRU - índice de resistência umbilical.

Os modelos PMC tinham 5 unidades escondidas *tan-sigmoidais* e 2 unidades de saída lineares (uma era o PFE e a outra o comprimento à nascença COM). Utilizou-se como função de erro o erro médio quadrático (EMQ) e o algoritmo de optimização é Levenberg- Marquardt, com paragem antecipada determinada pela evolução do erro num conjunto de validação e noutro de teste. Os conjuntos de teste usados no treino são sempre disjuntos dos que são usados para fazer os testes de performance.

Os modelos de FBR têm 5 unidades escondidas gaussianas e 2 unidades de saída lineares (tal como nos PMC, uma era o PFE e a outra o comprimento à nascença COM). Os pesos das conexões entre as entradas e as unidades escondidas foram adaptados pelo algoritmo EM, da expectativa-maximização. Os pesos das conexões entre as unidades escondidas e as unidade de saída foram adaptados pelo algoritmo dos mínimos quadrados. O número máximo de iterações foi fixado em 100 as quais não foram necessárias dum modo geral.

Os modelos SVM foram treinados com um algoritmo de decomposição rápido ($SVMTorch\ II$) implementado em C^{++} Collobert et al. (2002). Os parâmetros de configuração das MVS que se apresentam na Tabela 5.16 foram obtidos empiricamente de modo a que o ajustamento aos dados fosse apropriado, tendo-se adoptado uma regra heurística de que cerca de 30% a 40% dos exemplos de treino fossem vectores de suporte.

A Tabela 5.17 permite comparar os resultados entre dois modelos de rede neuronal, três modelos de MVS e dois modelos clássicos de regressão linear generalizada. Uma vez que o conjunto de exemplos utilizado para o treino e teste foi obtido por selecção de casos que não apresentavam muitas inconsistências entre si, os erros relativos de previsão do peso fetal obtidos com os modelos clássicos não se afastam muito dos obtidos com as redes e as MVS. Pode notar-se ainda que os desvios

Modelo	Er %	D.P.%	Min %	Max %	N
Rede neuronal **PMC**	7.19	5.3	0.0	25.2	280
Rede neuronal **FBR**	7.37	5.1	2.3	30.5	280
MVS polinomial	9.61	7.6	0.0	41.6	274
MVS radial	8.29	6.3	0.0	37.1	277
MVS sigmoidal	**6.71**	**4.7**	**0.0**	**23.4**	280
Regress. **Hadlock**	7.21	5.4	0.0	25.0	280
Regress. **Sheppard**	7.49	5.5	0.0	26.9	280

Tabela 5.17: Estatísticas dos erros relativos obtidos com 2 modelos de rede neuronal, 3 modelos de MVS e 2 modelos clássicos de regressão (Con. treino 210 casos, Conj. Teste 70 casos) (

padrão e os valores máximos e mínimos dos erros relativos são muito semelhantes entre os vários modelos. Porém, deve destacar-se que os melhores resultados foram obtidos com a máquina MVS com *kernel* sigmoidal.

5.5 Análise numérica do algoritmo de aprendizagem de uma MVS

5.5.1 Aproximação

O ajustamento de funções pode aplicar-se a dois tipos de dados, uns que são exactos e outros em que se sabe que existem erros. Nos primeiros o ajustamento de uma função deve ser exacto nos pontos que constituem o conjunto de dados. Nos segundos, quando se sabe que poderão existir erros, procura-se obter o melhor ajuste possível em conformidade com um dado critério baseado na a tendência dos dados. Nesta secção investiga-se mais em detalhe como se comportam as MVS com um e outro tipo de dados. Por outras palavras, tenta-se responder à questão de saber se uma MVS pode ser uma boa interpoladora de dados, quando é esse o interesse, e ser uma boa aproximadora dos dados, no caso alternativo, e, para cada um destes casos quais poderão ser as melhores estratégias de treino.

Um estudo deste tipo visando as propriedades interpoladoras de uma MVS de regressão inspira-se num estudo similar referido em Smola (1996) e em Vapnik et al. (1997) usando a função $sinc(x) = x^{-1}\sin x + \xi,\ \ x, \xi \in [0,1]$, em que ξ representa uma parcela de ruído aditivo com grandeza e distribuição variáveis.

Suponha-se que se tem um conjunto de dados numéricos constituídos por vectores e que a cada vector se associam uma etiqueta dada por um número real. Suponha-se que se pretende determinar uma função aproximada, pressupondo-se que se sabe que os dados são exactos, i.é., as medidas das componentes dos vectores de entrada e das etiquetas foram realizadas sem erros, e por consequência admite-se que não

existe ruído neste conjunto de vectores e etiquetas. Como se sabe é sempre possível obter-se uma função *interpoladora* que para uma dada precisão e para cada vector de entrada originará exactamente os mesmos valores das respectivas etiquetas.

Suponhamos um caso muito simples, o estudo de caso referido no Capítulo quatro, em que se realizam medições de características ecográficas de fetos com o objectivo de se estimar o peso fetal. Sabendo três medições do *dbp*, *cfe* e *pab* (expressas em centímetros), pode-se fazer a estimativa de uma função, usando os algoritmos para regressão com *MVS*, que designaremos simplesmente por uma *MVS de regressão*. Existe a seguinte fórmula muito simples Woo and Wan (1986) que permite obter uma estimativa do: em gramas:

$$\text{Peso-à-nascença} = 1.4 \times dbp \times cfe \times pab - 200 \text{ (g)}$$

Utilizando os valores médios de *dbp*, *cfe* indicados na tabela (ver Capítulo 4) para o $dbp = 9.2cm$ e para o $cfe = 7.2cm$, pode considerar-se uma nova função real *Peso-à-nascença*, $P_{\grave{a}n}(a)$, na variável perímetro abdominal, que se designa, em vez de *pab*, simplesmente por a :

$$P_{\grave{a}n}(a) = (1.4 \times 9.2 \times 7.2)\ a - 200 \text{ (g)} \tag{5.2}$$

$$P_{\grave{a}n}(a) = (92.736)\ a - 200 \text{ (g)} \tag{5.3}$$

Agora, vai-se introduzir mais duas parcelas que visam criar alguma complexidade na classe de funções que a máquina RVS vai ter que estimar:

$$y(x) = 100(x-2) + 300\sin\left(\frac{x}{3}\right) - 300\exp\left(-\frac{(x-33)^2}{18}\right) \text{ (g)} \tag{5.4}$$

A figura (5.10) representa, a traço contínuo, o *peso-à-nascença* dado pela a expressão (5.2) e, a traço interrompido, a função peso hipotética dada pela expressão (5.4). Os pontos nesta curva assinalados com quadrados representam os seis exemplos que vão ser considerados, e que se exprimem na matriz (5.5), em que a primeira coluna contém os valores do *pab*, x (cm), e a segunda os *pesos*, y (g):

$$\mathbf{z} = \begin{bmatrix} 15 & 1012 \\ 21 & 2097 \\ 27 & 2583 \\ 33 & 2500 \\ 39 & 3785 \\ 45 & 4495 \end{bmatrix} \tag{5.5}$$

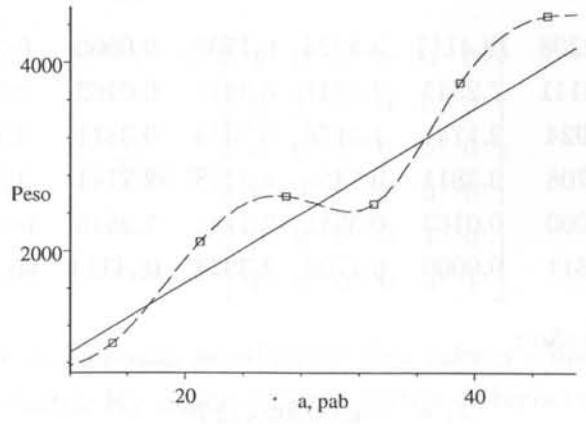


Figura 5.10: Peso-à-nascença (gramas) em função do PAB, perímetro abdominal a (cm). A Expressão 5.4 está representada a traço interronpido.

Normalização

Seguidamente, transforma–se a matriz **z** dos exemplos numa matriz contendo valores estandardizados $\mathbf{z}_{norm}$.As médias e os desvios padrão de x e y desta amostra têm os seguintes valores:

$$\begin{cases} m_x &= 30.0\,\text{cm} \\ std_x &= 11.225\,\text{cm} \\ m_y &= 2745.3\,\text{g} \\ std_y &= 1237.2\,\text{g} \end{cases}$$

Fazendo a transformação:

$$\begin{cases} x_n = \frac{(x-m_x)}{std_x} \\ y_n = \frac{(y-m_y)}{std_y} \end{cases}$$

obtém-se:

$$\mathbf{z}_{norm} = \begin{bmatrix} -1.3363 & -1.4009 \\ -0.8018 & -0.5241 \\ -0.2673 & -0.1313 \\ 0.2673 & -0.1984 \\ 0.8018 & 0.8406 \\ 1.3363 & 1.4141 \end{bmatrix}$$

Matriz de Gram

Seguidamente, apresenta-se a matriz de Gram, **K**, resultante da aplicação de um kernel polinomial de grau quatro. Um elemento K_{ij} desta matriz é dado por $K_{ij} = (x_i * x_j^T + 1)^4$, em que x_i e x_j são o i-ésimo e j-ésimo exemplos, respectivamente.

Obtém-se uma matriz 6×6 :

$$\mathbf{K} = \begin{bmatrix} 60.2208 & 18.4111 & 3.3924 & 0.1708 & 0.0000 & 0.3811 \\ 18.4111 & 7.2845 & 2.1741 & 0.3811 & 0.0163 & 0.0000 \\ 3.3924 & 2.1741 & 1.3178 & 0.7435 & 0.3811 & 0.1708 \\ 0.1708 & 0.3811 & 0.7435 & 1.3178 & 2.1741 & 3.3924 \\ 0.0000 & 0.0163 & 0.3811 & 2.1741 & 7.2845 & 18.4111 \\ 0.3811 & 0.0000 & 0.1708 & 3.3924 & 18.4111 & 60.2208 \end{bmatrix}$$

cujos valores próprios são :

$$\begin{aligned} \lambda_1 &= -0.00003\,412\,7 \\ \lambda_2 &= 0.295\,53 \\ \lambda_3 &= 1.836\,2 \\ \lambda_4 &= 3.056\,9 \\ \lambda_5 &= 65.846 \\ \lambda_6 &= 66.612. \end{aligned}$$

Para se ter uma melhor compreensão da estrutura desta matriz, e realçar a simetria desta uma matriz definida e positiva, procedeu-se ao corte (usando a função *floor)* da parte decimal nos seus elementos obtendo-se a seguinte matriz :

$$\mathbf{K}_f = \begin{bmatrix} 60 & 18 & 3 & 0 & 0 & 0 \\ 18 & 7 & 2 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 2 & 7 & 18 \\ 0 & 0. & 0 & 3 & 18 & 60 \end{bmatrix} \tag{5.6}$$

Uma questão que às vezes tem relevância na prática consiste em saber as condições que esta matriz Kernel deve obedecer no sentido de facilitar a optimização (quadrática).

Uma matriz ideal $\mathbf{K}_{ideal}$ num problema classificação dicotómica, sendo as etiquetas $\{\pm 1\}$, segundo Cristianini (2001) (slide75-80) deveria ter a forma:

$$\mathbf{K}_{ideal} \approx \begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

e uma matriz muito má $\mathbf{K}_{má}$ seria uma matriz unitária:

$$\mathbf{K}_{má} \approx \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Nos problemas de regressão as etiquetas têm valores numéricos diversos, e não apenas ± 1, logo a matriz $\mathbf{K}_f$ sugere ser uma matriz de kernel que se afasta de $\mathbf{K}_{má}$, não sendo portanto de prever problemas de redundância e incompatibilidade nos sistemas de equações da optimização.

Experimentalmente com algumas matrizes geradas por dados ecográficos provenientes do estudo de caso dos pesos fetais verificou-se que por vezes a programação quadrática não convergia alegando problemas de mau condicionamento. Outras vezes originavam valores estimados absurdos, situação que surgiu com frequência durante a experimentação de parâmetros dos modelos conducente aos resultados que se descreveram em Sereno et al. (2000a).

Definição do ε-tubo para o problema dual

Em conformidade com o problema dual referido no capítulo 3, expressão (3.35), e que se reproduz a seguir:

$$\begin{aligned} \textbf{Maximizar } W(\alpha, \alpha^*) = & \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\, k(\mathbf{x}_i, \mathbf{x}_j) - \\ & - \varepsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i) \end{aligned} \tag{5.7}$$

o primeiro somatório da expressão (3.35) é representado pela seguinte matriz 12×12:

$$\mathbf{K}_b = \begin{bmatrix} \mathbf{K} & -\mathbf{K} \\ -\mathbf{K} & \mathbf{K} \end{bmatrix}$$

As duas últimas parcelas, representadas em baixo da expressão (3.35), motivam a determinação das matrizes $(\varepsilon \times \mathbf{1}_6) - \mathbf{y}$ e $(\varepsilon \times \mathbf{1}_6) + \mathbf{y}$, em que $\mathbf{1}_6$ significa um vector coluna de com 6 células todas iguais a 1. Relembra-se que se utiliza o algoritmo *ε-insensitivo*, e que o valor de ε utilizado neste exercício é:

$$\varepsilon = 0.3$$

Para um valor de ε tão elevado a precisão obtida é baixa. Mas não se deve esquecer que o objectivo da presente análise é observar o comportamento do algoritmo num problema de ajustamento de uma função. Pode-se posteriormente aumentar a precisão, tanto quanto se desejar, dentro das limitações dadas pelo número de dados existentes. Obtém-se as seguintes matrizes:

$$[(\varepsilon \times \mathbf{1}_6) - \mathbf{y}] = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{bmatrix} - \begin{bmatrix} -1.4009 \\ -0.5241 \\ -0.1313 \\ -0.1984 \\ 0.8406 \\ 1.4141 \end{bmatrix} = \begin{bmatrix} 1.7009 \\ 0.8241 \\ 0.4313 \\ 0.4984 \\ -0.5406 \\ -1.1141 \end{bmatrix}$$

e,

$$[(\varepsilon \times \mathbf{1}_6) + \mathbf{y}] = \begin{bmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{bmatrix} + \begin{bmatrix} -1.4009 \\ -0.5241 \\ -0.1313 \\ -0.1984 \\ 0.8406 \\ 1.4141 \end{bmatrix} = \begin{bmatrix} -1.1009 \\ -0.2241 \\ 0.1687 \\ 0.1016 \\ 1.1406 \\ 1.7141 \end{bmatrix}.$$

Preparação dos dados para a programação quadrática

Cria-se a seguinte matriz 2×6:

$$\mathbf{c} = \begin{bmatrix} (\varepsilon \times \mathbf{1}_6) - \mathbf{y} \\ [(\varepsilon \times \mathbf{1}_6) + \mathbf{y}] \end{bmatrix} = \begin{bmatrix} 1.7009 & 0.8241 & 0.4313 & 0.4984 & -0.5406 & -1.1141 \\ -1.1009 & -0.2241 & 0.1687 & 0.1016 & 1.1406 & 1.7141 \end{bmatrix}$$

Considerando na expressão (3.35), referida no Capítulo 3, as restrições:

$$\sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0 \quad \mathbf{e} \quad \alpha_i, \alpha_i^* \in [0, C/m] \ \mathbf{e} \ i = 1, \ldots, m$$

Criam-se os vectores coluna de tamanhos 12×1, **vei** e **ves**, contendo, respectivamente, os valores do extremo inferior, $C = 0$, e do extremo superior, $C = 20000$, do intervalo a que podem pertencer os multiplicadores de Lagrange α_i^*, α_i:

$$\mathbf{vei} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$

$$\mathbf{ves} = 10^4 \times \begin{bmatrix} 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}^T$$

O vector coluna $\mathbf{x}_o$ especifica o vector inicial:

$$\mathbf{x}_o = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$$

A variável $NRI = 0$ estipula que o *número de restrições de igualdade* é zero, o que significa que não vai existir termo de polarização b. Uma vez que a restrição $Ax = b$ não existe, então, introduzem-se matrizes vazias $A = []$; $b = []$. Notar que estes parâmetros, NRI, A, e b, podem ser interpretados como detalhes técnicos da função com a qual se resolveu este problema pertencente á caixa de ferramentas de optimização do MATLAB, a optimização quadrática dada pela função **qp**:

$$[\mathbf{alfa}] = \mathbf{qp}\,(\mathbf{Kb}, \mathbf{c}, A, b, \mathbf{vei}, \mathbf{ves}, \mathbf{xo}, NRI)$$

A saída desta função com interesse é o vector [**alfa**], um vector coluna 12×1 contendo os valores de α_i nas primeira seis linhas e de $\alpha_i^* m$nas restantes.

Resultados determinados por programação quadrática

Obteve-se o seguinte vector coluna de multiplicadores de Lagrange:

$$[\mathbf{alfa}] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.099482 \\ 0 \\ 0.01804 \\ 0 \\ 0 \\ 0.084683 \\ 0 \\ 0 \end{bmatrix}.$$

A função estimadora pode exprimir-se em termos de um desenvolvimento em série dos vectores do suporte, o qual como se sabe tem a expressão seguinte (3.36) a qual foi referida no Capítulo 3:

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\; k(\mathbf{x}_i, \mathbf{x}_j) - b \tag{5.8}$$

Precisa-se de calcular

$$\mathbf{beta} = (\alpha_i^* - \alpha_i)\,,$$

ou,

$$\mathbf{beta} = alfa(1:n) - alfa(n+1:2*n),$$

em que n é o tamanho da amostra, neste caso temo o valor 6.

Efectuando os cálculos obtém-se o seguinte vector coluna **beta**:

$$\mathbf{beta} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.099482 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.01804 \\ 0 \\ 0 \\ 0.084683 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.018\,04 \\ 0 \\ 0 \\ -0.08\,468\,3 \\ 0.09\,948\,2 \\ 0 \end{bmatrix}$$

A partir deste resultado, utilizando a fórmula (3.31), sabendo-se que $b = 0$, e que o vector coluna com os exemplos de treino é:

$$x_{treino} = \begin{bmatrix} -1.\,336\,3 \\ -0.801\,8 \\ -0.267\,3 \\ 0.267\,3 \\ 0.801\,8 \\ 1.\,336\,3 \end{bmatrix},$$

podem obter-se para qualquer valor novo x_{teste} (perímetro abdominal, PAB) a respectiva estimativa do peso fetal usando o modelo de MVS de regressão que se acabou de treinar.

Estimativa de alguns valores de peso fetal

Sejam, por exemplo, os seguintes valores (normalizados) de x (pab) para teste:

$$\mathbf{x} = \begin{bmatrix} -1.2472 \\ 0 \end{bmatrix}$$

Seja a matriz $\mathbf{P}$, 6×2, cujos elementos são dados por:

$$P_{ij} = (x_i * x_j^T + 1)^4,$$

tem-se pois que:

$$\mathbf{P} = (\langle \mathbf{x}, \mathbf{x} \rangle + \mathbf{1})^4 = \left(\begin{bmatrix} -1.3363 \\ -0.8018 \\ -0.2673 \\ 0.2673 \\ 0.8018 \\ 1.3363 \end{bmatrix} \begin{bmatrix} -1.2472 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \right)^4 =$$

$$= \begin{bmatrix} 50.565 & 1 \\ 16.0 & 1 \\ 3.1609 & 1 \\ 0.19748 & 1 \\ 6.0524 \times 10^{-22} & 1 \\ 0.19749 & 1 \end{bmatrix}$$

A expressão (3.31) pode reescrever-se em notação matricial depois de introduzidas todas as simplificações:

$$\mathbf{y} = \mathbf{P} * \mathbf{beta}$$

A estimativa dos dois pesos fetais, agora designados y_{teste},obtêm-se em valores normalizados:

$$y_{teste} = \mathbf{P} * \mathbf{beta}$$

$$= \begin{bmatrix} \mathbf{50.565} & 16.000 & 3.161 & 0.198 & 0 & 0.198 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -0.01804 \\ 0 \\ 0 \\ -0.08468 3 \\ 0.09948 2 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -0.92896 \\ -0.003241 \end{bmatrix}$$

Transformação dos PFE para as unidades originais Vai agora proceder-se à transformação destes valores para as unidades originais, *cm* e *gramas*:

Sabe-se que:

$$\begin{cases} m_x = 30.0\,\text{cm} \\ std_x = 11.225\,\text{cm} \\ m_y = 2745.3\,\text{g} \\ std_y = 1237.2\,\text{g} \end{cases}, \quad \begin{cases} x_n = \frac{(X - m_x)}{std_x} \\ y_n = \frac{(Y - m_y)}{std_y} \end{cases}$$

Logo tem-se a transformação:

$$\begin{cases} X = x_n\, std_x + m_x \\ Y = y_n\, std_y + m_y \end{cases}$$

Se o *pab* for:

$$x_{teste} = \begin{bmatrix} -1.2472 \\ 0 \end{bmatrix}$$

Então, obtém-se:

$$\begin{aligned} X_{teste} &= \begin{bmatrix} -1.2472 \times 11.225 + 30.0 \\ 0 \times 11.225 + 30.0 \end{bmatrix} \\ &= \begin{bmatrix} 16.0 \\ 30.0 \end{bmatrix} \begin{bmatrix} 1\,\text{cm} \end{bmatrix} \end{aligned}$$

As estimativas dos pesos-à-nascença dadas pelo MVS regressora, podem ver-se na figura (5.11) representadas por cruzes:

$$\begin{aligned} \hat{Y}_{teste} &= \begin{bmatrix} y_1 \times std_y + m_y \\ y_2 \times std_y + m_y \end{bmatrix} \\ &= \begin{bmatrix} -0.928\,96 \times 1237.2 + 2745.3 \\ -0.003\,241 \times 1237.2 + 2745.3 \end{bmatrix} \\ &= \begin{bmatrix} 1596 \\ 2741 \end{bmatrix} \begin{bmatrix} 1\,\text{g} \end{bmatrix} \end{aligned}$$

Determinação dos erros

Os pesos-à-nascença exactos, dados pela expressão (5.4), podem ver-se na figura (5.11) representados por círculos.

$$\begin{aligned} Y_{teste} &= \begin{bmatrix} 100\left((16-2) + 3\sin\left(\frac{16}{3}\right) - 3\exp\left(-\frac{(16-33)^2}{18}\right)\right) \\ 100\left((30-2) + 3\sin\left(\frac{30}{3}\right) - 3\exp\left(-\frac{(30-33)^2}{18}\right)\right) \end{bmatrix} \\ &= \begin{bmatrix} 1156 \\ 2455 \end{bmatrix} \begin{bmatrix} 1\,\text{g} \end{bmatrix} \end{aligned}$$

O erro cometido:

$$erro = \begin{bmatrix} 1156 - 1596 \\ 2455 - 2741 \end{bmatrix} = \begin{bmatrix} -440 \\ -286 \end{bmatrix} \begin{bmatrix} 1\,\mathrm{g} \end{bmatrix}$$

Estão representados na figura (5.11) os gráficos dos pontos $X_{teste}, \hat{Y}_{teste}$ (quadrados), a curva dos valores estimados para os *pesos-à-nascença*, obtidos através da MVS-Regressora referida neste exemplo (linha contínua).

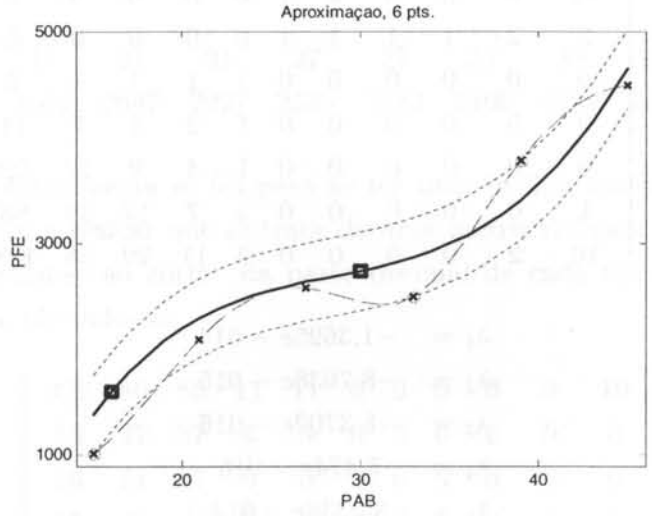


Figura 5.11: Peso-à-nascença (PFE,gramas). Pab (cm). Aproximação (linha contínua). Exacto (linha interrompida). $\varepsilon - tubo$ (linha a ponteado). $\varepsilon = 0.3$, $C = 20000$, kernel polinomial de grau 4. Vectores de suporte (círculos).

Notar que:

(*i*) Apesar da escolha do kernel dever obedecer a algum critério, existem várias regras mais ou menos heurísticas sobre a escolha apropriada a um dado problema, viu-se um exemplo de uma matriz de um kernel polinomial, e obteve-se alguma sensibilidade ao efeito deste kernel na solução do problema.

(*ii*) Um número pequeno de exemplos pode permitir identificar se um kernel será muito ou pouco apropriado a um problema.

(*iii*) Ilustrou-se um problema recorrente nesta técnica, a instabilidade numérica de certas operações matriciais, devidas aos erros de truncatura.

No caso de se utilizar um conjunto de 11 pontos de treino

$$\mathbf{z}_2 = \begin{bmatrix} 15 & 18 & 21 & 24 & 27 & 30 & 33 & 36 & 39 & 42 & 45 \\ 1012 & 1516 & 2097 & 2494 & 2583 & 2455 & 2500 & 3057 & 3785 & 4294 & 4495 \end{bmatrix}^T$$

procedendo como anteriormente obtém-se uma matriz de Gram, podendo-se trans-

formar os seus elementos por corte da parte decimal, obtendo-se a matriz $\mathbf{K}_{f2}$, que é uma matriz definida positiva e simétrica, de cada um dos seus elementos obtendo-se a seguinte matriz:

$$\mathbf{K}_{f2} \approx \begin{bmatrix} 329 & 178 & 86 & 36 & 11 & 2 & 0 & 0 & 0 & 0 & 10 \\ 178 & 100 & 51 & 22 & 8 & 2 & 0 & 0 & 0 & 0 & 2 \\ 86 & 51 & 27 & 13 & 5 & 1 & 0 & 0 & 0 & 0 & 0 \\ 36 & 22 & 13 & 7 & 3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 11 & 8 & 5 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 4 & 7 & 11 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 & 9 & 17 & 29 \\ 1 & 0 & 0 & 0 & 0 & 0 & 2 & 7 & 17 & 36 & 66 \\ 10 & 2 & 0 & 0 & 0 & 0 & 2 & 11 & 29 & 66 & 128 \end{bmatrix}$$

$$\begin{aligned} \lambda_1 &= -1.3695e-014 \\ \lambda_2 &= -8.7638e-015 \\ \lambda_3 &= -5.3702e-016 \\ \lambda_4 &= 5.474e-016 \\ \lambda_5 &= 3.3336e-014 \\ \lambda_6 &= 1.0265e-013 \\ \lambda_7 &= 1.6906 \\ \lambda_8 &= 5.3277 \\ \lambda_9 &= 11.017 \\ \lambda_{10} &= 171.45 \\ \lambda_{11} &= 456.96 \end{aligned}$$

5.5.2 Estimação

Nos problemas de *estimação* de funções reais o conjunto dos dados são encarados como obedecendo a uma dada distribuição de probabilidade que se desconhece, portanto estes dados não devem ser encarados com a mesma a confiança absoluta que se encara nos problemas de interpolação sem ruído da secção anterior. Admite-se, frequentemente, que estes poderão estar contaminados com erros aleatórios cuja origem se desconhece. Por exemplo, estes podem ter sido gerados por um processo a que se associa algum ruído aditivo. Na maioria dos casos não só o modelo deste ruído é mal conhecido, como apenas se conhecem algumas pistas sobre como sobre a sua representação funcional.

Note-se que no caso do problema de aproximação descrito na secção anterior os

pares (x, y) de dados definiam uma função, tal que a cada valor de x do seu domínio, constituído por um subconjunto dos números reais, $\mathbf{R}$, corresponde um e um só valor de y do contradomínio. O mesmo não se passa no problema da presente secção. Os exemplos, i.é., os pares (x, y), constituídos por um valor de entrada e a respectiva etiqueta, ambos reais, em geral não têm uma estrutura de função, mas satisfazem uma dada *relação ou correspondência não-univoca.*

Seja, por exemplo, o seguinte vector de exemplos, em que introduziu ruído na variável dependente, peso:

$$\mathbf{z}_3 = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 15 & 21 & 21 & 27 & 27 & 33 & 33 & 33 & 39 & 39 & 45 \\ 1012 & 2097 & 2297 & 2583 & 2783 & 2400 & 2600 & 2800 & 3585 & 3785 & 4495 \end{bmatrix}$$

Tal como anteriormente se fez para se ter uma melhor compreensão da estrutura desta matriz, e realçar que se trata de uma matriz definida positiva, mas não simétrica, procedeu-se ao corte da parte decimal de cada um dos elementos da matriz de Gram, obtendo-se:

$$\begin{bmatrix} 329 & 86 & 86 & 11 & 11 & 0 & 0 & 0 & 0 & 0 & 10 \\ 86 & 27 & 27 & 5 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 86 & 27 & 27 & 5 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 5 & 5 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 5 & 5 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 9 & 9 & 29 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 9 & 9 & 29 \\ 10 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 29 & 29 & 128 \end{bmatrix}$$

$$\begin{aligned} \lambda_1 &= -3.3092e-015 \\ \lambda_2 &= -1.4455e-016 \\ \lambda_3 &= -2.1678e-048 \\ \lambda_4 &= 9.182e-017 \\ \lambda_5 &= 7.5142e-016 \\ \lambda_6 &= 3.5542e-015 \\ \lambda_7 &= 1.7688 \\ \lambda_8 &= 6.1405 \\ \lambda_9 &= 11.546 \\ \lambda_{10} &= 142.91 \\ \lambda_{11} &= 378.25 \end{aligned}$$

Seja, agora, o seguinte vector de exemplos, em que introduziu ruído na variável independente, pab:

$$\mathbf{z}_4 = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 15 & 21 & 23 & 26 & 28 & 31 & 33 & 35 & 36 & 39 & 45 \\ 1012 & 2097 & 2097 & 2583 & 2583 & 2600 & 2600 & 2600 & 3685 & 3685 & 4495 \end{bmatrix}$$

Realça-se que se obtém uma matriz Gram em que se procedeu ao corte da parte decimal dos seus elementos, que sendo definida positiva, também não é simétrica como no caso anterior:

$$\mathbf{K}_{f4} = \begin{bmatrix} 329 & 86 & 86 & 1 & 11 & 0 & 0 & 0 & 0 & 0 & 10 \\ 86 & 27 & 27 & 5 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 5 & 5 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 11 & 5 & 5 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 9 & 9 & 29 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 9 & 9 & 29 \\ 10 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 29 & 29 & 128 \end{bmatrix}$$

$$\begin{aligned} \lambda_1 &= -3.3092e-015 \\ \lambda_2 &= -1.4455e-016 \\ \lambda_3 &= -2.1678e-048 \\ \lambda_4 &= 9.182e-017 \\ \lambda_5 &= 7.5142e-016 \\ \lambda_6 &= 3.5542e-015 \\ \lambda_7 &= 1.7688 \\ \lambda_8 &= 6.1405 \\ \lambda_9 &= 11.546 \\ \lambda_{10} &= 142.91 \\ \lambda_{11} &= 378.25 \end{aligned}$$

Os vectores de suporte versus as observações influentes em regressão

O estudo dos "resíduos" $(y_i - f(x_i))$, designação para os erros preferida pelos estatísticos, pode ter uma importância fundamental a vários títulos, uma vez que permite detectar observações que tanto podem ser aberrantes como podem desempenhar um papel importante na determinação da curva de regressão. Por outro lado,

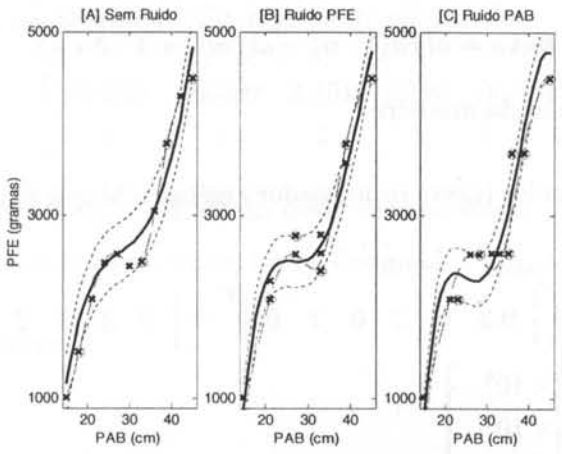


Figura 5.12: Estimativa por uma MVS regressora $\varepsilon = 0.3$, $C = 20000$, kernel Polinomial de grau 4. O conjunto de treino éconstituido por 11 pontos sem ruído [A] e com ruído [B] e [C]

o estudo dos resíduos é por vezes a única forma de verificar empiricamente a boa fundamentação de algumas hipóteses de um modelo de regressão (a sua linearidade, homocedasticidade, etc.). Deve verificar-se nos gráficos dos resíduos em função das variáveis explicativas quantitativas se existe uma expressão de alguma tendência particular.

5.5.3 Problemas de álgebra linear numérica

Considere-se os resultados obtidos pela programação quadrática no problema da subsecção anterior, para $\varepsilon = 0.1$.

Obteve-se o seguinte vector coluna dos multiplicadores de Lagrange:

$$[\mathbf{alfa}] = 10^4 \times \left[\begin{array}{cccccccccccc} 0.2 & 0 & 2 & 0 & 1 & 0 & 0 & 1 & 0 & 2 & 0 & 0.2 \end{array} \right]^T,$$

sabe-se que $b = 0$.

A função estimadora pode exprimir-se em termos de desenvolvimento em série dos vectores do suporte, como se sabe, tem a expressão que foi referida no Capítulo 3 (3.36).

$$f(\mathbf{x}) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, k(\mathbf{x}_i, \mathbf{x}_j) - b \tag{5.9}$$

Precisa-se de calcular

$$\mathbf{beta} = (\alpha_i^* - \alpha_i),$$

ou,

$$\mathbf{beta} = alfa(1:n) - alfa(n+1:2*n),$$

em que $n = 6$, tamanho da amostra.

Efectuando os cálculos (neste processador algébrico Maple 6) obtém–se o seguinte vector coluna **beta**:

$$\mathbf{beta} = 10^4 \times \left(\begin{bmatrix} 0.2 & 0 & 2 & 0 & 1 & 0 \end{bmatrix}^T - \begin{bmatrix} 0 & 1 & 0 & 2 & 0 & 0.2 \end{bmatrix}^T \right)$$

$$= \begin{bmatrix} 0.2 \times 10^4 \\ -1 \times 10^4 \\ 2 \times 10^4 \\ -2 \times 10^4 \\ 1 \times 10^4 \\ -0.2 \times 10^4 \end{bmatrix}$$

A partir deste resultado, utilizando a fórmula (3.31), sabendo-se que $b = 0$, e que o vector coluna com os exemplos de treino é:

$$x_{treino} = \begin{bmatrix} -1.3363 & -.8018 & -.2673 & .2673 & .8018 & 1.3363 \end{bmatrix}^T,$$

podem obter-se estimativas de pesos fetais dadas pela máquina de RVS que se acabou de gerar, para qualquer perímetro abdominal (*pab*) novo.

Estimativas de alguns PFE. Sejam, por exemplo, os dois valores (normalizados) seguintes de *pab* para teste:

$$\mathbf{x}_{teste} = \begin{bmatrix} -1.2472 & 0 \end{bmatrix}.$$

A expressão (3.31) pode reescrever-se em notação matricial depois de introduzidas todas as simplificações:

$$\mathbf{y} = \mathbf{P} * \mathbf{beta}$$

em que cada elemento da matriz $\mathbf{P}$ é dado por:

$$P_{ij} = (x_i * x_j^T + 1)^4,$$

Por exemplo, se *pab*$=-1.2472$,

$$P_{11} = (-1.2472 \times (-1.3363) + 1)^4 = 50.565$$

tal como se mostra na primeira linha e primeira coluna de:

$$\mathbf{P} = \begin{bmatrix} 50.565 & 16.000 & 3.161 & 0.198 & 0 & 0.198 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

os restantes elementos foram calculados deste mesmo modo.

A estimativa dos dois pesos fetais agora designados y_{teste},obtêm-se:

$$\begin{aligned} y_{teste} &= \mathbf{P} * \mathbf{beta} \\ &= \begin{bmatrix} \mathbf{50.565} & 16.000 & 3.161 & 0.198 & 0 & 0.198 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.2 \times 10^4 \\ -1 \times 10^4 \\ 2 \times 10^4 \\ -2 \times 10^4 \\ 1 \times 10^4 \\ -0.2 \times 10^4 \end{bmatrix} \\ &= \begin{bmatrix} -6.0 \\ 0 \end{bmatrix} \end{aligned}$$

Porém, notar que o valor absoluto de y_{teste} (1) , 6.0, é muito elevado, certamente que é um resultado inaceitável. Notar ainda que, por exemplo, se introduzirmos uma pequena variação $\delta P_{11} = 0.001$ na célula P_{11}da matriz $\mathbf{P}$:

$$P_{11} = 50.565 + 0.01$$

obtém-se uma grande variação, $-4 - (-6) = 2$, na primeira célula de y_{teste}.

$$\begin{aligned} y_{teste} &= \mathbf{P} * \mathbf{beta} \\ &= \begin{bmatrix} \mathbf{50.566} & 16.000 & 3.161 & 0.198 & 0 & 0.198 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.2 \times 10^4 \\ -1 \times 10^4 \\ 2 \times 10^4 \\ -2 \times 10^4 \\ 1 \times 10^4 \\ -0.2 \times 10^4 \end{bmatrix} \\ &= \begin{bmatrix} -4.0 \\ 0 \end{bmatrix} \end{aligned}$$

Estes resultados indiciam a possibilidade de existência de uma instabilidade devido à potencia $\times 10^4$ da matriz **beta**, consequentemente a necessidade de se aumentar a precisão com que os cálculos são efectuados para evitar a propagação indesejável de erros de truncatura. Esta situação é muito frequente em problemas

mal formulados, tal como se refere em Tikhonov and Arsenin (1977) ou em Twomey (1977).

Assim, utilizando a formatação longa dada pelo MATLAB, introduzindo-se na matriz **P** e **beta** uma precisão de 9 e 14 algarismos significativos, respectivamente, obtém-se:

$$
\begin{aligned}
y_{teste} &= \mathbf{P} * \mathbf{beta} \\
&= \begin{bmatrix} 50.567901234 & 1 \\ 16.000000000 & 1 \\ 3.160\,493827 & 1 \\ 0.\,197\,530864 & 1 \\ 0.000000000 & 1 \\ 0.197\,530864 & 1 \end{bmatrix}^T * \begin{bmatrix} 0.19999511881227 \times 10^4 \\ -0.99998636847713 \times 10^4 \\ 1.99997179714216 \times 10^4 \\ -2.00000000000000 \times 10^4 \\ 1.00002899439448 \times 10^4 \\ -0.20000638870116 \times 10^4 \end{bmatrix} \\
&= \begin{bmatrix} -1.191\,3 \\ 3.152\,3 \times 10^{-2} \end{bmatrix}
\end{aligned}
$$

5.6 Limites da capacidade de generalização

5.6.1 Complexidade amostral

A complexidade amostral de um conjunto de exemplos com respeito a uma dada classe de funções, pode ser abordada sob o ponto de vista donúmero de exemplos da amostra, ou da dimensionalidade dos exemplos ou de ambos os aspectos. Considerando apenas o primeiro aspecto vai determinar-se um limite do número de exemplos.

Seja um algoritmo de *aproximação-MEA*, (*"approximate-SEM"* em inglês, Anthony and Bartlett (1999), pp.258-267), isto é, minimizador do erro de amostragem para uma dada classe de funções $\mathbb{F}$, que tem como entradas uma amostra $\mathbf{z}$ e um número real positivo ϵ, e que determina uma função $\mathcal{A}(\mathbf{z},\epsilon) \in \mathbb{F}$ tal que

$$\widehat{er}\left(\mathcal{A}(\mathbf{z},\epsilon)\right) < \inf_{f \in \mathbb{F}} \widehat{er}\left(f\right) + \epsilon$$

Pode provar-se, Anthony and Bartlett (1999) (p.237), que se uma se uma classe de funções for totalmente limitada com respeito à métrica L_∞, então ela pode ser aprendida por meio de um algoritmo similar a um algoritmo de *aproximação-MEA*. Utilisando a convergência uniforme e limites dos números de cobertura pode ampliar-se este resultado a outras classes de funções mostrando que desde que $\mathbb{F}$ tenha uma *dimensão de FAT-estilhaçamento* finita, então qualquer algoritmo de *aproximação-MEA* pode ser usado para construir um algoritmo de aprendizagem para $\mathbb{F}$. Pode também obter-se limites inferiores da complexidade amostral de qualquer algoritmo

de aprendizagem em termos da *dimensão de FAT-estilhaçamento* dessa classe de funções.

Considere-se uma classe de funções $\mathbb{F}$ que aplicam um domínio $\mathbb{X}$ no intervalo real $[0,1]$, e suponha-se que a *dimensão de FAT-estilhaçamento* de $\mathbb{F}$, $\text{fat}_{\mathbb{F}}$, é finita. Seja $\mathcal{A}$ um dado algoritmo de *aproximação-MEA* para $\mathbb{F}$ e defina-se, para $z \in Z^m$, $L(z) = \mathcal{A}(\mathbf{z}, \epsilon_0/6)$, em que $\epsilon_0 = 16\sqrt{m}$. Então L é um algoritmo de aprendizagem para $\mathbb{F}$, e a sua complexidade amostral verifica:

$$m_L(\epsilon, \delta) \leq m_0(\epsilon, \delta) = \frac{26}{\epsilon^2}(18)\ln^2\left(\frac{128}{\epsilon}\right) + \ln\left(\frac{16}{\delta}\right)$$

para todo o $\epsilon, \delta > 0$.

Recordando que um conjunto $\mathbb{S} = \{x_1, x_2, \ldots x_m\}$ é estilhaçado por uma classe de funções $\mathbb{F}$ se existirem os números $r_1, r_2, \ldots, r_m$ tal que todas as 2^m combinações 'superiores/inferiores' com respeito aos r_i forem possíveis. Pode dar-se mais rigor a esta condição requrendo não só que os valores das funções x_i estejam ora acima ora abaixo, mas acima ou abaixo de pelo menos uma dada 'permissão' (*clearance*) Anthony and Bartlett (1999) (p.159).

Considere-se classes de funções de variação limitada. Diz-se que uma função $f : [0,1] \to \mathbf{R}$ tem *variação limitada* se existir um V tal que para todo o inteiro m e para toda a sucessão de números $y_1, y_2, \ldots, y_m$, com $0 \leq y_1 < y_2 < \ldots < y_m \leq 1$, tem-se que:

$$\sum_{i=1}^{m-1} |f(y_{i+1}) - f(y_i)| \leq V$$

Diz-se neste caso que f tem uma variação total no máximo de V.

Pode-se mostrar que sendo $\mathbb{F}$ um conjunto de todas as funções f que aplicam o intervalo$[0,1]$ ao intervalo $[0,1]$ e cuja variação total é majorada por V então, Anthony and Bartlett (1999) (p.160):

$$\text{fat}_{\mathbb{F}}(\gamma) = 1 + \left\lfloor \frac{V}{2\gamma} \right\rfloor$$

Para mais detalhes pode consultar-se Bartlett et al. (1996), Marques_de_Sá (2001a) e Marques_de_Sá (2001b).

5.6.2 Escolha óptima dos parâmetros C e ε

Um outro aspecto que nos interessou analisar detalhadamente diz respeito ao comportamento da aproximação dada pela MVS de regressão em face dos diferentes valores dos parâmetros de ajuste, a constante de regularização C, a tolerância ε, e o grau do kernel polinomial . A Figura (5.13) ilustra algumas ideias que, de uma

forma ou doutra, e na inexistência de uma melhor teorização, influenciaram alguns dos ajustes adoptados noutras experiências descritas mais à frente. O parâmetro C nunca pode ter um valor nulo, porque isso significaria que se desprezavam os dados, daí ter-se considerado como valor mínimo 0.01. Um aumento do parâmetro C provoca um melhor ajuste aos dados, e quando ele tem um valor ∞ significa que só os dados são importantes e se despreza a minimização da distância, quer dizer, a primeira parcela do funcional de risco regularizado. Para $C = \infty$ a margem $\|\mathbf{w}\|^2 = 0.0650$, e para $C = 0.01$ obteve-se $\|\mathbf{w}\|^2 = 0.0186$.

O parâmetro ε de precisão da função de custo insensitiva tem um efeito directo no número de vectores de suporte, os quais para para $\varepsilon = 0.01$ são todos os seis exemplos, enquanto que para $\varepsilon = 0.5$ são apenas três exemplos. Os efeitos deste parametro nas margens são, para $\varepsilon = 0.01$ e $\varepsilon = 0.5$, respectivamente, de $\|\mathbf{w}\|^2 = 0.1884$ e $\|\mathbf{w}\|^2 = 0.0297$. Note-se que existe evidencia, apesar da literatura não se referir explicitamente, que estes valores de ε devem entender-se como erros absolutos, portanto, que têm significado no espaço de valores (normalizados) que se está a usar, neste caso o quadrado centrado na origem e de lados $[-1.5, 1.5]$.

No que diz respeito ao grau do polinómio pode observar-se na figura (5.13) que quando o grau aumentou a forma de curva aproxima-se mais da verdadeira função mas ajusta-se menos aos dados. Para $grau = 4$ e $grau = 10$, respectivamente, obteve-se $\|\mathbf{w}\|^2 = 0.065$ e três vectores de suporte e $\|\mathbf{w}\|^2 = 0.0035$ e dois vectores de suporte.

Saliente-se que o número de exemplos é muito reduzidos para se fazer um estudo quantitativo dos erros, limitando-nos apenas a uma mera percepção intuitiva dada pela vizualização das curvas da figura (5.13).

5.6.3 Selecção de Modelos

O melhor modelo para um dado conjunto de exemplos é central neste capítulo. Pode experimentar-se cada um dos métodos, medir criteriosamente os respectivos desempenhos com conjunto novos e escolher aquele com maior performance, mas subsiste o problema de se saber como medir o desempenho (*performance*).

Comparações empíricas. Para uma boa compreensão dos métodos adaptativos de regressão é necessário efectuar-se comparações empíricas com o cuidado de identificar quaisquer factores subjectivos ou objectivos que possam afectar os resultados das comparações. Para que um procedimento seja aprendido com sucesso tem de ser criteriosa a escolha das funções usadas para obter a aproximação, o princípio indutivo e a forma como se realiza a optimização dos parâmetros. Através do uso de experiências bem concebidas é possível obterem-se respostas a questões bem profundas sobre a performance dos componentes individuais, pelo que as comparações empíricas são muitas vezes mais um ponto de partida para um aprofundamento do

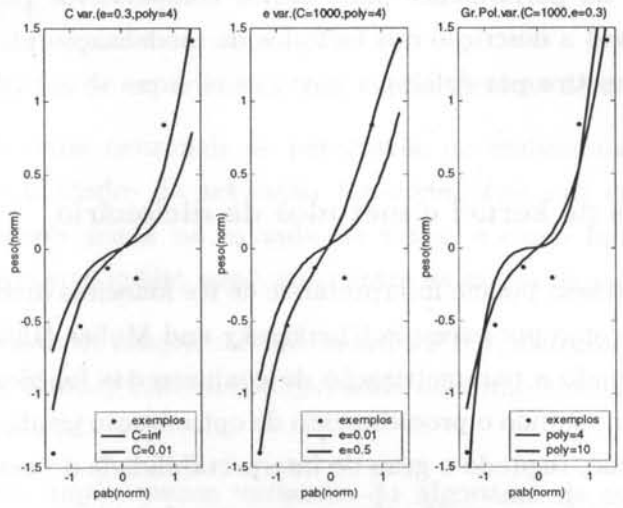


Figura 5.13: Aproximação por uma MVS-Regressora C, ε e grau do polinómio vriáveis (n^{o}casos=6).

estudo do que um ponto de chegada. A título de exemplo enumeram-se algumas regras de boa prátca na experimentação

Objectivo da comparação. Na maioria dos casos pretende-se determinar a performance preditiva dos diferentes métodos . Em geral não se tem em consideração as capacidades explicativas dos métodos, o esforço computacional ou a complexidade do algoritmo e assume-se que o treino é *off-line.*

Metodologia da comparação. Faz-se correr cada método num mesmo conjunto de treino, usando, por exemplo, quatro valores diferentes de um dado parâmetro que controla a sua complexidade. Com base na estimativa do *risco previsto* que foi obtida num conjunto de teste independente prodede-se à escolha do melhor modelo preditivo, ou parâmetro de complexidade. Seguidamente, comparam-se os quatro melhores modelos de cada um dos métodos e determina-se o vencedor. Nesta metodologia não se faz uma comparação em termos de selecção do modelo, uma vez que esta é determinada através do ajuste que o utilizador fez aos parametros de complexidade do modelo, enquanto que o interesse neste estudo reside na performance preditiva do modelo atendendo, por exemplo, às funções de aproximação usadas, ou ao procedimento de optimização especifico do treino de cada modelo.

Concepção da experiência. Tem a ver com (*i*) a especificação dos tipos de funções usadas par gerar os exemplos (no caso destes serem artificiais), (*ii*) as propriedades dos conjuntos de treino (p.ex., tamanho e ruído) e teste (p.ex., para funções de duas variáveis supondo que temos 100 pontos escolhe-se aleatóriamente pontos uniformemente espaçados numa grelha quadrada 10×10) (*iii*) a especificação

da métrica usada na performance para efeitos comparativos (em geral é o risco empírico, RMS), (*iv*) a descrição dos métodos de modelização utilizados, incluindo os ajustes de parametros por defeito.

5.6.4 Métodos de kernel e métodos de dicionário

Os métodos de regressão podem interpretar-se de tês maneiras diferentes, na opinião de alguns autores como por exemplo Cherkassky and Mulier (1998). Pode fazer-se uma distinção segundo a parametrização do conjunto das funções de aproximação implementáveis, ou segundo o procedimento de optimização usado na estimação dos parametros ou, então, segundo o grau de interpretabilidade do modelo.

Este terceiro critério parece não oferecer um grande interesse uma vez que parece ser muito subjectiva a avaliação da capacidade de interpretação da realidade subjacente a um problema de um dado modelo preditivo. Mesmo nos modelos de mais fácil interpretação, como é o caso das árvores de regressão ($CART$), esta interpretabilidade esvanece-se quando a complexidade dos modelos ultrapassa um determinado limite.

O segundo critério parece também não ser muito apropriado na medida em que o método de optimização que minimiza o funcional de risco penalizado pode ser diverso sem que se saiba bem, à partida, qual será mais vantajoso. Por exemplo, os pesos das redes neuronais PMC podem ser adaptados por um dado tipo de descida do gradiente, ou por algoritmos genéticos ou, por optimização com restrições.

Portanto, resta o primeiro critério que se baseia na parametrização de um conjunto de funções aproximadoras. É possível distinguir-se duas classes de métodos, uma baseada numa representação por meio de um "*dicionário*" de funções de base escolhidas previamente, e a outra na representação por funções "*kernel.*" Como exemplo para a primeira classe pode sugerir-se o perceptrão de multicamada PMC, e para a segunda classe pode referir-se as MVS.

Os métodos de dicionário e kernel são procedimentos de aprendizagem construtivos e universais, em que o termo "*universais*" se deve entender como qurendo significar procedimentos que dispõem de um número de parâmetros adequados e que podem aprender a partir de um número de exemplos determinável qualquer tipo de função com a precisão que se desejar. A forma como cada um destes dois métodos é parametrizada pode ser muito diferente, por exemplo, a parametrização das MVS fornece, através do número de vectores de suporte, uma caracterização significativa da complexidade das funções, a qual é independente da dimensionalidade do problema, circunstância que não acontece com as redes neuronais PMC e as FBR, representativas do método de dicionário Cherkassky and Mulier (1998).

5.7 Sumário

Relatam-se resultados de experiências com várias técnicas de modelação:

- Modelos de redes neuronais de perceptrão de multicamada, PMC, em que se utilizam unidades de activação *tan-sigmóidais* nas camadas escondidas, funções *lineares* puras na camada de saída, e como função de custo a ser minimizada o erro médio quadrático entre as saídas da rede e os valores alvo.
- Modelos de rede de funções de base radiais, FBR, utilizando o software *Netlab*, de Bishop e Nabney (2002), incorporando uma implementação muito eficiente do algoritmo *EM*.
- Modelos que implementam variantes do algoritmo de máquina de vectores de suporte, MVS, tendo por base algumas noções práticas da teoria das estatísticas de amostras finitas, na qual se considera que a quantidade verdadeiramente crítica é a dimensão VC do conjunto de funções disponibilizadas pela máquina.

O principal objectivo de investigação tratado neste capítulo são, de facto, as experiências realizadas com vista à concepção de um método neuronal de processamento das medidas ecográficas, que permita obterem-se estimativas do peso fetal mais precisas do que as que são dadas pelos métodos tradicionais, assim como apresentar análises de alguns problemas de natureza numérica computacional que surgiram na sequência destas experiências.

5.8 Referências bibliográficas e notas históricas

Um aspecto da afinação da aprendizagem pelo método da rectropropagação relaciona-se com o controlo da taxa de aprendizagem e do *momentum*. O escalonamento da taxa de aprendizagem baseado no número de entradas da rede foi sugerido por Plaut et al. (1986). A regra de actualização do delta foi apresentada por Rumelhart et al. (1986), tendo surgido nos anos seguintes várias regras propostas, por exemplo, Vogl et al. (1988) e Jacobs (1988), que propuseram o conhecido método do "*Delta-Bar-Delta*". Igualmente surgiu o método de aceleração de Silva and Almeida (1990). As técnicas de inicialização dos pesos foram também objecto de estudos vários que aparecem revistos em Thimm and Fiesler (1995), por exemplo, existe Fahlman (1988) entre muitos outros.

Uma introdução didáctica a muitos dos aspectos aplicacionais das redes neuronais e das máquinas de vetores de suporte pode encontrar-se no capítulo quinto de Marques_de_Sá (2001a) e alguns complementos sobre classificação com MVS e dimensão VC em Marques_de_Sá (2001b).

Capítulo 6

Estimação Baseada na Teoria da Informação

Introdução

No capítulo *segundo* fizeram-se múltiplas referências ao conceito de complexidade: complexidade amostral, complexidade de classes de funções (binárias e reais), complexidade da aprendizagem, complexidade como sinónimo de adaptabilidade, complexidade no contexto do compromisso viés-variância, complexidade de um modelo no sentido de graus de liberdade, e complexidade como tamanho ou número de bits que descrevem um modelo.

No capítulo *terceiro* voltou-se a fazer referência à complexidade, em dois contextos diferentes. Primeiramente, na secção 3.1 quando se aborda a regressão linear com MVS usando a função de custo insensitiva, e se sugerem duas interpretações para os valores absolutos dos coeficientes de Lagrange α, como uma medida do *conteúdo de informação* de um dado exemplo, e como uma medida da *complexidade* da função alvo na representação dual. Depois voltou-se a faze referência, na secção 3.2, quando se aborda a teoria da regularização e se estabelece a expressão da função de custo dada pela soma de duas parcelas, uma que penaliza a complexidade do modelo (a rugosidade das funções adaptadas aos exemplos), e a outra que se exprime pelo quadrado da norma do vector de pesos.

Na secção 4.4 introduziu-se a ideia, sugerida por Bishop (1995), de que os métodos tradicionais de treino de redes neuronais baseados na minimização de uma função de custo deduzida a partir de um princípio subjacente ao processo de amostragem, tal como o princípio da máxima verosimilhança, são susceptíveis de terem várias deficiências, tais como, por exemplo, a dificuldade de determinação do nível apropriado da complexidade de um modelo. Um modelo mais complexo, seja porque tem um número excessivo de unidades escondidas, seja porque o valor do parâmetro

de regularização é mais pequeno, permite um melhor ajustamento aos dados, mas, quando a complexidade do modelo é demasiado grande este pode ter uma capacidade de generalização pobre, porque se verifica o fenómeno do sobreajustamento. Tal como se referiu no quinto capítulo, secção 5.1, este fenómeno indesejável pode evitar-se reservando no conjunto de dados um subconjunto com o qual se faz a validação e optimização da performance e complexidade do modelo.

A *performance* de um modelo pode ser medida de duas maneiras diferentes. No fim da aprendizagem pode medir-se a probabilidade do modelo cometer um erro num exemplo novo, medida que pode ser vista como uma função do número de exemplos m, e que é conhecida por *curva de aprendizagem* do modelo. Pode também medir-se pela contagem cumulativa dos erros do modelo, o número total dos casos nos quais os erros cometidos pelo modelo $f(\cdot)$ são superiores a um dado limite δ, isto é, pelo cardinal do conjunto$\{(x_i, y_i) \mid |y_i - f(x_i)| \geq \delta\}$. Estas duas medidas estão intimamente relacionadas uma com a outra, em ambas está-se interessado no comportamento assimptótico do algoritmo de aprendizagem à medida que o número de exemplos m vai crescendo. A curva de aprendizagem pode servir para determinar o tamanho que m deve ter de forma a que a probabilidade de erro seja inferior a um dado valor e, neste caso, pode dizer-se que o estudo das curvas de aprendizagem de um modelo também permite uma visualização da *complexidade amostral* da aprendizagem.

A investigação recente baseada na teoria estatística da aprendizagem conduziu ao desenvolvimento de pelo menos dois pontos de vista bastante gerais, e bem sucedidos, sobre o processo de aprendizagem em termos de curvas de aprendizagem e de erros cumulativos. Um destes pontos de vista remonta ao conceito de distribuição livre de Valiant (ver por exemplo Haussler et al. (1994b)), ou modelo *PAC*, o qual tem raízes, quer na teoria do reconhecimento de padrões Duda et al. (2001), quer na teoria "*minimax*" da decisão Berger (1985). Nesta perspectiva a complexidade amostral da aprendizagem de um conjunto de exemplos é caracterizada por uma distribuição livre, em termos pessimísticos, no *pior dos casos,* através de uma parâmetro combinatorial, que se definiu no segundo capítulo, a *dimensão VC.* O outro ponto de vista tem sido investigado num contexto essencialmente Bayesiano, sofreu influências de ideias provenientes, quer da física estatística, quer da teoria da informação, apresenta uma caracterização da complexidade amostral da aprendizagem de uma distribuição livre em termos da *média dos casos.* Cada um destes pontos de vista e respectivas teorias tem as suas vantagens e inconvenientes próprios, contudo, segundo Haussler et al. (1994a), as relações entre as duas não são fáceis de compreender.

Tendo em consideração a abundância, em capítulos anteriores, de referências ao conceito de complexidade de um modelo, procura-se, nas secções seguintes, apresen-

tar uma revisão dos principais conceitos e fundamentos da teoria da informação no contexto dos modelos de rede neuronal.

Neste capítulo, faz-se primeiramente uma revisão de algumas utilizações que o termo *entropia* tem na teoria estatística da aprendizagem. Esta parte contém uma abordagem de algumas variantes do conceito de entropia no contexto da estatística matemática, tais como, a entropia empírica, a entropia de um conjunto de funções e a entropia estatística de arrefecimento ("*annealed*", in Vapnik (1998), p.95,118), as quais se podem aplicar na análise das condições de convergência uniforme dos processo empíricos. Seguidamente, faz-se uma revisão das principais variantes do conceito de *entropia*, no contexto da teoria da informação.

A entropia, em ambas as situações, permite uma quantificação e manipulação da redundância dos exemplos existente num conjunto de aprendizagem, a qual pode ser devida a muitos factores tais como, a existência eventual de várias fontes de informação. Serão revistos os conceitos de *entropia de uma variável aleatória,* e de informação mútua de *duas variáveis aleatórias,* os quais são devidos a Shannon e Renyi, o conceito de *entropia relativa,* ou divergência, que é devido a Kullback e a Leibler, o de potencial informativo e força informativa sugeridos recentemente por Principe et al. (2000b). Estas últimas medidas permitem investigar os valores extremos de uma dada função de custo, tal como vem referido em Haykin (1999) (p.484), Principe et al. (2000b) e Haussler and Opper (1997a).

A secção 6.1 apresenta uma perspectiva teorético-informacional da aprendizagem segundo Mackay (2002) e Deco and Obradovic (1996).

A secção 6.2 faz uma revisão dos principais conceitos da teoria da informação: entropia, entropia relativa e informação mútua

A secção 6.3 faz uma revisão de um método de determinação não-paramétrico de funções de distribuição designado método das janelas de Parzen (gaussianas).

Finalmente, a secção 6.4 descreve três experiências de estimação do peso fetal com modelos PMC treinados por um critério informático-teorético.

6.1 Perspectiva teorético-informacional da aprendizagem

Esta secção apresenta uma reformulação do problema da aprendizagem supervisada de um modelo de rede neuronal num enquadramento novo. A ideia consiste em se partir de uma dada arquitectura de rede, verificando-se a ausência de informação *a priori* sobre o fenómeno em que se está interessado, saber-se com e se é possível extrair a estrutura estatística implícita dos dados, em termos de *informação* a ser memorizada pela rede.

Seja um modelo não-paramétrico gerado por uma um perceptrão simples. O modelo resulta de uma adaptação de um vector de pesos $\mathbf{w}$ em resposta a um conjunto de exemplos, constituído por m etiquetas, $E_m = \{y_i\}_{i=1}^m$, que estão associadas a determinados pontos do espaço das entradas $\{\mathbf{x}_i\}_{i=1}^m$. Após a adaptação dos pesos aos exemplos utilizam-se esses pesos num processo de previsão da etiqueta que irá corresponder a um novo exemplo. Este processo pode ser visto como um *processo de comunicação*, no qual um emissor examina os exemplos E_m nos pontos $\{\mathbf{x}_i\}_{i=1}^m$ e cria uma "*mensagem*" $\mathbf{w}$ que é dependente desses dados. Por sua vez, o receptor pode tentar uma reconstrução com base nos pesos $\mathbf{w}$ das etiquetas originais E_m, podendo dizer-se neste caso que a unidade neuronal fez uma "*memorização*" (ver Tabela 6.1).

$\{y_i\}_{i=1}^m \longrightarrow$	Algoritmo de Aprendizagem	$\longrightarrow \mathbf{w} \longrightarrow$	$\mathbf{w} \longrightarrow$	$\{\hat{y}_i\}_{i=1}^m$
	$\uparrow$		$\uparrow$	
	$\{\mathbf{x}_i\}_{i=1}^m$		$\{\mathbf{x}_i\}_{i=1}^m$	

Tabela 6.1: Aprendizagem de uma rede neuronal vista como um processo de comunicação

Contudo, esta não é a tarefa mais relevante do modelo. A tarefa mais nobre do modelo consiste em adquirir a capacidade de "*generalização*", a qual significa que essa unidade de processamento tendo por base o conjunto de exemplos usados na adaptação deverá ser capaz de fazer a extrapolação de uma etiqueta desconhecida y_{m+1} a partir de um novo ponto x_{m+1}, que não era pertencente ao conjunto de treino. Da mesma forma que os estados físicos das regiões de um disco de memória, ou de um dispositivo *RAM* podem ser vistos como um canal de comunicação, os pesos $\mathbf{w}$ de uma rede adaptada podem também ter o papel de canal de comunicação, fornecendo *informação* a um utilizador futuro sobre os exemplos de aprendizagem. Uma das questões que se pode colocar é a de saber qual será a "*capacidade* " deste canal, por outras palavras qual será, a "*quantidade de informação*"que uma unidade neuronal treinada poderá armazenar.

Se um dado algoritmo de aprendizagem originasse um modelo neuronal tal que para todas as entradas possíveis a sua resposta fosse sempre $+1$, ou alternativamente, sempre -1, dependendo de ter sido utilizado na aprendizagem o conjunto A ou o conjunto B, respectivamente, então, os pesos permitiam neste caso distinguir estas duas espécies de conjuntos diferentes. A informação máxima capaz de ser fornecida sobre estes dados por este algoritmo seria então del *bit* no caso de os conjuntos A e B serem acontecimentos equiprováveis.

A questão de saber a quantidade de informação adicional que se poderá fornecer utilizando uma rede neuronal para representar outras funções ultrapassa o âmbito desta secção, mas, no caso simples que se vai referir a seguir existe uma única unidade

de processamento de exemplos binários com n entradas, e uma única função de activação em degrau de Heaviside, a qual pode mostrar-se possuir uma capacidade de *dois bits por conexão*, logo, podendo armazenar $2n$ bits de informação.

Para se estabelecer este resultado é necessário definirem-se algumas regras que excluam afirmações menos interessantes, tais como a de que "*a capacidade de uma unidade de processamento pode ser infinita uma vez que os seus pesos são números reais e, portanto, pode ter um número infinito de bits*". Esta afirmação pode excluir-se notando que o receptor não é capaz de examinar os pesos directamente, nem o receptor é autorizado a testar os pesos através da observação da saída da unidade de processamento relativa a entradas escolhidas arbitrariamente. Torna-se necessário impor a restrição ao receptor de observância das saídas da unidade de processamento no mesmo conjunto dado de m pontos $\{x_i\}_{i=1}^{m}$ existentes no conjunto de treino.

O número de funções distintas que uma unidade de processamento é capaz de produzir, dado que os valores da função só são conhecidos em m pontos, consiste em saber o número de conjuntos de etiquetas binárias diferentes para os m pontos e que se poderão obter com uma função em degrau de Heaviside linear:

$$y = f\left(\sum_{i=1}^{m} w_i x_i\right)$$

em que

$$f(a) = \left\{ \begin{array}{cc} 1 & a > 0 \\ -1 & a \leq 0 \end{array} \right\}$$

O número máximo de etiquetas binárias é 2^m. Se quase todas estas 2^m etiquetas podem ser realizadas pela unidade de processamento, então, pode dizer-se que ela constitui um canal de comunicação e que é capaz de transmitir todos os m bits (os valores das etiquetas $\{t_i\}_{i=1}^{m}$) com uma pequena probabilidade de erro. Para se identificar a capacidade de uma unidade de processamento, isto é, o máximo valor de m tal que a probabilidade de erro seja muito pequena, considera-se o seguinte cenário.

Tem-se um emissor com uma unidade de processamento de n entradas, e um conjunto de dados Z constituído por uma etiquetagem de m pontos em posição geral. O emissor usa um algoritmo adaptativo para tentar encontrar um vector de pesos $\mathbf{w}$ que possa reproduzir exactamente esta etiquetagem. Deve assumir-se nesta análise a existência de um algoritmo perfeito para descobrir esse tal vector $\mathbf{w}$, caso tal vector .exista. Caso este não exista os valores de $\mathbf{w}$ são ajustados aleatoriamente. O receptor faz a avaliação da função em degrau para cada um dos m exemplos de entrada, interessando-se em conhecer quer probabilidade de todos os m bits serem correctamente reproduzidos, quer o máximo valor de m para o qual, dado um número n de entradas da unidade de processamento, se verifica que esta probabilidade está

muito próxima da unidade. Estas são questões que pertencem claramente à *teoria da informação* e às quais se procura abordar nas secções seguintes.

6.2 Entropia e informação

O conceito de entropia foi originalmente introduzido em física, para exprimir uma medida do grau de desordem de um sistema termodinâmico, tendo posteriormente adquirido, no âmbito da matemática, um estatuto mais abstracto e formalizado em diversas teorias. Salientam-se, a título de exemplo, a teoria da informação de Shannon e de Renyi, as quais por sua vez constituem uma ampliação de teorias anteriormente desenvolvidas por Nyquist e Hartley. Também a teoria das equações diferenciais em derivadas parciais (condição de entropia), a teoria do caos em sistemas dinâmicos (entropia de Kolmogorov, ou entropia empírica), a estatística matemática (entropia de Fisher ou entropia de uma distribuição de probabilidade) apresentam contribuições para uma elucidação deste conceito. Como se verá o conceito de entropia define-se com base no conceito de probabilidade.

Existem duas utilizações comuns do conceito de probabilidade. A primeira ocorre quando é preciso descrever sucessões de ensaios aleatórios. É por vezes difícil obter-se neste caso uma definição dos termos *frequência* e *aleatório* que não seja circular, uma vez que surge a necessidade de se definir sempre o conceito de "*em média*", o qual, por sua vez, também não é fácil de explicitar sem recorrer a termos que por sua vez são sinónimos de probabilidade. A segunda utilização ocorre quando se faz referência a *graus de crença,* quando se produzem afirmações em que não estão envolvidas variáveis aleatórias. Esta situação também não é fácil de explicitar, uma vez que se admite que podem existir crenças que não são racionais, uma vez que se requer um espaço probabilizado que satisfaça certas regras de consistência simples, tais como os axiomas de Cox (1946). Estes axiomas visam garantir o chamado ponto de vista Bayesiano, para o qual duas pessoas possuindo os mesmos pressupostos devem inferir conclusões idênticas sempre que os dados recebidos são os mesmos (Mackay (2002), p.30). Uma das questões que se pode colocar, atendendo a esta dualidade de definições do conceito de probabilidade, consiste em saber porque será mais vantajoso associar-se ao conceito de "*informação*", em lugar do conceito de função de densidade de probabilidade (fdp), um outro conceito, "*entropia*", que se define a partir da fdp, mas que é tão pouco intuitivo.

Uma das razões, referida por Cover (1991) (p.5), poderá prender-se com o facto de as questões que inicialmente eram abordadas pela teoria da informação estarem nas áreas da compressão e transmissão de dados, e as respostas a tais questões assentarem em determinadas grandezas, tais como a entropia e a informação mútua, que por sua vez são funções das distribuições de probabilidade subjacentes ao processo

de comunicação.

Uma outra razão, referida por Duda et al. (2001) (p.32), poderá ter a ver com o facto de existir uma forte relação entre a distribuição gaussiana e a entropia. Sendo a entropia uma medida da incerteza fundamental existente nos valores de pontos seleccionados aleatoriamente de uma distribuição, no caso de esta ser gaussiana:

$$P(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{\frac{-x^2}{2\sigma^2}}$$

então:

$$H(x) = -\int f(x) \ln(f(x))\, dx = \ln\sqrt{2\pi e\sigma^2} \text{ nats}$$

(demonstrada em Deco and Obradovic (1996) (p.13), ou em Shannon (1948) (p.37)) podendo ainda demonstrar-se que este é o valor máximo de entropia que é possível obter-se entre todas as distribuições possíveis com os mesmos valores fixos de média e variância.

6.2.1 Informação de Fisher

Sendo conhecida uma amostra casual de uma variável aleatória X, com distribuição de frequência $f(X, \theta)$, desde 1935 que Fisher (Fisher (1973), pp., 214-218) já vinha a referir-se, na teoria da estimação (paramétrica), a uma grandeza que designou *informação observada* sobre um parâmetro θ:

$$-\ell(\theta) = -\frac{\partial^2}{\partial\theta^2} \log f(x \mid \theta)$$

A *informação de Fisher* $I(\theta_0)$ define-se como o valor esperado desta informação obtido no verdadeiro valor do parâmetro:

$$I(\theta_0) = \mathbf{E}\left[-\frac{\partial^2}{\partial\theta^2} (\log f(X \mid \theta))\Big|_{\theta=\theta_0}\right] \tag{6.1}$$

Esta grandeza pode interpretar-se como uma medida da curvatura esperada da função de verosimilhança $\log f(X \mid \theta)$ no valor verdadeiro do parâmetro. Sob determinadas condições gerais de regularidade o recíproco de $I(\theta_0)$ define um *limite inferior* da variância de um estimador de θ não enviesado (um estimador de máxima verosimilhança pode ser enviesado), de tal forma que quanto maior for a variância tanto menor será a "informação". Por outras palavras, se T for um estimador não enviesado de θ, tal que $\mathbf{E}(T) = \theta$, então a variância:

$$var(T) \geq \frac{1}{m\int_{-\infty}^{+\infty} \frac{\partial^2}{\partial\theta^2} (\log f) f dx}$$

O conceito de entropia está na base do conceito de informação, assim como de muitas ideias estatísticas tais como "*predição*", "*equivalência entre modelação e compressão de dados*", e "*verosimilhança*" (Mackay (2002), pp.30-33).

Existe um princípio de verosimilhança em que se afirma ser possível determinar a informação em qualquer amostra a partir da função que especifica a probabilidade da amostra observada e que se chama *função de verosimilhança.* Por exemplo, seja a observação de m experiências independentes de lançamento de uma moeda ao ar, em que a probabilidade de se obter uma face voltada para cima se supõe ser constante, isto é, igual a p. Se nestas experiências ocorrerem X faces voltadas para cima, então, a função de verosimilhança define-se:

$$L(p) = kp^X(1-p)^{m-X}$$

em que a constante de proporcionalidade k não tem nenhum significado especial sendo de facto o parâmetro p que é relevante. Em particular, este princípio de inferência sugere que não é importante saber-se se o número de ensaios foi m, ou se foi por capricho que se decidiu interromper os ensaios. A única coisa importante consiste em saber-se *qual é o valor* da probabilidade p com que os dados surgem para cada uma das hipóteses possíveis (Mackay (2002) refere não ser este princípio aceitável para alguns estatísticos frequencistas).

6.2.2 Entropia de Vapnik

Referiu-se no Capítulo2 (secção 2.2.3) que mesmo desconhecendo-se a função de distribuição de probabilidade das fontes de informação que geram os exemplos de treino, pode demonstrar-se, como se faz na teoria da aprendizagem estatística, Vapnik (1998) (pp.79-95), que existe *consistência* no princípio da minimização do risco empírico (MRE). Com base num número mínimo m de exemplos de uma amostra :

$$\{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m .$$

pode sempre determinar-se o valor mínimo do funcional *risco esperado* $R[f(x, w_m)]$ através da determinação do valor mínimo do funcional *risco empírico*, $R_{emp}(w_m)$. Por outras palavras, para um conjunto de funções de custo $c(x, y, f(x, w))$, $w \in \mathbb{W}$, e para uma distribuição de probabilidade $P(x, y)$, os riscos verdadeiro e empírico são convergentes *em probabilidade* para um mesmo limite:

$$\inf_w R[f(x, w)] .$$

Excluindo-se agora o caso trivial, pode dizer-se que, para um conjunto de funções de custo $c(x, y, f(x, w_m))$, o método da minimização do risco empírico (MRE)

é *estritamente consistente,* se para qualquer subconjunto não vazio, $\mathbb{X}(c)$, $c \in (-\infty, +\infty)$, do conjunto de funções de custo, tal que:

$$\mathbb{X}(c) = \left\{ w : \int c(x, y, f(x, w))\, dP(x, y) \geq c \right\} \tag{6.2}$$

existe convergência verificando-se apenas uma condição:

$$\inf_{w \in \mathbb{X}(c)} R_{emp}(w_m) \xrightarrow[m \to \infty]{P} \inf_{w \in \mathbb{X}(c)} R(w) \tag{6.3}$$

Quer dizer, para o método MRE ser estritamente consistente tem de verificar-se a convergência indicada na expressão 6.3, quer se trate do conjunto dado de funções, quer de um qualquer seu subconjunto de funções, $H(c)$, obtido através da eliminação de todas as funções cujos riscos sejam inferiores a c.

Recorde-se o *teorema chave* da teoria da aprendizagem o qual estabelece que, para uma dada função de distribuição $P(x, y)$, se existirem os números reais a e A que limitam, inferior e superiormente, todos os funcionais de risco esperado que se podem obter com o conjunto de funções de custo $c(x, y, f(x, w))$, $w \in \mathbb{W}$, então, o método MRE é estritamente consistente para o conjunto de todas as funções de custo, o que é equivalente a dizer-se que se verifica a convergência uniforme *simples* (*one-sided*) das médias para as respectivas esperanças matemáticas.

Para se traçar o caminho sugerido por Vapnik para o conceito de entropia de um conjunto de funções (indicatrizes) é preciso fazer-se uma referência a condições de convergência uniforme no modelo mais simples.

Suponha-se que dado um conjunto de acontecimentos contendo um número finito N de elementos,

$$A_k = \{(x, y) : c(x, y, f(x, w_k)) > 0\}, k = 1, 2, \ldots N,$$

se verifica a convergência uniforme, e por consequência que são válidas as seguintes desigualdades:

$$P\left\{ \max_{1 \leq k \leq N} |P\{c(x, y, f(x, w_k)) > 0\} - \nu_m\{c(x_i, y_i, f(x_i, w_k)) > 0\}| > \varepsilon \right\} \tag{6.4}$$

$$\leq \sum_{k=1}^{N} P\{|P\{c(x, y, f(x, w_k)) > 0\} - \nu_m\{c(x_i, y_i, f(x_i, w_k)) > 0\}| > \varepsilon\} \tag{6.5}$$

$$\leq 2N \exp\{-2\varepsilon m\} = 2 \exp\left\{ \left(\frac{\ln N}{m} - 2\varepsilon^2 \right) m \right\} \tag{6.6}$$

em que se utilizou a desigualdade de Chernoff referida no Capítulo 2 para se obter a expressão 6.6. A última expressão sugere que para existir convergência uniforme

para qualquer ε deve verificar-se:

$$\frac{\ln N}{m} \underset{m\to\infty}{\to} 0 \tag{6.7}$$

No caso de ser finito o número N de acontecimentos e o número de observações m tender para infinito é evidente que se verifica a convergência para o limite nulo. Isto também prova que para qualquer conjunto com um número finito de acontecimentos existe convergência uniforme. Contudo, numa amostra $\{z_i\}_{i=1}^m$ uma relação do tipo 6.8 será indicadora de convergência uniforme neste caso, em que é infinito o número de acontecimentos do conjunto, porque serão *distintos* (ou, *distinguíveis*) apenas um número finito de *aglomerados* de acontecimentos, entendendo-se por acontecimentos *distintos* aqueles para os quais se verifica existir pelo menos um elemento da amostra que pertence a um acontecimento mas não pertence ao outro. Seja o número de aglomerados, $N^{\mathbb{W}}(\{z_i\}_{i=1}^m)$, o qual depende quer da amostra, quer do conjunto de funções. Seja um conjunto de funções indicatrizes:

$$C(x_i, y_i, f(x_i, w)), w \in \mathbb{W},\ i = 1, \ldots, m,,$$

e um conjunto de exemplos $\{z_i\}_{i=1}^m$. Com estes exemplos e funções indicatrizes pode determinar-se o conjunto dos vectores binários $m-$dimensionais:

$$c(w) = \{C(x_1, y_1, f(x_1, w)), \ldots, C(x_m, y_m, f(x_m, w))\},\quad w \in \mathbb{W} \tag{6.8}$$

Para um dado valor $w = w^*$ o vector binário $c(w^*)$ determina um dado vértice do cubo unitário. O número de vértices diferentes induzidos quer pela amostra, quer pelo conjunto de funções indicatrizes, é dado pela função:

$$N^{\mathbb{W}}(\{z_i\}_{i=1}^m) \equiv N^{\mathbb{W}}(z_1, \ldots, z_m)$$

a qual por sua vez se sabe que:

$$N^{\mathbb{W}}(\{z_i\}_{i=1}^m) \leq 2^m.$$

Supondo que esta função é mensurável com respeito à medida de probabilidade:

$$P(\{z_i\}_{i=1}^m) = P(z_1, \ldots, z_m) = \prod_{i=1}^{m} P(z_i) \tag{6.9}$$

então, chama-se à quantidade:

$$H^{\mathbb{W}}(\{z_i\}_{i=1}^m) = \ln N^{\mathbb{W}}(\{z_i\}_{i=1}^m) \tag{6.10}$$

entropia aleatória do conjunto de funções indicatrizes $C(x_i, y_i, f(x_i, w))$, $w \in \mathbb{W}$, na amostra $\{z_i\}_{i=1}^{m}$.

A *entropia* do conjunto destas funções, e na amostra, define-se pela quantidade:

$$H^{\mathbb{W}}(m) = \int H^{\mathbb{W}}(\{z_i\}_{i=1}^{m})\ dP(z_1, \ldots, z_m). \tag{6.11}$$

Existe um teorema que diz que observando-se certas condições apropriadas de mensurabilidade de um conjunto de funções, para que se verifique no conjunto de funções indicatrizes $C(x_i, y_i, f(x_i, w))$, $w \in \mathbb{W}$, a convergência uniforme:

$$P\left\{\sup_{w\in\mathbb{W}}\left|\int c(x, y, f(x, w))\, dP(x, y) - \frac{1}{m}\sum_{i=1}^{m} c(x_i, y_i, f(x_i, w_k))\right| > \varepsilon\right\} \underset{m\to\infty}{\to} 0$$

é condição necessária e suficiente:

$$\frac{H^{\mathbb{W}}(m)}{m} \underset{m\to\infty}{\to} 0 \tag{6.12}$$

Vai a seguir fazer–se uma extensão deste resultado a um conjunto de funções reais. Começa-se, primeiro, por considerar funções uniformemente limitadas:

$$C(x, y, f(x, w)), w \in \mathbb{W},$$

tal que:

$$|C(x, y, f(x, w))| < c, \quad w \in \mathbb{W} \tag{6.13}$$

para as quais se vai introduzir uma generalização da definição de entropia dada pela expressão (6.11). Considerando-se sucessões de vectores $m-$dimensionais tal como se considerou na expressão 6.9:

$$c^*(w) = \{C(x_1, y_1, f(x_1, w)), \ldots, C(x_m, y_m, f(x_m, w))\}, \quad w \in \mathbb{W} \tag{6.14}$$

em que o conjunto de vectores $c^*(w)$ é induzido a partir da amostra e do conjunto de funções limitadas $C(x, y, f(x, w))$. No caso anterior considerou-se o conjunto de vectores binários $c(w)$, que tinham sido induzidos pelo conjunto de funções indicatrizes, uma vez que para a definição de entropia se pretendia que o conjunto de vectores $c(w)$ contivesse um número de elementos finito. Presentemente, o conjunto $c^*(w)$ contém um número infinito de elementos interiores a um cubo $m-$dimensional com arestas de comprimento $2c$, em que c é o limite definido na expressão (6.13), não tendo necessariamente que preencher todo o cubo.

Seja uma $\varepsilon-$rede mínima (para uma definição ver em Vapnik (1998), p.99) seja

o conjunto de vectores $c^*(w)$ cujo número dos seus elementos é dado por:

$$N^{\mathbb{W}}(\varepsilon,\ \{z_i\}_{i=1}^m)$$

isto é, este número depende de ε no conjunto de funções limitadas $C(x, y, f(x, w))$ e da amostra aleatória $\{z_i\}_{i=1}^m$. Usa-se este número nas definições dos conceitos de entropia aleatória e de entropia para um dado conjunto de funções reais. Supondo que para qualquer dimensão da amostra m a função $\ln N^{\mathbb{W}}(\varepsilon,\ \{z_i\}_{i=1}^m)$ é mensurável, então, define-se a $\varepsilon-$*entropia aleatória* do conjunto das funções uniformemente limitadas sobre uma conjunto amostral, pela quantidade dada pela expressão:

$$H^{\mathbb{W}}(\varepsilon,\ \{z_i\}_{i=1}^m) = \ln N^{\mathbb{W}}(\varepsilon,\ \{z_i\}_{i=1}^m)$$

Igualmente se define a $\varepsilon-$*entropia* do conjunto das funções uniformemente limitadas sobre uma conjunto amostral, pela quantidade dada pela expressão

$$H^{\mathbb{W}}(\varepsilon, m) = \int H^{\mathbb{W}}(\varepsilon,\ \{z_i\}_{i=1}^m)\, dP(\{z_i\}_{i=1}^m)$$

Para se verificar a convergência uniforme do funcional risco esperado para o funcional risco empírico, no caso de se ter um conjunto de funções uniformemente limitadas sobre uma conjunto amostral de tamanho m, é condição necessária e suficiente que, para qualquer $\varepsilon > 0$, se verifique:

$$\frac{H^{\mathbb{W}}(\varepsilon, m)}{m} \underset{m\to\infty}{\to} 0 \tag{6.15}$$

Para o caso da generalização a conjuntos de funções não limitadas pode ver-se Vapnik (1998), pp.101-105).

6.2.3 Entropia de Shannon

Mais de cinquenta anos estão passados desde que Claude Shannon formalizou os fundamentos da *teoria da informação*. Presentemente esta é uma das teorias que mais se tem enriquecido, particularmente devido às suas inúmeras aplicações na melhoria da eficiência e eficácia dos processos de transmissão e comunicação. Os dois principais temas focados nesta teoria são a eficiência da representação da informação e as limitações existentes na transmissão fiável da informação através de um canal de comunicação.

Um dos princípios básicos da teoria é o da *máxima informação*, devido a Linsker (1988). Este princípio também se pode aplicar na aprendizagem de uma rede neuronal, podendo então dizer-se que a adaptação dos pesos das conexões sinápticas de uma rede de perceptrão de multicamada (PMC) se processa de modo a maximizar a

quantidade de informação existente ao longo de todas as transformações sofridas pelos sinais nas diferentes fases de processamento, transformações que podem envolver ajustamentos ou restrições.

Uma interpretação deste princípio no contexto da percepção visual de cenas, referida por Haykin (1999) (p.484), diz uma das principais funções da maquinaria perceptual consiste em:

> *"Eliminar alguma redundância inerente à simulação através de uma descrição, ou codificação, de toda a informação recebida para uma forma mais económica do que a que afecta directamente os receptores, (in Attneabe, 1954)"*

Isto é, a identificação das características específicas de uma dada cena está relacionada com a forma como os dados de entrada são codificados de forma a minimizar o mais possível a redundância que neles possa existir.

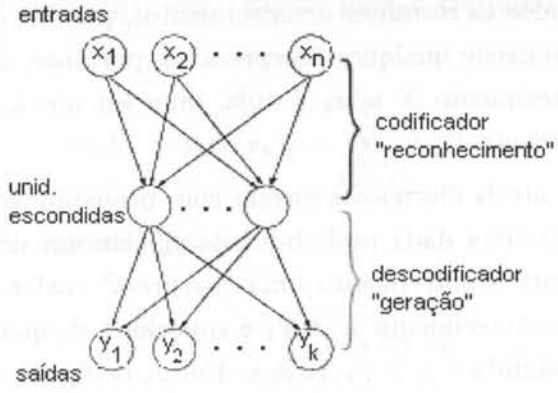


Figura 6.1: Codificador e descodificador

A Figura (6.1) mostra um esquema que ilustra a aplicação dos conceitos de codificação e descodificação da teoria da informação numa rede neuronal (Hinton and Ghahramani (2000)).

Como se afirmou anteriormente um dos conceitos mais utilizados para medir a quantidade de informação é o conceito de *entropia*. Considerando uma variável aleatória X, e supondo que cada realização (ou apresentação) desta variável constitui uma mensagem, em termos estritos, pode dizer-se que, se a amplitude da variável aleatória X for contínua, então, ela contém uma quantidade de informação infinita. Contudo, física e biologicamente reconhece-se não ter muito sentido uma medida de amplitude com precisão infinita, o que sugere que os valores possíveis de X só devem ser quantificados uniformemente num número finito de níveis discretos. Seja,

portanto, uma variável aleatória discreta X, modelada pelo seguinte conjunto de componentes x_k:

$$X = \{x_k |\ k = 0, \pm 1, ..., \pm K\}, \tag{6.16}$$

em que x_k é um número discreto, e $(2K+1)$ é o número total de níveis discretos que podem ocorrer.

Suponha-se, agora, que a separação δx entre dois níveis discretos sucessivos é suficientemente pequena para permitir ao modelo dado pela expressão (6.16) ter uma representação adequada da variável. Considerando a passagem para o limite, supondo a existência de continuidade, e deixando δx tender para zero, e K aproximar-se de infinito, obtém-se, assim, uma variável aleatória contínua uma vez que o somatório se transforma num integral.

Para completar o modelo, seja o acontecimento $X = x_k$ que ocorre com a probabilidade $p_k = P(X = x_k)$, sujeita à condição de se verificar $0 \leq p_k \leq 1$, e $\sum_{k=1}^{m} p_k = 1$.

Quando um acontecimento $X = x_k$ ocorre com uma probabilidade $p_k = 1$ e, por consequência, para todos os restantes acontecimentos, para os quais $i \leq k$, se sabe que $p_i = 0$, então, não existe qualquer "*surpresa*" e, portanto, a informação devida à ocorrência do acontecimento $X = x_k$ é nula, uma vez que a sua mensagem já é conhecida antecipadamente.

Quando os vários níveis discretos ocorrem com probabilidades diferentes e, em particular, se sabe que uma dada probabilidade p_k tem um valor muito pequeno, então, existirá para este acontecimento uma "surpresa" maior e, portanto, a "informação" devida ao acontecimento $X = x_k$ é superior à de qualquer outro acontecimento x_i, com probabilidade $p_i > p_k$, $i \neq k$. Então, os significados de "*surpresa*", "*incerteza*" e "*informação*" deverão estar relacionadas entre si. Existe uma certa quantidade de surpresa e incerteza antes de se verificar a ocorrência de um acontecimento $X = x_k$, mas, depois deste ter ocorrido, verifica-se que a quantidade de informação sofreu um acréscimo e, portanto, será de esperar que exista uma relação entre estas três quantidades.

A quantidade de informação está relacionada, como se acaba de ver, com o inverso da probabilidade da ocorrência de um dado acontecimento. Formalmente, diz-se que a quantidade de informação I que se ganhou pela ocorrência de um dado acontecimento $X = x_k$, de probabilidade p_k, é definida por:

$$I(x_k) = \log_2 \left(\frac{1}{p_k} \right) = -\log_2 p_k \tag{6.17}$$

A base dois dos logaritmos que é em geral utilizada permite exprimir os resultados em *bits* por acontecimento. Pode usar-se outra qualquer base arbitrária, por exemplo,

os logaritmos naturais, obtendo-se neste caso os resultados em *nats*.

A definição de informação dada pela expressão (6.17) tem as seguintes propriedades:

1. $I(x_k) = 0$, para $p_k = 1$. Obviamente que quando se tem a certeza absoluta da ocorrência de um acontecimento, então, não se obtém nenhuma informação com o facto de se ter a notícia da sua ocorrência.

2. $I(x_k) \geq 0$, para $0 \leq p_k \leq 1$. Isto é, após a ocorrência de um acontecimento $X = x_k$, das duas uma, ou ele forneceu alguma informação, ou não forneceu nenhuma informação, mas, o que nunca pode acontecer, é que haja perda de informação (ou informação negativa).

3. $I(x_k) > I(x_i)$ para $p_k < p_i$. Isto é, quanto menor for a probabilidade de um acontecimento tanto maior será a informação que se ganha com a ocorrência desse acontecimento.

Se a quantidade de informação $I(x_k)$ for uma variável aleatória discreta com probabilidade p_k, então, o seu valor médio para o conjunto dos $2K + 1$ valores discretos é dada por:

$$H(X) = \mathbf{E}\left[I(x_k)\right] = \sum_{k=-K}^{K} p_k I(x_k) \tag{6.18}$$

logo, a entropia de Shannon define-se:

$$H(X) = -\sum_{k=-K}^{K} p_k \log p_k \tag{6.19}$$

A entropia da variável aleatória X é dada por $H(X)$ e pode ter um conjunto de valores discretos que se exprimem em *bits* por mensagem, ou por acontecimento. A entropia $H(X)$ exprime o valor médio da quantidade de informação enviada pela mensagem, note-se, contudo, que a variável aleatória X, na expressão (6.19) de $H(X)$, não constitui o argumento da função mas apenas a designação da variável aleatória em causa. Notar também que na definição da expressão (6.19) se considera-se que $0log0 = 0$.

Suponha-se que o número total de níveis discretos é $2K + 1$, então, os valores da entropia $H(X)$ têm como limites:

$$0 \leq H(X) \leq \log_2(2K + 1) \tag{6.20}$$

A anulação da entropia verifica-se quando $p_k = 1$ para um dado k, e $p_k = 0$ para os restantes k, logo este limite inferior significa que não existem incertezas. O

valor máximo da entropia verifica-se quando $p_k = 1/(2K+1)$ para qualquer k, isto é, todos os níveis discretos são equiprováveis e, portanto, este limite superior corresponde à incerteza máxima. A demonstração do limite superior pode obter-se a partir de um *lema* (Gray (1990), p.25) sobre uma desigualdade a partir da qual muitas propriedades básicas da entropia se podem deduzir.

Lema 69 (Gray, 1990) *Dadas duas funções de massa de probabilidade $\{p_i\}$ e $\{q_i\}$, isto é, duas sucessões finitas e enumeráveis de números não-negativos cuja soma é unitária, então:*

$$\sum_i p_i \ln \frac{p_i}{q_i} \geq 0$$

verificando-se a igualdade se e só se $p_i = q_i$, para todo o i.

A razão usada neste lema tem tanta importância que sugere um novo conceito. Suponha-se que $p_X(x)$ e $q_X(x)$ são as funções de probabilidades da variável aleatória X estar sob a influência de duas condições (distribuições induzidas) diferentes. Define-se *entropia relativa (cross-entropy), entropia, ou divergência, de Kullback-Leibler,* entre as duas distribuições de probabilidade $p_X(x)$ e $q_X(x)$:

$$K(p,q) = \sum_{x \in \mathcal{H}} p_x \log \left(\frac{p_X(x)}{q_X(x)} \right) \tag{6.21}$$

em que o somatório se estende a todos os estados possíveis do sistema (i.é, ao alfabeto $\mathcal{H}$ da variável aleatória discreta X). A função $q_X(x)$ desempenha o papel de medida de referência. É interessante notar-se que a *entropia de Kullback-Leibler* pode ser representada no contexto da geometria diferencial como uma métrica de Riemann no espaço das distribuições Deco and Obradovic (1996), (p.10), segundo, não é uma verdadeira distância uma vez que não é simétrica, $K(p,q) \neq K(q,p)$, podendo-se dizer que é uma quase-distância.

Notar que a definição de entropia de Shannon, dada pela expressão (6.19), tem estado intrinsecamente ligada à estimação de funções de distribuição de probabilidade (fdp) de variáveis aleatórias. Todas as tentativas de utilizar esta definição de entropia pressupõem que se utiliza, ou um modelo analítico para a fdp, ou um modelo não paramétrico para a estimação da fdp. Por sua vez estes modelos têm um comportamento pobre em espaços de elevada dimensionalidade, razão pela qual se introduz, numa outra subsecção, uma definição alternativa de entropia, devida a Renyi, para a qual a entropia é função de uma norma da fdp em espaços de probabilidade, que é muito mais fácil de estimar no caso de os espaços serem de dimensão elevada (Principe et al. (2000b)).

6.2.4 Entropia de Kullback-Leibler diferencial

A definição de entropia de Shannon e entropia de Kullback-Leibler são extensíveis ao caso de variáveis aleatórias *contínuas*. Por exemplo, seja a variável X descrita pela sua função de densidade de probabilidade $f_X(x)$, e para um conjunto C que constitua o suporte da variável contínua x. Por analogia com o caso discreto pode escrever-se que a entropia da variável aleatória contínua X tem por expressão:

$$\begin{aligned} h(X) &= -\int_C f_X(x) \log f_X(x) dx \\ &= -\mathbf{E}\left[\log f_X(x)\right] \end{aligned} \tag{6.22}$$

Do critério da entropia pode obter-se o princípio da máxima entropia ("*MaxEnt*") enunciado por Jaynes (Jaynes (1994), capítulo 11), o qual é um dos mais potentes princípios de optimização tendo uma extensa aplicação em mecânica estatística, ciências físicas, economia e engenharia. O objectivo deste princípio consiste em maximizar a incerteza sobre tudo o que se desconhece acerca das restrições dum dado problema. Jaynes mostrou que a maioria das distribuições usadas na estatística podem ser deduzidas a partir do princípio *MaxEnt*.

A entropia pode também ser utilizada no processamento de sinal para ampliar várias metodologias há muito estabelecidas, e que se baseiam em momentos de segunda ordem, a variância e a correlação. Quando uma variável aleatória tem distribuição gaussiana os dois primeiros momentos nunca são nulos razão pela qual se adopta predominantemente a suposição gaussiana nos modelos de sistemas em engenharia, e tal se fez no modelo simples da distribuição dos pesos fetais descrito no Capítulo 4.

6.2.5 Informação mútua

Introduz-se a noção de informação mútua $I(X;Y)$ com o objectivo de medir a independência estatística entre duas variáveis aleatórias X e Y, definidas sobre os alfabetos discretos $\aleph$ e $\Im$, respectivamente, conhecendo as respectivas distribuições de probabilidades $p(x)$ e $p(y)$. A informação mútua $I(X;Y)$ utiliza a *entropia de Kullback-Leibler* entre a distribuição de probabilidade conjunta das duas variáveis e o produto das distribuições de probabilidade individuais, anulando-se no caso de X e Y serem independentes. Tem a expressão:

$$\begin{aligned} I(X;Y) &= K\left(p(x,y)\ ,\ p(x)\,p(y)\right) = \\ &= \sum_{x\in\aleph}\sum_{y\in\Im} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \end{aligned} \tag{6.23}$$

A informação mútua é simétrica, isto é, $I(X;Y) = I(Y;X)$. Obviamente que $I(X;X) = H(X)$, portanto, constitui uma medida de quantidade de informação que a variável aleatória Y exprime acerca de X (ou vice-versa). Por outras palavras $I(Y;X)$ dá uma medida da correlação estatística existente entre X e Y. Quando X e Y são definidas como a entrada e saída de um canal estocástico, respectivamente, então $I(X;Y)$ é uma medida da informação transmitida nesse canal.

Na aprendizagem (particularmente não supervisada) a avaliação da informação mútua tem mais utilidade do que da entropia, uma vez que a primeira envolve a estimação de uma distância entre funções de distribuição de probabilidade (fdp). A expressão 6.23 pode ser vista como uma distância assimétrica entre duas *fdps*, que é sempre positiva, anulando-se quando $p_X(x) = q_X(x)$. A divergência de Kulback-Leibler está no centro de um critério teorético informático enunciado por Kullback e que se designa por princípio da *mínima entropia cruzada* ("*MinxEnt*"). O objectivo deste critério consiste em determinar-se a distribuição de probabilidade que está mais se aproxima de uma dada distribuição. No caso especial de $p_X(x)$ ser a distribuição conjunta de duas variáveis X e Y , $p_X(x) = p_{XY}(x,y)$, e de $q_X(x)$ ser o produto das variáveis marginais correspondentes, $q_X(x) = p_X(x)\,p_Y(y)$,então, a divergência de Kulback-Leibler exprime a informação mútua entre as variáveis X e Y, isto é:

$$I(X;Y) = \iint p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)\,p_Y(y)} dxdy$$

Uma questão que se pode levantar consiste em saber-se em que medida o critério da informação mútua é apropriado para a aprendizagem. Começando por realçar o carácter inovador desta aproximação e a dificuldade existente em encontrar argumentos suficientes que justifiquem teoricamente a utilização deste critério, Principe et al. (2000b) consideram que existe uma fundamentação robusta para a utilização da teoria da informação, uma vez que esta é proveniente de duas teorias bem estabelecidas, a teoria da comunicação e a teoria da mecânica estatística. Contudo, estes autores reparam que estas teorias não contemplam aspectos básicos da teoria da aprendizagem tais como, por exemplo, a estimação e a inferência estatística Vapnik (1998).

Por exemplo estes autores salientam que em estimação paramétrica, sabe-se hoje em dia, que a variância de um qualquer estimador não enviesado está limitada inferiormente pelo limite de Cramer-Rao Deco and Obradovic (1996) (p.176). Seria igualmente importante saber-se em que medida é apropriado usar-se o critério da informação mútua para minimizar o erro de classificação de Bayes. Supondo que se pretende fazer a estimação da variável x, com uma fdp discreta $p(x)$, através do cálculo de uma estimativa $\hat{x}$ a partir de uma outra variável aleatória y caracterizada

por $p(x|y)$, sob condições moderadas, Fano demonstrou , que:

$$P(x \neq \hat{x}) \geq \frac{H_S(x|y) - 1}{\log(\Theta(x))} \tag{6.24}$$

em que $\Theta(x)$ significa o número de exemplos possíveis de x. A relação (6.24) mostra que a probabilidade de erro está limitada inferiormente pela entropia condicional de x dado y. Substituindo nesta relação a entropia condicional pela expressão com a informação mútua obtém-se:

$$P(x \neq \hat{x}) \geq \frac{H_S(x) - I(x,y) - 1}{\log(\Theta(x))} \tag{6.25}$$

Pode reparar-se que sendo difícil controlar-se a entropia e o número de exemplos é possível melhorar o limite inferior da probabilidade de erro através da maximização da informação mútua entre x e y. Uma vez que o objectivo é a minimização da probabilidade de erro, pode concluir-se que a relação (6.25) poderá não ser um resultado muito útil, uma vez que não nos dá um limite superior. Mas, tal como no limite de Cramer-Rao, a relação (6.25) mostra a existência de um limite inferior possível de ser atingido, enquanto que um limite superior dependeria muito em particular do estimador escolhido. Principe et al. (2000b) referem que é possível obter-se para o erro de Bayes um limite superior dependente da entropia condicional e, graças a todas estas subtis propriedades, estes autores interrogam-se porque razão a teoria da informação não tem tido uma aplicação mais alargada na aprendizagem automatizada, sugerindo que talvez uma parte das dificuldades resida na determinação da entropia e da informação mútua directamente a partir dos exemplos .

6.2.6 Aprendizagem teorético-informática

A aprendizagem teorético-informática (ATI) parece ser um processo natural de aprendizagem dos parâmetros de uma máquina de aprendizagem uma vez que a finalidade desta consiste em transferir a informação existente nos exemplos, a relação funcional entre as entradas e as etiquetas destes, para o sistema dos parâmetros adaptados. Uma definição de aprendizagem teorético-informática, proposta por Principe et al. (2000b), refere que esta consiste num procedimento que adapta os parâmetros livres w de uma máquina de aprendizagem $g(\cdot, w)$ com base num dado critério teorético-informático (6.2), que pode ser a *entropia* (maximização ou minimização), ou *a informação mútua* (maximização ou minimização) medidas no espaço de saídas do sistema de aprendizagem.

Uma vez que o argumento da função *entropia* depende do sistema de coordenadas utilizado na representação dos exemplos, a sua manipulação constitui um paradigma *da aprendizagem supervisada.* A separação às cegas das fontes de um sinal (" *blind*

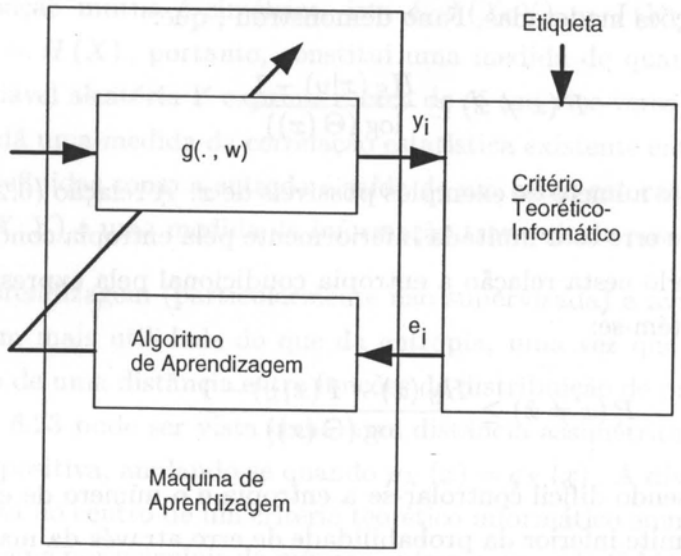


Figura 6.2: Treino de uma máquina de aprendizagem pelo processo de adaptação dos parâmetros livres w.

source separation") e a análise das componentes principais pode resolver-se por *maximização* da entropia, uma vez que, formalmente, esta é uma extensão da maximização da energia de saída de um sistema linear adaptativo utilizando o critério do erro quadrático médio. A previsão, redução da redundância e a discriminação de aglomerados ("*clustering*") pode resolver-se por *minimização* da entropia.

A medida de informação potencialmente mais útil em engenharia é a informação mútua sendo independente do sistema de coordenadas, uma vez que tem por base a estimação de uma medida de divergência entre as funções densidade de probabilidade de duas variáveis aleatórias. Consoante a natureza destas variáveis os critérios de informação mútua *podem ser aplicados a problemas de classificação supervisados e não supervisados*, tais como a análise dos componentes independentes ("ICA, *independent component analysis*") e a extracção de características Xu (1999).

A Figura (6.2) mostra esquematicamente o treino de uma máquina de aprendizagem pelo processo de adaptação dos parâmetros livres w da máquina, $g(\cdot, w)$, usando um critério teorético informático, que pode ser a maximização da informação mútua $I(X;Y)$ para problemas supervisados, a maximização (ou minimização) da entropia $h(X)$ no caso de problemas não supervisados. O objectivo consiste em encontrar-se um estimador não-paramétrico que tem por base os exemplos, e utiliza uma dada função de custo baseada na entropia de Shannon ou na divergência de Kullback-Leibler, para obter a maximização da transferência de informação das entradas estimador para os pesos w Principe et al. (2000b).

6.2.7 Entropia de Renyi

Definiu-se a entropia de Shannon pela expressão (6.19), no entanto, existe uma definição alternativa, devida ao matemático Húngaro Renyi (1970), tal como se referiu atrás. Na teoria geral das médias, a média $\bar{x}$ dos números reais $x_1 \ldots x_n$, com pesos $p_1 \ldots p_n$ tem a forma:

$$\bar{x} = \varphi^{-1}\left(\sum_{k=1}^{n} p_k \varphi(x_k)\right)$$

em que $\varphi(x)$ é a função de Kolmogorov-Nagumo, que por sua vez é uma função arbitrária contínua estritamente monotónica definida para números reais. Em geral, uma medida de entropia obedece à relação:

$$H = \varphi^{-1}\left(\sum_{k=1}^{n} p_k \varphi(I(p_k))\right) \tag{6.26}$$

em que $\varphi(I(p_k))$ é uma medida de informação, que não pode ser arbitrária uma vez que tem que satisfazer à condição de aditividade. Poderá ser $\varphi(x) = x$ e, neste caso, a expressão (6.26) transforma-se na entropia de Shannon. Alternativamente, poderá ser $\varphi(x) = 2^{(1-\alpha)x}$, para a qual a expressão (6.26) se transforma na medida de informação, $H_{R\alpha}$, designada entropia de Renyi de ordem α, que se exprime pela fórmula:

$$H_{R\alpha} = \frac{1}{1-\alpha} \log\left(\sum_{k=1}^{n} p_k^{\alpha}\right) \quad \alpha > 0, \quad \alpha \neq 1 \tag{6.27}$$

Quando $\alpha = 2$ a expressão (6.27) transforma-se em:

$$H_{R2} = \frac{1}{1-2} \log\left(\sum_{k=1}^{n} p_k^2\right) \tag{6.28}$$

que se designa *entropia quadrática*. Como em geral está-se interessado na determinação da entropia ou da informação mútua à saída dos sistemas dá-se preferência à designação da variável por Y . Dada uma variável aleatória contínua Y com função de distribuição de probabilidade $f_Y(y)$ pode obter-se uma versão diferencial da entropia de Renyi, e da entropia quadrática, seguindo uma via análoga à que se

utilizou para a entropia diferencial de Shannon, obtendo-se as expressões:

$$H_{R\alpha}(Y) = \frac{1}{1-\alpha} \log \left(\int_{-\infty}^{\infty} f_Y(y)^\alpha dy \right) \tag{6.29}$$

$$H_{R2}(Y) = -\log \left(\int_{-\infty}^{\infty} f_Y(y)^2 dy \right). \tag{6.30}$$

A entropia de Renyi tem um interesse particular na estimação uma vez que requerendo que se calcule o integral da função de distribuição elevado à potência α, permite tornar os cálculos mais simples do que a estimação pela fórmula da entropia de Shannon.

A entropia de Renyi traz uma perspectiva diferente para o problema da estimativa da entropia de uma distribuição de probabilidade $P = (p_1, p_2, ..., p_N)$ quando considerada como um ponto no espaço N-dimensional. Supondo que se verificam as condições canónicas de mensurabilidade, $p_k \geq 0$, $\sum_{k=1}^{n} p_k) = 1$, P reside sempre no primeiro quadrante de um sistema de coordenadas ortogonais de N dimensões, intersectando no ponto de coordenada $+1$ cada um dos eixos coordenados do sistema. Definindo a distância de P à origem do sistema, que se designa *α-norma da distribuição de probabilidade,* pela expressão:

$$\|P\|_\alpha = \sqrt[\alpha]{\sum_{k=1}^{n} p_k^\alpha} = \sqrt[\alpha]{V_\alpha}$$

pode exprimir-se a entropia de Renyi como uma função de V_α, através da expressão:

$$H_{R\alpha} = \frac{1}{1-\alpha} \log V_\alpha$$

Ao escolher-se na expressão da entropia de Renyi valores diferentes para α está-se de facto a escolher *α-normas* diferentes e, no caso limite, quando a *α-norma da distribuição de probabilidade* converge para um ($\alpha \to 1$), então, está-se a considerar a entropia de Shannon. Outros valores de α permitem medir a distância à origem de maneira diferente, tal como acontece na selecção de diferentes normas para o erro de um critério de aprendizagem. O valor $\alpha = 2$ além de parecer ser um bom compromisso para a simplicidade do modelo é também muito usado em dinâmica não linear Principe et al. (2000b). Os valores de α mais elevados apesar de permitirem aumentar a robustez da estimação em regiões com baixa densidade de amostragem, em contrapartida, fazem aumentar a complexidade algorítmica, a qual cresce exponencial mente com α.

6.2.8 Entropia quadrática e o seu estimador não paramétrico

A entropia quadrática de Renyi tem uma grande aplicação em virtude da facilidade com que se determina a 2-norma da *fdp* usando o conhecido estimador da *janela de Parzen* (Duda et al. (2001), pp.164-174). Seja $y_i \in R^k$, $i = 1 \ldots N$, um conjunto de exemplos extraídos de uma variável aleatória, num espaço k-dimensional, que podem ser os valores de uma função não linear multivariável, tal como, por exemplo, a que se obtém na saída de um perceptrão PMC. Pode estimar-se a *2-norma da distribuição de probabilidade* deste conjunto de exemplos y_i usando um *kernel* gaussiano na janela de Parzen:

$$f_Y(y) = \frac{1}{N} \sum_{i=1}^{N} G\left(\mathbf{y} - \mathbf{y}_i, \sigma^2\right). \tag{6.31}$$

$G(\cdot, \cdot)$ é um kernel gaussiano num espaço k-dimensional, e em que o segundo parâmetro, a matriz de covariância Σ, por simplicidade se escolheu um *kernel* com simetria esférica, para o qual se verifica $\Sigma = \sigma^2 I$:

$$G\left(\mathbf{y} - \mathbf{y}_i, \sigma^2\right) = \frac{1}{(2\pi)^{k/2}\, \sigma^k} \exp\left(-\frac{(\mathbf{y} - \mathbf{y}_i)^T (\mathbf{y} - \mathbf{y}_i)}{2\sigma^2}\right) \tag{6.32}$$

Substituindo a *fdp* $f_Y(\mathbf{y})$ da expressão (6.30), da entropia de Renyi para $\alpha = 2$, pela expressão dada na janela gaussiana (6.31) obtém-se:

$$\begin{aligned}
H_{R2}(\{\mathbf{y}_i\}) &= -\log\left(\int_{-\infty}^{\infty} f_Y(\mathbf{y})^2 dy\right) \\
&= -\log\left(\int_{-\infty}^{\infty} \frac{1}{N} \sum_{i=1}^{N} G\left(\mathbf{y} - \mathbf{y}_i, \sigma^2\right) \frac{1}{N} \sum_{j=1}^{N} G\left(\mathbf{y} - \mathbf{y}_j, \sigma^2\right) dy\right) \\
&= -\log\left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{-\infty}^{\infty} G\left(\mathbf{y} - \mathbf{y}_i, \sigma^2\right) G\left(\mathbf{y} - \mathbf{y}_j, \sigma^2\right) dy\right) \\
&= -\log\left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G\left(\left(\mathbf{y}_i - \mathbf{y}_j\right), \sigma^2\right)\right)
\end{aligned} \tag{6.33}$$

Introduzindo uma grandeza nova, $V(\{\mathbf{y}_i\})$, dada pela expressão dentro do parênteses da função logaritmo da entropia quadrática (6.33):

$$V(\{\mathbf{y}_i\}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G\left(\left(\mathbf{y}_i - \mathbf{y}_j\right), \sigma^2\right) \tag{6.34}$$

obtém-se um resultado de mais simples determinação, no qual se procede ao cálculo

dos valores do *kernel* nos pontos $\mathbf{y}_i - \mathbf{y}_j$ em substituição do cálculo do integral. Assim, passa-se a determinar os valores de $V(\{\mathbf{y}_i\})$ em vez dos valores da entropia $H(\{y_i\})$, os quais são mais difíceis de determinar e dispensáveis e uma vez que o logaritmo não afecta o processo de optimização. A precisão deste estimador da entropia quadrática está dependente da aproximação conseguida na estimação da função densidade de probabilidade, *fdp*. Principe et al. (2000b) sugerem que considerando uma possível analogia entre as "*amostras de dados*" e as "*partículas físicas*", faz sentido interpretar-se $V(\{\mathbf{y}_i\})$ como uma energia potencial num campo potencial dado pela amostra de dados. Propõem a designação de *potencial informativo* para esta energia potencial, logo a maximização da entropia quadrática de Renyi é equivalente à minimização do potencial informativo. Estendendo a analogia mecânica à derivada da energia informativa, obtém-se uma *força informativa* que "desloca" os exemplos no espaço das saídas de forma a obter-se um estado de equilíbrio o qual fica determinado pelo critério escolhido. Assim, pode considerar-se que a *força* que o exemplo $\mathbf{y}_j$ transmite ao exemplo $\mathbf{y}_i$ é dada pela expressão:

$$\frac{\partial}{\partial \mathbf{y}_i} G\left(\left(\mathbf{y}_i - \mathbf{y}_j\right), \sigma^2\right) = \frac{G\left(\left(\mathbf{y}_i - \mathbf{y}_j\right), \sigma^2\right)\left(\mathbf{y}_i - \mathbf{y}_j\right)}{2\sigma^2}.$$

Se qualquer exemplo for livre de se deslocar numa certa região do espaço, então, as forças informativas forçarão que os exemplos atinjam todos um estado com potencial informativo *mínimo*.

Supondo que os exemplos são as saídas de um sistema adaptativo, como por exemplo um perceptrão PMC, e que se pretende adaptar esse PMC através de um algoritmo iterativo tal que seja maximizada a entropia de saída $H_{R2}(\{\mathbf{y}_i\})$, o problema é equivalente a determinar-se os parâmetros do PMC que minimizam o potencial informativo $V(\{\mathbf{y}_i\})$. Por consequência, a força informativa aplicada a cada exemplo pode ser rectro-propagada para os parâmetros utilizando a conhecida regra da função composta ("*chain rule*"). A título de exemplo, o seguinte gradiente pode ser interpretado como uma força de rectropropagação:

$$\frac{\partial}{\partial w_{ij}} V(\{y(n)\}) = \sum_{n=1}^{N} \sum_{p=1}^{k} \frac{\partial V(\{y(n)\})}{\partial y_p(n)} \frac{\partial y_p(n)}{\partial w_{ij}}$$

em que $y(n) = [y_1(n), \ldots, y_k(n)]^T$, e w_{ij} é um dos pesos no PMC. A maximização da entropia equivale à minimização do potencial informativo. Notar que o perceptrão PMC está a ser adaptado num modo não-supervisado, com uma propriedade da sua saída. A função de custo é totalmente independente da função a estimar, portanto, pode ser aplicada em qualquer caso. A complexidade deste algoritmo é $O(N^2)$.

A força informativa a que estão sujeitos os exemplos $y(n)$ é dada pela expressão:

$$\frac{\partial}{\partial y(n)} V(\{y(n)\}) = \left[\frac{\partial V(\{y(n)\})}{\partial y_1(n)}, \ldots, \frac{\partial V(\{y(n)\})}{\partial y_k(n)}\right]^T.$$

Pode reparar-se que a sensibilidade da saída $\frac{\partial y_p(n)}{\partial w_{ij}}$, com respeito aos parâmetros do PMC, constitui o mecanismo de transmissão através do qual as forças informativas sofrem uma rectropropagação para o parâmetro. Principe et al. (2000b) sugerem que esta analogia com o formalismo da rectropropagação permite concluir que as forças informativas desempenham um papel semelhante ao de erros injectados, e que a manipulação da entropia origina uma metodologia geral, não paramétrica, baseada em exemplos, com a qual é possível fazer a adaptação de um sistema não linear arbitrário desde que este tenha não linearidades suaves.

A informação mútua entre duas variáveis aleatórias X e Y, cujas funções densidade de probabilidade marginais são $f_X(x)$ e $f_Y(y)$, respectivamente, e a função densidade de probabilidade conjunta é $f_{XY}(x,y)$, pode determinar-se usando a divergência de Kullback-Leibler (6.21) entre a função de probabilidade conjunta e o produto das funções de probabilidade marginais. Porém, surge o problema de a divergência de Kullback-Leibler em espaços de dimensão elevada ser de muito difícil estimação não-parametricamente. Para resolver este problema pode utilizar-se na aprendizagem outras medidas alternativas para a divergência que sejam teoricamente consistentes com a de Kullback-Leibler. Isto é, medidas de divergência que conduzam a valores extremos (i.é., mínimos ou máximos), quer da entropia, quer da informação mútua, iguais aos que se poderiam obter pela determinação da divergência de Kullback-Leibler. Tendo em consideração as restrições quadráticas ($\alpha = 2$) e baseando-se na desigualdade de Cauchy-Schwartz, Principe et al. (2000b) propõe a medida I_{CS} para a estimativa da *informação mútua quadrática* entre as variáveis aleatórias X e Y :

$$I_{CS}(X,Y) = \log \frac{\left(\iint f_{XY}(x,y)^2\, dxdy\right)\left(\iint f_X(x)^2 f_Y(y)^2\, dxdy\right)}{\left(\iint f_{XY}(x,y) f_X(x) f_Y(y)\, dxdy\right)^2}$$

Em alternativa a esta medida estes autores também desenvolveram uma outra expressão para $I_{CS}(X,Y)$ baseada na distância euclidiana.

6.3 Estimação de uma fdp pelo método das janelas de Parzen

A determinação de uma *função densidade de probabilidade* (fdp) cuja forma é desconhecida pode ser obtida através de técnicas não paramétricas. A ideia básica assenta

na suposição de que a probabilidade P de um vector $\mathbf{x}$ cair numa dada região R pode exprimir-se pela expressão:

$$P = \int_R p(\mathbf{x})\, d\mathbf{x}$$

a qual permite obter-se uma versão mais suavizada da função densidade $p(\mathbf{x})$. Extraindo-se n amostras $x_1 \ldots x_n$, (i.i.d.) de uma população que obedeça à densidade $p(\mathbf{x})$ a probabilidade de que k destas n amostras caiam na região R, supondo $p(\mathbf{x})$ dada por uma lei binomial, tem um valor esperado $\mathbf{E}(k) = nP$. Supondo que esta distribuição de k é muito apertada na vizinhança do valor médio, pode esperar-se que a razão k/n seja uma boa estimativa da probabilidade P e, por consequência, da função de densidade suavizada Duda et al. (2001) (p.162). Esta estimativa é particularmente precisa quando o valor de n é muito elevado. Se, agora, se admitir que $p(\mathbf{x})$ é contínua, e que a região R tem um volume V e é suficientemente pequena de tal modo a que $p(\mathbf{x})$ não tenha variações apreciáveis, então, num ponto $\mathbf{x}$ interior a R tem-se que:

$$P = \int_R p(\mathbf{x})\, d\mathbf{x} \simeq p(\mathbf{x})\, V \tag{6.35}$$

e por consequência tem-se que:

$$p(\mathbf{x}) \simeq \frac{k/n}{V} \tag{6.36}$$

Do ponto de vista prático o número de amostras n é sempre limitado e o volume V não pode tornar-se tão pequeno quanto se desejar. Por esta razão tem de se aceitar uma margem de variação quer no valor da variância da razão k/n, quer no valor de variação da média de $p(\mathbf{x})$.

O método de Parzen determina o valor de k_n, correspondente ao número de amostras n que caem na região R_n, usando-se uma definição para o volume que varia inversamente com o número n de amostras, $V_n = 1/\sqrt{n}$.

Suponha-se que se considera uma região R_n definida por um hipercubo d-dimensional, de volume $V_n = (h_n)^d$, em que h_n é o comprimento da aresta Marques_de_Sá (2001a) (p.110). Pode ter-se uma expressão analítica para k_n, definindo-se um hipercubo unitário centrado na origem através de uma função janela $\varphi(\mathbf{u})$:

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq \frac{1}{2}, \quad j = 1, \ldots d \\ 0 & \textit{caso contrário} \end{cases} \tag{6.37}$$

Isto é, $\varphi\left(\left(\mathbf{x} - \mathbf{x}_i\right)/h_n\right)$ tem um valor unitário se $\mathbf{x}_i$ pertencer ao interior do hipercubo de volume V_n, centrado em $\mathbf{x}$, e zero no caso contrário. O número de amostras neste hipercubo é, portanto, dado por $k_n = \sum_{i=1}^{n} \varphi\left(\left(\mathbf{x} - \mathbf{x}_i\right)/h_n\right)$ e a n-ésima estimativa

de:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\varphi\left(\left(\mathbf{x}-\mathbf{x}_i\right)/h_n\right) \tag{6.38}$$

A expressão 6.38 sugere uma aproximação mais geral da estimativa da função densidade, na qual em lugar de se estar limitado a uma função janela hipercúbica, dada pela expressão 6.37, se pode permitir uma classe mais geral de funções janela. Neste caso a expressão 6.38 exprime uma estimativa de $p(\mathbf{x})$ como uma média de funções de $\mathbf{x}$ e de amostras $\mathbf{x}_i$. No essencial a função janela está a ser usada para interpolação – cada amostra contribui para a estimativa tendo em conta a sua distância de $\mathbf{x}$.

Para a estimativa $p_n(\mathbf{x})$ ser uma função de densidade legítima tem que se verificar $\varphi(\mathbf{x}) \geq 0$ e $\int \varphi(\mathbf{u})\, d\mathbf{u} = 1$ a qual é satisfeita automaticamente quando se verifica a relação $V_n = (h_n)^d$.

Para se examinar o efeito da largura da janela h_n em $p_n(\mathbf{x})$ define-se a função:

$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right)$$

e, a seguir pode escrever-se $p_n(\mathbf{x})$ como a média:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n\left(\mathbf{x}-\mathbf{x}_i\right).$$

Quer a amplitude, quer a largura de $\delta_n(\mathbf{x})$ são afectados pelo valor de h_n uma vez que $V_n = h_n^d$. Quando o valor h_n é muito grande, então, a amplitude de δ_n é muito pequena e $\mathbf{x}$ pode estar longe de $\mathbf{x}_i$ antes que $\delta_n(\mathbf{x}-\mathbf{x}_i)$ sofra uma variação muito pequena de $\delta_n(\mathbf{0})$. Neste caso, $p_n(\mathbf{x})$ é uma sobreposição de n funções amplas e lentamente variáveis e torna-se uma estimativa de $p(\mathbf{x})$ muito suave e desfocada. Por outro lado quando h_n tem um valor muito pequeno, então, o valor máximo de $\delta_n(\mathbf{x}-\mathbf{x}_i)$ torna-se amplo e ocorre na vizinhança de $\mathbf{x} = \mathbf{x}_i$. Neste caso $p(\mathbf{x})$ torna-se a sobreposição de n impulsos muito abruptos centrados nos exemplos – obtendo-se assim uma estimativa errática e ruidosa. Como a distribuição é normalizada, qualquer que seja o valor de h_n :

$$\int \delta_n\left(\mathbf{x}-\mathbf{x}_i\right) d\mathbf{x} = \int \frac{1}{V_n}\varphi\left(\frac{\mathbf{x}}{h_n}\right) d\mathbf{x} = \int \varphi(\mathbf{u})\, d\mathbf{u} = 1$$

então, quando h_n converge para zero, $\delta_n(\mathbf{x}-\mathbf{x}_i)$ aproxima-se da função delta de Dirac centrada em $\mathbf{x}_i$, e $p_n(\mathbf{x})$ aproxima-se de uma sobreposição de funções delta centradas nas amostras.

A escolha adequada de h_n, ou de V_n, tem um efeito importante em $p_n(\mathbf{x})$. Se V_n for muito grande a estimativa terá uma resolução muito baixa; Se V_n for muito

pequeno, a estimativa terá uma variabilidade estatística muito grande. Com um número de amostras muito limitado o melhor que se pode conseguir é um compromisso aceitável. Adicionalmente existem condições de bom comportamento a impor-se a $\varphi(\cdot)$ para que se verifique a convergência de $p_n(\mathbf{x})$ para $p(\mathbf{x})$ Duda et al. (2001) (pp.166–8).

6.4 Experiências com redes *PMC* estimadoras do peso fetal treinadas por um algoritmo informático-teorético

6.4.1 Estudo do erro de treino e generalização

O objectivo desta investigação consiste em saber até que ponto a aprendizagem baseada num critério teorético-informático permite uma redução dos valores de erro de previsão dos pesos fetais obtidos com redes neuronais *PMC* treinadas com o critério de minimização do erro médio quadrático Sereno et al. (2002b).

No capítulo anterior descreveram-se experiências e resultados que mostram que o erro médio das redes neuronais *PMC* e *FBR*, treinadas e testadas por validação cruzada, usando subconjuntos do conjunto inicial de 414 exemplos de pesos fetais referido no Capítulo 4, permitiam obter erros relativos da ordem dos 6.5%. Estes valores constituem um melhoramento em relação aos valores dados pelos modelos de regressão linear generalizada, descritos no Capítulo 4, com os quais se obtinham erros relativos da ordem dos 7.5%. Da mesma forma que se fizeram tentativas de melhoramento destes resultados utilizando *MVS*, que se descrevem no Capítulo 5, também neste capítulo se apresentam os resultados de algumas experiências com os métodos baseados na maximização do potencial informativo.

A figura 6.3 representa alguns resultados do treino de um *PMC* com 3 unidades escondidas, treinado com 158 exemplos, usando-se apenas três entradas ecográficas, o perímetro abdominal, o comprimento do fémur e o diâmetro biparietal.

O gráfico 6.3 (1), representado na primeira linha e primeira coluna desta figura, mostra a evolução ao longo de 30 iterações do potencial informativo. O seu valor máximo corresponde ao valor mínimo da entropia do erro (Erdogmus and Principe (2001)).

O gráfico 6.3 (2) mostra a evolução da abertura σ da janela gaussiana usada no método de Parzen para a determinação da fdp do erro, começando com $\sigma = 1.0$ e decrescendo ao longo das iterações.

O gráfico 6.3 (3) mostra alguns valores estimados do peso fetal (cruzes) *versus* os correspondentes valores observados de peso fetal (círculos). Notar que os casos, representados em abcissa, estão ordenados por ordem crescente dos pesos fetais observados.

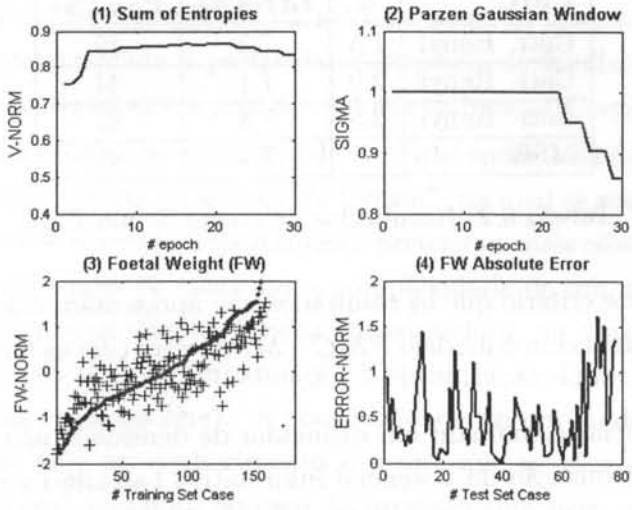


Figura 6.3: Resultados de um PMC treinado por minimização do potencial informativo ($\alpha = 1.5$).

O gráfico 6.3 (4) mostra os valores absolutos do erro de estimativa num conjunto disjunto de casos de teste constituído por 78 exemplos.

Utilizaram-se três valores distintos para o parâmetro α na expressão da entropia de Renyi, tendo-se obtido o erro mais baixo para o caso de $\alpha = 2.0$, que corresponde à entropia quadrática. Tal como se pode ver na figura 6.3 (1) o número de iterações parece ter sido excessivo uma vez que o treino foi prolongado para além do valor máximo de potencial informativo V. A entropia (ou o potencial informativo), tal como qualquer função de custo, pode ter mínimos (máximos) locais, podendo os resultados que se apresentam ser relativos a um desses valores óptimos locais. Para obstar a este efeito o algoritmo de aprendizagem reduziu a largura da janela de Parzen.

A tabela 6.2 mostra, para uma mesma arquitectura de PMC, alguns resultados obtidos em quatro modelos de redes PMC. Os três primeiros modelos foram optimizados por um critério baseado na entropia de Renyi. Na última linha desta tabela apresentam-se resultados obtidos como o critério do erro médio quadrático (MSE). Os resultados representados na terceira coluna (**Erro**(%)) são erros relativos, verificando-se que os valores obtidos são sensivelmente inferiores nos três modelos baseados num critério teorético informático, comparativamente com os obtidos quer com perceptrões PMC treinados por um critério MSE, quer com as fórmulas de regressão de Hadlock (7.8%). Os resultados representados na quarta coluna ($\mathbf{P}_{0.05}(\%)$) são valores da percentagem de casos cujo erro é inferior a $0.05 \times PFE$, isto é, dos casos cujos erros absolutos são inferiores a 5% do pesos fetais observados.

PMC	α	Erro(%)	$P_{0.05}$(%)
Entr. Renyi	1.5	7.3	39
Entr. Renyi	2.0	7.1	44
Entr. Renyi	2.5	7.3	42
MSE		7.5	41

Tabela 6.2: Resultados de treino de um PMC

Verificando-se neste critério que os resultados não apresentam diferenças significativas nem entre si, nem com o modelo $PMC-MSE$, nem com as fórmulas de Hadlock (41%).

Estas experiências combinam um estimador de densidade não paramétrico com um critério de optimização do potencial informativo baseado na entropia de Renyi permitiram a obtenção de um estimador do peso fetal cujos resultados por serem ligeiramente melhores dos que foram descritos anteriormente, nas experiências referidas no Capítulo 5, e em que se utilizaram $PMCs$ com topologias similares e se usava a rectropropagação do erro médio quadrático, sugerem que esta aproximação quando devidamente implementada pode permitir obter-se resultados interessantes.

6.4.2 Comparação de duas redes PMC

Os resultados que se vão referir a seguir com redes PMC constituem um primeiro estudo comparativo entre redes treinadas pelo método neuronal clássico, baseado no critério de erro dado pela norma L_2, e redes treinadas pelo método baseado no critério teorético-informático Sereno et al. (2002a). Para além de serem apresentados resultados de experiências com um algoritmo de treino minimizador da entropia do erro, apresenta-se um estudo do problema da selecção do modelo, procurando-se determinar em que medida um aumento da complexidade, para o conjunto de treino existente, permite obter uma melhor generalização.

Em secções anteriores foi referido que podem observar-se desvantagens na utilização no treino pelo critério de erro quadrado mínimo dado pela norma L_2 (*minimum square error*, MSE, para uma abordagem mais desenvolvida pode ver-se em Haykin (1999), Principe et al. (2000b) ou Marques_de_Sá (2001a)). Recorde-se, por exemplo, desvantagens devidas à: (*i*) sensibilidade significativa em relação a *outliers* e a grandes desvios verificados na regra de actualização dos valores dos parâmetros (pesos) ; (*ii*) dependência das estatísticas de segunda ordem que se verifica na função de densidade de probabilidade do erro.

Os critérios entrópicos têm vantagens, não só em problemas de classificação nos quais as saídas se podem interpretar como probabilidades *a posteriori*, mas mais importante ainda, a entropia é dependente de toda a fdp e não apenas dos valores da média e variância desta Erdogmus and Principe (2001). O critério de performance

PMC	α	Erro(%)	$P_{0.05}$(%)
Entr. Renyi	1.5	7.3	39
Entr. Renyi	2.0	7.1	44
Entr. Renyi	2.5	7.3	42
MSE		7.5	41

Tabela 6.2: Resultados de treino de um PMC

Verificando-se neste critério que os resultados não apresentam diferenças significativas nem entre si, nem com o modelo $PMC-MSE$, nem com as fórmulas de Hadlock (41%).

Estas experiências combinam um estimador de densidade não paramétrico com um critério de optimização do potencial informativo baseado na entropia de Renyi permitiram a obtenção de um estimador do peso fetal cujos resultados por serem ligeiramente melhores dos que foram descritos anteriormente, nas experiências referidas no Capítulo 5, e em que se utilizaram $PMCs$ com topologias similares e se usava a rectropropagação do erro médio quadrático, sugerem que esta aproximação quando devidamente implementada pode permitir obter-se resultados interessantes.

6.4.2 Comparação de duas redes PMC

Os resultados que se vão referir a seguir com redes PMC constituem um primeiro estudo comparativo entre redes treinadas pelo método neuronal clássico, baseado no critério de erro dado pela norma L_2, e redes treinadas pelo método baseado no critério teorético-informático Sereno et al. (2002a). Para além de serem apresentados resultados de experiências com um algoritmo de treino minimizador da entropia do erro, apresenta-se um estudo do problema da selecção do modelo, procurando-se determinar em que medida um aumento da complexidade, para o conjunto de treino existente, permite obter uma melhor generalização.

Em secções anteriores foi referido que podem observar-se desvantagens na utilização no treino pelo critério de erro quadrado mínimo dado pela norma L_2 (*minimum square error*, MSE, para uma abordagem mais desenvolvida pode ver-se em Haykin (1999), Principe et al. (2000b) ou Marques_de_Sá (2001a)). Recorde-se, por exemplo, desvantagens devidas à: (*i*) sensibilidade significativa em relação a *outliers* e a grandes desvios verificados na regra de actualização dos valores dos parâmetros (pesos) ; (*ii*) dependência das estatísticas de segunda ordem que se verifica na função de densidade de probabilidade do erro.

Os critérios entrópicos têm vantagens, não só em problemas de classificação nos quais as saídas se podem interpretar como probabilidades *a posteriori*, mas mais importante ainda, a entropia é dependente de toda a fdp e não apenas dos valores da média e variância desta Erdogmus and Principe (2001). O critério de performance

entrópico pode, pois, ser baseado numa medida da própria entropia, ou então na entropia cruzada determinada a partir da divergência de Kullback-Leibler entre a *fdp* dos valores das etiquetas e a *fdp* dos valores estimados Cover (1991).

O problema da selecção do modelo pode ser visto como uma formalização da conhecida metáfora da "*lâmina de barbear de Occam*", na qual se realça a importância em dar preferência aos modelos mais simples e preterir os mais complexos. Portanto, deve existir a preocupação de controlar a complexidade de um modelo no sentido de se obter sempre a mais simples função aproximadora, ou o modelo indutivo (e, inclusivamente, a mais simples aproximação à optimização do processo adaptativo, particularmente em redes *on line*). A complexidade de um modelo pode também ser controlada com base no conhecimento *a priori* que eventualmente possa existir sobre o modelo real, contudo, no caso da previsão com base num número finito de exemplos, a precisão é determinada principalmente pela quantidade e qualidade dos exemplos de treino. O conhecimento *a priori* pode ser útil apenas em modelos preditivos se ele permitir exercer um controlo explícita, ou implicitamente, sobre a complexidade do modelo. Deve dar-se preferência aos métodos que originam um controlo explícito da complexidade do modelo, seja porque permitem medir adequadamente a sua complexidade, seja porque permitem obter uma boa parametrização duma família de funções aproximadoras Cherkassky and Mulier (1998).

Estima-se a *fdp* dos erros pelo método de Parzen utilizando uma janela $\varphi(\cdot)$ Gaussiana Marques_de_Sá (2001a):

$$f_Y(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} G\left(\mathbf{y} - x_i, \sigma^2\right) \tag{6.39}$$

Tal como se referiu na subsecção 6.2.8, e para mais detalhes pode ver-ser, por exemplo, Principe et al. (2000a), no caso simples em que $\alpha = 2$ a entropia pode exprimir-se por:

$$H_2(\{\mathbf{x}_i\}) = -\log V(\{\mathbf{x}_i\}) \tag{6.40}$$

em que o potencial informativo V é dado pela expressão 6.34 sendo optimizado pelo algoritmo de aprendizagem, isto é, optimiza-se a informação mútua entre as entradas e as saídas da rede Principe et al. (2000b).

A aproximação utilizada recorre a uma validação cruzada do tipo "*deixar-um-de-fora*" sobre dois conjuntos um de tamanho 60, e o outro100 casos, extraídos aleatoriamente e *iid* de uma população de 236 casos, considerados apropriados para o treino de redes. Utilizaram-se quatro medidas ecográficas como entradas, uma camada escondida e duas saídas (peso e comprimento fetal). O risco empírico considerado é a média da soma dos quadrados destas duas saídas.

$N_{u.esc}$	2	5
Entropico	0.295	0.204
Standard	0.048	0.014

Tabela 6.3: Riscos previstos com PMC

Utilizaram-se duas arquitecturas de modelos, uma com uma complexidade de 16 parâmetros ($N_{u.esc} = 2$ unidades escondidas) e outra de 37 parâmetros ($N_{u.esc} = 5$ unidades escondidas). As médias dos riscos estimados que foram obtidas com uma amostra de 100 casos e optimizações *MSE* e *entrópica* estão representadas na tabela 6.3. Os *gráficos-em-caixa* da figura 6.4 mostram com mais detalhe as distribuições dos riscos estimados.

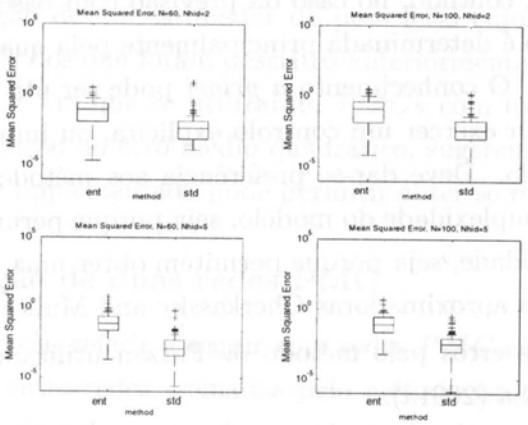


Figura 6.4: Risco esperado, PMC: "ent"significa que foi treinado com optimização entrópica e "std"com optimização MSE

O algoritmo entrópico de optimização do erro que se utilizou foi originalmente concebido por Erdogmus and Principe (2001) tendo sido ligeiramente adaptado ao problema de previsão do peso fetal. Como era de esperar o esforço computacional necessário na utilização deste algoritmo foi muito superior ao necessário pelo o algoritmo estandardizado, MSE. Uma possível explicação para o facto de se terem obtido valores de erro relativo superiores no método entrópico pode estar num treino insuficiente. O número de iterações utilizadas no treino, 10 nesta experiência, é reduzido porque o método de deixar-um-de-fora exige um esforço computacional muito grande. Uma outra razão pode ter a ver com o facto de os métodos terem sido comparados pelos valores médios quadráticos dos erros respectivos, em conformidade com o que está sugerido em Cherkassky and Mulier (1998) (pp.119-127). Na opinião de Sereno (2002) deve ser considerado um outro método de comparação de modelos que não favoreça o treino por MSE e apresente um viés que valoriza este

método em relação ao entrópico.

Finalmente, vai referir-se alguns aspecto que foram considerados com respeito ao controlo da complexidade do modelo.

Segundo Cherkassky et al. (1999), deve ter-se em linha de conta os seguintes três aspectos:

- Índice de complexidade que pode ser dado pelo número de parâmetros livres significativos;
- Método de estimação do risco esperado ;
- Quando o método de estimação do risco esperado é uma estimativa do risco empírico esta não ser um mínimo local.

Utilizou-se como índice de complexidade dos modelos o número de pesos das conexões da rede; na determinação do risco esperado utilizou-se um conjunto de teste disjunto do conjunto de treino da rede, e procedeu-se a um número de iterações de treino que teve em consideração as experiências anteriores, de forma a não criar nem um sobre ajustamento aos dados, nem cair-se num mínimo local.

A aproximação por reamostragem utilizada na estimação do risco esperado tem a vantagem de não estar dependente quer de pressupostos sobre as estatísticas dos exemplos, quer de propriedades específicas das funções aproximadoras. A sua principal desvantagem reside no esforço computacional necessário, e na variabilidade das estimativas a qual pode advir da maneira como se fixa o tamanho do conjunto de treino.

Existe também a possibilidade de comparação entre os resultados empíricos com os resultados que se podem obter analiticamente, com base na teoria da aprendizagem estatística sendo conhecida uma estimativa da dimensão VC, d_{VC}, do conjunto de funções de custo. Segundo Cherkassky et al. (1999) existe um limite superior que com um nível de confiança de $1-\eta$ é dado pela fórmula:

$$R^*_{emp} \leq \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \cdot \tag{6.41}$$

$$\left(1 - c\sqrt{a_1 \frac{d_{VC}\left(\ln\left(\frac{a_2 N}{d_{VC}}\right) - \ln\left(\frac{\eta}{4}\right)\right)}{N}}\right)^{-1}_{+} \tag{6.42}$$

em que c é uma constante que reflecte as extremidades da função e distribuição dos custos e a_1 e a_2 são constantes teóricas e em que $u_+ = \max(0, u)$ Cherkassky and Mulier (1998).

O método de optimização informático-teorético alia uma técnica de estimação não-paramétrica com um método de adaptação de pesos baseado num critério de maximização do potencial informativo, baseado na entropia de Renyi. O algoritmo parece dar resultados robustos no caso da estimação do peso fetal.

6.4.3 Combinação de modelos usando critérios de informação mútua

Descrevem-se agora as experiências em que se utilizam critérios teorético-informáticos para adaptar os pesos de um conjunto de redes neuronais estimadoras do peso fetal.

Está-se interessado em investigar como se poderiam treinar *comités* constituídos por três tipos de redes estimadoras de pesos fetais, *PMC*, *FBR* e *MVS*, treinados com o mesmo conjunto de treino.

O interesse por este tipo de comités reside no facto de cada saída da rede neuronal poder ser interpretada como um voto e, por consequência, na perspectiva da regressão poder dizer-se que se utiliza a média dos valores produzidos por cada rede, de forma a decrementar a variância da decisão final.

Realizaram-se experiências (muito preliminares) em que se treinou uma rede com um perceptrão simples de três entradas. Cada entrada recebia a saída das três redes treinadas no seu valor óptimo de minimização do erro e maximização da capacidade de generalização.

A figura 6.5 mostra algumas observações relativas ao estudo da performance de um perceptrão simples treinado por um critério informático-entrópico, optimizador da entropia quadrática do erro de treino, obtido com 158 exemplos de treino e uma aprendizagem durante 30 épocas. Obteve-se como pesos, ou votos, para cada uma das entradas, os valores 0.3393, 0.3417 e 0.3311, os quais apresentam ligeiras diferenças em relação aos valores 0.3333 que se utilizariam numa média igualmente pesada.

Cada uma das três entradas deste perceptrão recebe a saída de uma das três redes neuronais estimadoras dos pesos fetais, previamente treinadas. Pode ver-se alguns resultados preliminares na figura 6.5, que mostra quatro gráficos, dos quais, o primeiro é a evolução do potencial informativo, o segundo mostra a abertura da janela de Parzen gaussiana, o terceiro a evolução dos pesos fetais (os círculos representando os valores verdadeiros e as cruzes os valores estimados), e o quarto, representa o erro absoluto do peso fetal estimado num conjunto de teste com 78 exemplos.

O erro de estimativa do comité num conjunto disjunto foi de 8.7%, valor ligeiramente superior ao que se obteve em cada uma das redes componentes do comité: 7.5%, 7.1% e 8.5%. Destes resultados que não se pode inferir muitas conclusões uma vez que não foram realizados testes sobre a capacidade de generalização do comité.

Estes resultados são aqui referidos apenas para salientar que a apreciação que

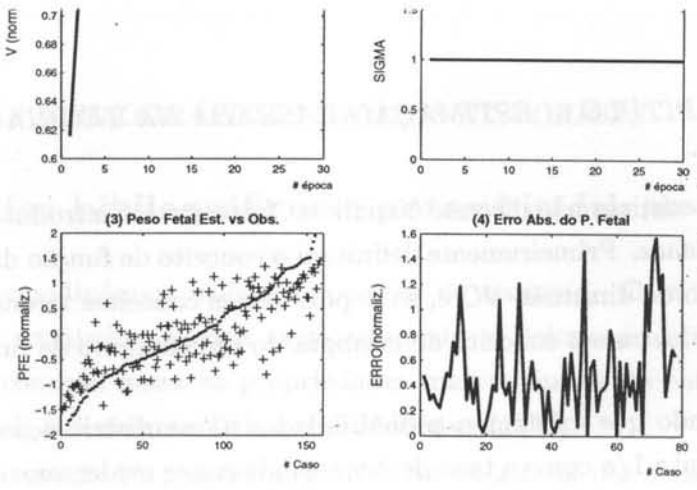


Figura 6.5: Resultados preliminares obtidos com um perceptrão simples treinado por um critério informático-entrópico ($\alpha = 2$).

se faz do critério torético-informático de treino de redes neuronais parece merecer uma atenção significativa, havendo ainda muitos outros aspectos em que importa conseguir uma melhor compreensão e introduzir melhoramentos.

6.5 Sumário

A *teoria da informação* no seu nível sintáctico, tal como foi introduzida por Shannon em 1949, tem muitas limitações quando aplicada ao processamento de informação, contudo, não obstante algumas limitações óbvias, os princípios subjacentes têm uma importância fundamental. A engenharia de sistemas e a aprendizagem automatizada utilizam abordagens baseadas na teoria da informação que, por um lado, visam obter uma melhor compreensão da natureza dos problemas de manuseamento da informação, a sua aquisição, armazenamento, processamento e interpretação e, por outro lado, procuram um constante aperfeiçoamento das técnicas de codificação, detecção, reconhecimento de padrões e filtragem.

O número N das diferentes partições por hiperplanos de um cubo n-dimensional, sabia-se em meados dos anos 1960, que estava limitada por $N \leq e^{n^2}$. Vapnik e Chervonenkis (1964) demonstraram, para um caso particular, que separando sem erro o conjunto de treino com um hiperplano num espaço de características binário e n-dimensional, então, com probabilidade $1-\eta$ se poderia afirmar que o erro de teste estaria limitado pela razão $(\ln N - \ln \eta)/m$. Neste limite o termo de capacidade, $\ln N$, tem uma grandeza da ordem de n^2. Ao deduzirem este limites para um espaço

de características não discreto Vapnik e Chervonenkis introduziram vários conceitos de *capacidade*. Primeiramente definiram o conceito de função de crescimento, depois o conceito de dimensão VC e, só depois destes conceitos terem sido introduzidos, é que introduziram o conceito de *entropia* de um conjunto de funções indicatrizes.

Supondo que existem n probabilidades numa distribuição discreta, cada uma sendo igual a $1/n$ como a taxa de diversidade cresce evidentemente com n, para obter uma medida desta adopta-se uma função de n crescente monótona e que tenha o valor zero quando $n = 1$, sendo escolhida, de entre um conjunto de funções, a função $k \ln n$, ao qual se chama a *entropia* de um conjunto de n probabilidade iguais, sem referência à natureza dos atributos. Em Física os atributos são os estados quânticos. A entropia, em teoria da informação, é simplesmente uma medida logarítmica do carácter aleatório ou incerteza preliminar acerca dos resultados de um ensaio. Para provas independentes ela é aditiva. Quando se considera a entropia conjunta de uma série completa de N ensaios repetidos e independentes, caracterizados por uma variável aleatória Y o postulado da adição dará uma entropia total de $-N \sum P(Y) \ln P(Y)$, logo a entropia média por ensaio será $H(Y) = -\sum P(Y) \ln P(Y)$. Esta expressão para $H(Y)$ pode ser interpretada como a entropia da distribuição de probabilidades $P(Y)$, anula-se quando todas as probabilidades do conjunto são nulas, com excepção de uma única, e é máxima para um número dado de probabilidades, quando estas são todas iguais. Se o número de valores de Y for infinito então $H(Y)$ pode não ter um limite superior.

O conteúdo de informação de uma mensagem pode ser definido pela capacidade mínima necessária para o seu armazenamento. Pretende-se calcular a capacidade média mínima necessária para uma mensagem cujas probabilidades dos estados se supõem conhecidas. Quando o número de estados é dois e as suas probabilidades são p e q, então, o conteúdo médio de informação de uma mensagem binária será $h = -p \ln p - q \ln q$. Quando existem mais que dois estados Shannon propôs a fórmula $h = -\sum P_i \ln P_i$, em que P_i é a probabilidade a priori do estado i.

Pode definir-se uma generalização da entropia devida a Renyi, $h_r(Y)$, de ordem r dada pela expressão $h_r(Y)) = \frac{1}{1-r} \log \left[\int f^r(x)\, dx\right]$, para $0 < r < \infty$ e $r \neq 1$. A entropia de Renyi de ordem 0 é a medida do conjunto de suporte da densidade f, e a entropia de Renyi de ordem 1 é a entropia de Shannon.

Recentemente foram propostos para sistemas dinâmicos dominados por ruído estacionário novos métodos de determinação não paramétrica dos parâmetros de um sistema através da minimização da entropia dos residuais dos erros, e introduzem as noções de *potencial informativo* e *força informativa*.

6.6 Referências bibliográficas e notas históricas

A segunda lei da termodinâmica utiliza o conceito de *entropia.* Com o aparecimento da mecânica estatística, devida principalmente a Boltzmann (1844-1906), estabeleceu-se uma conexão entre as propriedades macroscópicas da entropia e o estado microscópico de um sistema Tribus and McIrvine (1971).

A medida da informação foi pela primeira vez caracterizada nos anos 1920 por Nyquis (1924), Hartley (1928) e Fisher (1925). Nos anos 1940 o assunto foi retomado por Wiener (1948) que relacionou entre si os conceitos gémeos de informação e retroacção (feedback), cristalizando esta ideia na designação *cibernética,* a qual significa uma teoria geral e estrutural centrada sobre processos que se podem verificar quer em organismos vivos, quer em máquinas, e inclusivamente em sistemas sociais.

Shannon em 1948, publicou o ensaio "*The mathematical theory of comunications* Shannon (1948)", considerado como sendo a base da teoria moderna da comunicação e um dos mais citados e célebres artigos da história das comunicações, tecnologia e ciência.

Nos anos 1960 Kolmogorov, Solomonoff e Chaitin desenvolveram de forma independente teorias da informação baseadas no tamanho dos programas para a máquina de Turing Kolmogorov (1950), Kolmogorov (1965), Kolmogorov (1968), Solomonoff (1964), Chaitin (1974), e Chaitin (1990),

O princípio da máxima entropia foi pela primeira vez enunciado explicitamente em 1957 por E. T. Jaynes, mas, só nos anos 1970, Jaynes (1978), Jaynes (1994), é que este autor começou a publicar e, na opinião de Kapur (1993), as potencialidades deste princípio estão ainda por explorar.

Shannon estudou modelos estocásticos para o Inglês no seu ensaio seminal e Cover (1991) e seus colaboradores desenvolveram uma nova técnica de estimação da entropia do Inglês. Não obstante a exclusão do aspecto semântico verificou-se uma intensa investigação sobre a aplicação da teoria da informação na psicologia e educação, particularmente devida a Attneave (1953), Miller (1958) e Quastler (1955). Estes investigadores empregaram com sucesso os conceitos de unidade de medida da teoria da informação na interpretação e análise estatística e sintáctica dos resultados de psicologia experimental

Na Europa a teoria da informação desenvolvia-se segundo esta linha educacional. Em meados dos anos 1960 Cube (1960), Frank (1960) e Weltner (1973) sugeriam o conceito de *informação subjectiva* que era considerado aplicar-se a textos informativos segundo diferentes graus de dificuldade, grupos etários, e graus de domínio da língua e conhecimento prévio do domínio. A informação subjectiva de um dado texto era considerada como uma medida da relação entre o texto e o seu leitor, opondo-se assim a uma medida baseada numa fundamentação meramente estatística.

O problema da aprendizagem activa foi estudado extensivamente na teoria económica e estatística nos anos 1980 e seguintes. Por exemplo, Mandelbrot (1999) utiliza uma *dimensão de capacidade*, que também se designa por dimensão fractal, de Hausdorff ou de Hausdorff-Besicovitch, a qual é um limite superior da *dimensão de informação*, designação adoptada por Weisstein (2002) para a entropia.

O delineamento experimental num quadro Bayesiano usando a informação de Shannon como uma função objectivo foi estudado inicialmente por Lindley (1956). Uma característica notável desta abordagem consiste em tornar a optimização do delineamento experimental independente dos testes que são aplicados aos dados e às funções de custo associadas com qualquer decisão (Sivia (1997), pp.159-176)

O trabalho de Baum (1991) sobre a selecção activa de dados ("*query learning*") relaciona-se com problemas de classificação perfeitamente separáveis e apresenta resultados empíricos do algoritmo utilizado.

Nos finais dos anos 1990 surgiu um novo interesse pela teoria da informação, Principe et al. (2000b) e Fisher et al. (1999) apresentam um novo enquadramento teorético-informático na categorização e modelação de processos dinâmicos. As aplicações ao processamento de imagem são descritas, por exemplo, na revisão sobre "*formação informático-teorética da imagem*" de O'Sullivan et al. (1998) . Recentemente Guan et al. (2002) editam um número especial da "*IEEE Transactions on Neural Networks (vol. 13, No.4, July 2002)*" dedicado ao processamento multimídia inteligente, no qual a teoria de informação parece ter sido esquecida.

A partir de meados dos anos 1990 Haussler et al. (1994b), Haussler and Opper (1997a), Haussler (1997) e Haussler and Opper (1997b), usam as teorias da informação e da mecânica estatística para investigar as relações entre a aprendizagem Bayesiana, a complexidade de uma amostra e a dimensão VC. Em 2001 a contribuição de David Haussler é publicamente reconhecida como uma das mais relevantes na obtenção da primeira versão preliminar do genoma humano.

Algumas ideias referidas neste capítulo sobre a teoria da complexidade e os métodos entrópicos são provenientes de textos sobre processamento de sinal, em particular, escritos em língua portuguesa, salientam-se Abrantes (1998), Abrantes (2000), Campani and Menezes (2002) e Cardoso (2001).

Capítulo 7

Discussão e Conclusões

7.1 Discussão dos Resultados

As experiências com modelos de rede neuronal PMC, referidas em Sereno et al. (1999) e descritas mais detalhadamente na subsecção 5.1.1, em que se utiliza um conjunto de treino com 140 casos, sem qualquer pré-selecção, mostrou que era possível uma redução de cerca de 1% do erro relativo de estimação do peso fetal em relação ao método clássico. A melhor rede, desta série de experiências, apresentava um erro relativo de 8.08%, enquanto uma das fórmulas de regressão múltipla, a de Hadlock et al. (1985), apresentava um erro relativo de 9.30%, para um mesmo conjunto de teste (disjunto do de treino). Estas experiências apontaram na direcção de se tentar um melhoramento do pré-processamento dos casos, no sentido de o treino do modelos se fazer com casos mais apropriados e, ao mesmo tempo, reconhecia-se a necessidade de aumentar o número de casos de treino.

O conjunto de experiências seguintes foram focadas ainda em modelos de rede neuronal PMC, mas introduziu-se um novo modelo de estimadores, baseados na rede de funções de base radiais FBR. Fez-se referência a estes modelos em Sereno et al. (2000b) e Sereno et al. (2000a), descrevendo-se mais detalhadamente os resultados experimentais na subsecção 5.1.2. Os progressos na colheita de dados permitiu quase que duplicar o tamanho do conjunto de treino, continuram-se os estudos estatísticos e exploratórios, que se referem no quarto capítulo, mas, por falta de uma compreensão suficiente acerca do conhecimento *a priori* sobre os erros eventualmente existentes nos dados, optou-se por dar prioridade à quantidade dos casos, em detrimento de uma selecção baseada num dado critério qualitatitivo de selecção dos casos. A técnica de avaliação cruzada da performance com subconjuntos de 70 casos permitiu tirar um melhor partido dos casos existentes. Os resultados que se obtiveram nos modelos PMC revelaram que os modelos de rede neuronal PMC continuavam a ter uma vantagem percentual no erro relativo de cerca 1%, comparativamente com os modelos

clássicos. Em conjuntos de teste, disjuntos dos de treinos, os valores das médias dos erros relativos eram de 7.17% para as redes, enquanto que para as fórmulas de regressão atingiam os 8.16%. Por influência duma avaliação positiva das vantagens dos modelos de rede FBR, e que são referidas por exemplo em Bishop (1995), Haykin (1999) eReed and Marks (2000), surgiu a motivação de realizar ensaios com este tipo de redes, os quais aparecem descritos no aspecto teórico no terceiro capítulo, e, no que diz respeito à aplicação prática, no quinto capítulo, particularmente na subsecção 5.2.1. Entretanto, e com base em estudos exploratórios identificaram-se alguns casos que eram *outliers* e que se justificava serem suprimidos, e introduziram-se novos casos que entretanto tinham sido recolhidos. Usando ainda uma técnica de validaçãos cruzada, e conjuntos de teste disjuntos dos de treino, os valores das médias dos erros relativos das FBR desceram para 6.95%, e os das fórmulas de regressão desceram também para 7.54%, uma vez que se tinham filtrado alguns dos principais casos desviantes. Simultaneamente, sob a influência das recomendações referidas em Hudson and Cohern (2000), começou-se a dar mais atenção a aspectos quantitativos do raciocínio médico, e a técnicas de epidemologia clássicas, em particular procedeu-se a uma análise dos erros considerando os valores da sensibilidade e especificidade . O conceito de sensibilidade, por exemplo, teve uma grande relevância na avaliação da performance de modelos de estimação clínicos. Seja por exemplo os casos em que acontece serem as estimativa de peso fetal, por exemplo, baixas, portanto com risco, e que se verificava durante o parto serem verdadeiras, isto é, consideravam-se casos positivos verdadeiros. Um modelo que apresenta uma razão elevada entre o números destes casos, positivos verdadeiros, e o número total de casos de pesos baixos, pode ter uma maior importância na decisão clínica do que no caso desta taxa ser muito baixa.

Esta preocupação conduziu à realização de experiências numa nova perspectiva do problema. Considerou-se uma partição do conjunto de dados em três subconjuntos, pesos fetais baixos, médios e altos, e encarou-se a estimação do peso fetal como um problema de classificação de cada novo caso numa destas três partições. Em Sereno et al. (2000a) e na subsecção 5.2.2 apresenta-se os estudos e as experiências realizadas segundo esta perspectiva com modelos estimadores. Pela primeira vez pareceu ser artificial a separação, que muitas vezes se considera, entre os paradigmas da classificação de padrões e da regressão. Na prática um sistema de ajuda à decisão clínica para ser útil parecia dever estar concebido tendo em vista estas duas perspectivas, primeiro classificar e depois estimar o peso fetal, condicionado à classe a que o caso pertence. Fizeram-se experiências dentro desta linha de pensamento utilizando os dados classificados apenas em duas categorias, ou pesos baixos ou pesos altos, desprezando-se, portanto, os casos com pesos médios.

Treinaram–se *máquinas de vectores de suporte* MVS para fazer a classificação,

tal como se descreveu em Sereno et al. (2000a) e em 5.3.2, e analisaram-se as matrizes de confusão dos resultados da classificação. As percentagens de erros na classificação de novos casos foram, para os pesos baixos 11.2% e, para os pesos altos 10.0%. Mais importante do que estes resultados não serem muito animadores era o facto de se verificar que as máquinas MVS que se estava a treinar para esta classificação incorporavam a quase totalidade dos casos de treino como vectores de suporte. Deste modo as máquinas memorizavam os dados de treino e tinham muito poucas possibilidades de vir a ter uma boa generalização. Por um lado tinha-se um número muito reduzido de casos, quer de pesos baixos, quer de pesos altos, para treino e teste, por outro, a implementação em $MATLAB$ da MVS que se estava a utilizar tornava o treino demasiadamente lento, chegava a exigir por vezes mais horas de computação (a 500 MHz). Estas dificuldades não encorajavam a utilização de uma metodologia de análise de resultados dos modelos mais rigorosa a validação pelo *método de deixar-um-de-fora.*

Admitiu-se que se deveriam prosseguir dois objectivos: (*i*) Os dados precisavam de ser ainda mais filtrados; (*ii*) Era preciso ter-se um conhecimento mais detalhado sobre como funcionava de facto algébrica e numericamente o algoritmo das MVS que estava a ser utilizado.

O primeiro objectivo motivou a utilização de uma MVS de regressão para fazer uma espécie de edição nos dados, por outras palavras, melhorar a estrutura e consistência entre a matriz das entradas e os valores alvo correspondentes a cada caso. Todos os casos que apresentavam resíduos em valor absoluto superiores a 50 g quando estimados por uma MVS que foi treinada para ter um sobreajustamento a todos os dados, esses casos eram considerados inapropriados, quer para o treino, quer para o teste. Os melhores resultados que se obtiveram, os quais foram apresentados em Sereno et al. (2001a) e Sereno et al. (2001b) e são referidos mais detalhadamente na subsecção 5.3.3, são relativos a três MVS especializadas na estimação em três grupos de 100, 80 e 98 casos. Um terço dos casos de cada um destes grupos foi escolhido aleatoriamente para servir de conjunto de teste. Para cada um destes grupos verificaram-se erros relativos de 6.3%, 2.2% e 4.8%, respectivamente, para o primeiro grupo, constituído por cerca de 50% de casos cujos pesos eram baixos, para o segundo grupo constituído todo ele por casos de pesos médios e, para o terceiro grupo, constituídos por cerca de 50% de casos de pesos altos.

O segundo objectivo, uma dissecação do algoritmo implementado em $MATLAB$ da MVS constituiu a motivação para um estudo aprofundado da teoria das MVS,o qual que se esboçou, nas suas linhas gerais, na quase totalidade do terceiro capítulo, podendo notar-se que até a descrição do perceptrão é feita mais segundo uma perspectiva de máquina de *kernel,* do que segundo uma perspectiva inspirada na teoria clássica, que poderia ser sugerida por Haykin (1999), Bishop (1995) e Hertz et al.

(1991). Adicionalmente, em complemento deste estudo procedeu-se a uma tentativa de "*análise numérica*" do algoritmo de aprendizagem de uma MVS, a qual se apresenta na subsecção 5.5 do quinto capítulo.

Esta análise numérica permitiu identificar alguns dos problemas que podem surgir quando se faz uma aproximação, ou uma estimação, de funções definidas por conjuntos de dados (univariáveis, na referida análise), nomeadamente situações de instabilidade numérica devido a erros de truncatura. Estes deveriam constituir uma das razões pelas quais o algoritmo que se estava a utilizar, por vezes, não era convergente e obrigava a uma interrupção voluntária do processo de computação.

Um outro objectivo de investigação que se prosseguiu, referido em Sereno et al. (2002b), consistiu em saber até que ponto a aprendizagem baseada num critério teorético-informático permitia uma redução dos valores de erro de previsão dos pesos fetais obtidos com modelos redes neuronais PMC treinadas com o critério de minimização do erro médio quadrático. Em Sereno et al. (2002a) apresenta-se os resultados de um estudo comparativo entre modelos de rede PMC treinadas pelo método neuronal clássico, baseado no critério de erro dado pela norma L_2, e redes treinadas pelo método baseado no critério teorético-informático.

7.2 Discussão epistemológica de quatro tópicos sobre a Teoria Estatística da Aprendizagem

A Teoria da Aprendizagem Estatística (TEA) investiga os processos de estimação de dependências funcionais existentes em conjuntos de exemplos, e analisa quais são os factores responsáveis, e como se pode ter um controlo sobre estes, de forma a garantir que um dado modelo, estimador de dependência funcional, tenha uma boa *generalização*.

Ao contrário de muitos métodos estatísticos clássicos, que foram desenvolvidos preferencialmente para uma inferência em amostras *grandes* e condicionadas por vários tipos de informação *a priori*, a teoria da aprendizagem e generalização interessa-se pela inferência em amostras *pequenas* e pelo controlo da capacidade da estrutura de conjuntos de algoritmos que se podem obter através de uma "*máquina*" construída segundo princípios teoricamente bem estabelecidos.

Pode distinguir-se duas abordagens sobre a aprendizagem. Uma considera que esta constitui um problema de *minimização* de um funcional do risco esperado, numa situação em que se desconhece uma medida das probabilidades desse risco, mas conhece-se um conjunto de exemplos *i.i.d.* A outra abordagem considera que a aprendizagem consiste num problema de *identificação* de uma função ideal (densidade de probabilidade) da qual se conhecem alguns exemplos a partir dos quais se pretende obter a função que melhor aproxima a relação funcional subjacente,

abordagem que conduz necessariamente a um tipo de problemas conhecidos por problemas de formulação incompleta (*ill-posed problems*).

Existem ainda outros aspectos relacionados com a maneira como se faz a abordagem à TEA e que não foi possível fazer uma referência adequada nos capítulos anteriores. Uma vez que constituíram uma preocupação recorrente em muitas situações procura-se sumariamente expor a seguir particularmente quatro aspectos. O primeiro, tem a ver com a atitude do investigador em aprendizagem automatizada em face das do método de demonstração usado em TEA, não só dos teoremas que são referidos no segundo capítulo desta dissertação, mas de muitos outros que aparecem na literatura relacionados com a TEA. O segundo aspecto diz respeito às relações entre a TEA, a lógica indutiva e outras lógicas (em particular a lógica difusa '*fuzzy*'). O terceiro aspecto relaciona-se com a estatística bayesiana e os processos gaussianos e nas similaridades que estes poderão ter com os métodos da TEA. Finalmente, o quarto aspecto diz respeito ao problema das medidas indirectas e da matemática da inversão que se pode dizer constituir um antepassado a redescobrir para uma compreensão adequada de muitos resultados actualmente estabelecidos pela TEA .

7.2.1 Interpretação das demonstrações dos teoremas em TEA

Nesta tese são referidos vários teoremas básicos da TEA nos capítulos segundo e terceiro, mas não se apresentam as respectivas demonstrações. Este facto não deve ser interpretado no sentido de que se considerou que as demonstrações não eram relevantes, mas, que devido a várias particularidades da demonstração em estatística esta tarefa parecia desviar atenção dos aspectos aplicacionais.

Considera-se que a aplicação de um determinado resultado que foi obtido dedutivamente em conformidade com uma dada axiomática num dado campo da matemática só poderá ser devidamente aplicado se este processo dedutivo poder ser claramente interpretado por um leitor fortemente interessado. Acontece que a dimensão didáctica de textos de referência da literatura sobre TEA, talvez devido à recente emergência desta, parece ter sido posta em segundo plano.

Vai-se apresentar um exemplo que se destinam apenas a ser uma anotação para uma posterior investigação mais aprofundada, sobre este mesmo assunto, a fazer-se no futuro.

Vai-se considerar, por exemplo, o teorema (43) enunciado no segundo capítulo desta tese, designado *teorema chave da TEA*. O texto de base deste teorema é o terceiro capítulo de Vapnik (1998) (pp.79 – 120), em que se abordam as importantes condições para a consistência do princípio MRE, da minimização do risco empírico. O autor, primeiramente, apresenta uma definição clássica de *consistência* (definição 42) a esperança matemática e o risco empírico devem convergir em probabilidade

para um mesmo ínfimo do funcional de risco:

> "O princípio MRE é consistente para o conjunto de funções $c(x, y, f(x, w))$, em que os parâmetros w pertencem a um dado conjunto finito $\mathbb{W}$, e para a distribuição de probabilidades $P(x, y)$, se o método de minimização do risco empírico aplicado a cada amostra de dimensão m, produz uma sucessão de funções $c(x, y, f(x, w_m))$ tal que:

$$R[f(x, w_m)] \xrightarrow[m\to\infty]{P} \inf_{f\in\mathbb{F}} R[f] \tag{7.1}$$

$$R_{emp}[f(x, w_m)] \xrightarrow[m\to\infty]{P} \inf_{f\in\mathbb{F}} R[f] \tag{7.2}$$

Seguidamente, estabelece uma distinção entre o caso de consistência *trivial* e o caso de consistência *restrita,* demonstra um lema aplicável ao método de minimização do risco empírico estritamente consistente, e utiliza um exemplo para mostrar que a convergência do risco esperado 7.1 pode existir mas não existir a da equação 7.2.

Esta explicação exige bastante reflexão, o exemplo sugerido é abstracto porque não se visualiza facilmente um conjunto de funções indicatrizes para um funcional de risco, $f(x, w), w \in \mathbb{W}$, definidas no intervalo $[0, 1]$, tal que o parâmetro w define os intervalos nos quais a função é igual a zero, para se verificar um ajuste a um conjunto de valores de x amostrados. Se um leitor não acompanhar a sua leitura de uma reescrita das expressões usadas na exposição, ou não estiver muito familiarizado com este tópico, provavelmente não avança mais, tentou ler seis páginas e adia a leitura. Se prosseguir, entra-se numa abordagem do conceito de *processo empírico ("two-sided empirical process")* aplicado a uma sucessão de variáveis aleatórias, em relação às quais se pretende conhecer as condições de convergência em probabilidade para zero. Por outras palavras, pretende-se estabelecer as condições de convergência uniforme absoluta da média para a respectiva esperança matemática para um dado conjunto de funções. A exposição prossegue com uma observação sobre a lei dos grandes números e a sua generalização, a qual se fosse omitida pensa-se ninguém daria pela falta, e, finalmente, chega-se ao teorema chave da aprendizagem, que que tem o seguinte enunciado (Teorema 43):

> "Seja o conjunto de funções $c(x, y, f(x, w))$ em que os parâmetros w pertencem a um dado conjunto finito $\mathbb{W}$. Suponha-se que as funções de custo c são limitadas inferior e superiormente:

$$A \leq \int c(x, y, f(x, w))\, dP(x, y) \leq B.$$

Então, uma condição necessária e suficiente de consistência não trivial do

princípio de minimização do risco empírico consiste em o risco empírico $R_{emp}[f(x,w)]$ ser uniformemente convergente para o risco verdadeiro $R[f(x,w)]$ no conjunto de funções $c(x,y,f(x,w))$ e no sentido em que:

$$\lim_{m\to\infty} P\left\{\sup_{w\in\mathbb{W}}\left(R[f(x,w)]-R_{emp}[f(x,w)]\right)>\varepsilon\right\}=0, \quad \forall\varepsilon>0"$$

Coloca o problema da medida de probabilidade que se está a considerar, e enuncia um corolário que descreve as condições de convergência para um conjunto de funções de distribuição. Segue-se a demonstração da primeira parte do teorema chave que começa pela aplicação da definição de consistência estrita, para uma dada medida de probabilidade. A sequência da demonstração omite-se, dado que existem passos que requeriam investigações complementares, para permitir uma explanação compreensiva, mais completa do que a simples tradução do texto Vapnik (1998) (pp.89-92).

7.2.2 A lógica indutiva

Só a partir dos anos 1920 é que se deu o desenvolvimento da ideia seminal da *lógica indutiva* relacionando-a com os fundamentos matemáticos do cálculo das probabilidades. A ideia de lógica indutiva teve uma forte influencia de Jeffreys (1939), Popper (1972) e aparece exposta por Rudolf Carnap na sua obra seminal "*Logical Foundations of Probability* (1950). Esta teoria parece conter matéria de reflexão importante para um aprofundamento da teoria estatística da aprendizagem no sentido de melhorar algumas das suas deficiências, nomeadamente as didácticas que foram esboçadas na subsecção anterior. Segue-se nesta exposição um texto muito sintético devido a Dubucs (1998).

Entre os seus aspectos mais importantes figuram o conceito chave de implicação *parcial*, no qual, em certo sentido, uma dada conclusão B está dedutivamente implicada por uma "*parte*" de uma dada premissa A, e o facto de a lógica indutiva estar associada a noções *métricas*.

Diz-se que uma hipótese H é implicada pela evidencia E com um certo grau, que se designa por *grau de confirmação*, e representa-se por $c(H,E)$. O problema central da lógica indutiva consiste em definir, para uma dada uma linguagem $\mathbb{L}$, uma função, $c:\mathbb{L}\times\mathbb{L}\to[0,1]$, que associe a todo o par de enunciados oriundos da linguagem $\mathbb{L}$, um número real não negativo, como *medida* do grau em que o primeiro enunciado é uma implicação do segundo. Esta função deve ser calculável e satisfazer as leis formais do cálculo de probabilidades.

Uma semelhança entre a lógica indutiva e a dedutiva reside no facto de ambas serem concebidas como disciplinas formais e independentes, quer dos estados empiricamente realizados das coisas, quer das nossas crenças subjectivas.

Uma diferença em relação à lógica dedutiva é que nesta última existe a regra de *modus ponens* (do tipo: P, P implica Q; logo Q), enquanto que não é possível na lógica indutiva uma regra equivalente (do tipo: E, E implica parcialmente H com um grau elevado; logo H), uma vez que por maior que seja o grau $c(H,E) < 1$, verifica-se sempre que E permanece compatível com a falsidade de H.

No plano das aplicações da lógica indutiva só se pode enunciar a regra metodológica "E, E confirma H com um *grau* r; portanto, verifica-se H com *grau* r" se se acompanhar com uma restrição, designada *requisito de evidência total.* Este deve incluir, por exemplo, mencionar-se o *momento* em que o juizo de credibilidade é pronunciado, e o enunciado E deve ter a totalidade da evidência disponível nesse mesmo momento.

Considere-se uma linguagem $\mathbb{L}_m^\pi$ com um número finito π de predicados e um número também finito m de indivíduos. Chama-se *descrição de estado* a uma conjunção do tipo

$$\bigwedge_{i,j} \varepsilon_{ij}\ \varphi_i\ x_j, \quad i = 1, ..., \pi, \quad j = 1, ..., m$$

em que ε_{ij} ou é um símbolo vazio, ou é o símbolo negação. Esta descrição traça *um estado possível do mundo* ao dizer que um dado indivíduo x_j é, ou não é um φ_i . Chama-se característica (*rang*) $R(A)$ de um enunciado A o conjunto das descrições de estado para as quais A é verdadeiro. Afecta-se cada descrição com uma certa *massa* de tal maneira que a massa total seja 1. Esta repartição das massas pode evidentemente efectuar-se de várias maneiras. Ela induz sobre a álgebra de Boole de todas as características uma medida m de probabilidade. Quando cada elemento do conjunto $\mathbb{S}$ das descrições recebe uma massa estritamente positiva, fala-se de medida *regular.* Define-se então o grau de confirmação por :

$$c(H,E) = \frac{m(H,E)}{m(E)} = \frac{\sum_{S \in R(HE)} m(S)}{\sum_{S \in R(E)} m(S)}$$

Esta função $c(H,E)$ chama-se função de confirmação ou *c-função* e sempre que a medida subjacente seja regular a função é também *regular.* Como a medida da característica de E tem o valor 1 quando E for uma tautologia, pode identificar-se $m(H)$ a $c(H,t)$,em que t designa um enunciado identicamente verdadeiro.

Obtém-se, assim, um significado preciso para a noção de implicação parcial. Dizer que E implica dedutivamente H é equivalente a dizer que H é verdadeiro em todos os mundos possíveis em que E for verdadeiro. Dizer que E implica H com o grau $r\,\%$, quer dizer que H é verdadeiro em $100r$ dos estados do mundo em que E é verdadeiro, ou ainda que uma proporção r da característica de E está na de H. As *c-funções* tornam-se pois instrumentos que medem o suporte conferido a H pela evidencia E: elas determinam a relação entre o número de estados do mundo em

que E e H se verificam simultaneamente, e o número de estados em que apenas E se verifica.

Existem ainda condições suplementares de adequação das c-funções das quais se destacam.

Primeiramente, tem-se *a simetria.* As $c-$funções simétricas são as que cujos valores permanecem inalterados quando se permuta as constantes dos indivíduos. Num certo sentido, a exigência da simetria é uma expressão do principio da indiferença, segundo o qual deve tratar-se de maneira equivalente objectos dos quais não se tem razões para pensar que são diferentes. Por exemplo, o valor de $c(H, E)$ deveria ser igual ao de $c(H, E')$ se H for a hipótese universal $\forall x\,(\varphi x \supset \psi x)$ e se E e E' designam as evidencias respectivas φa & ψa e φb & ψb. A limitação às c-funções simétricas reverte na possibilidade de substituição de uns pelos outros, dos nomes dos indivíduos em $\mathbb{L}^{\pi}_{m}$.

Depois tem-se a condição de *ajustamento.* A definição de uma c-função para uma linguagem $\mathbb{L}^{\pi}_{m}$ que comporta uma infinidades de indivíduos, levanta um certo número de dificuldades que não se examinará (por exemplo, as descrições de estado tornam-se infinitas, e passam a não ser fórmulas da linguagem). Resta que a exigência de uma tal definição induz condicionamentos sobre as c-funções relativas às linguagens do tipo $\mathbb{L}^{\pi}_{m}$. De facto uma c-função c_{∞} relativa a $\mathbb{L}^{\pi}_{m}$ será definida como o limite de uma sucessão $\{\}$, $n \in \mathbf{N}$ de c-funções definidas sobre $\mathbb{L}^{\pi}_{m}$: $c_{\infty}(H, E) = \lim_{n\to\infty} c_n(H, E)$. Ainda é preciso que as c-funções consideradas se "prolonguem" umas nas outras. De facto, um enunciado quantificado,como $\forall x\,\varphi x$, mudas as condições de verdade quando se passa de $\mathbb{L}^{\pi}_{n}$ para $\mathbb{L}^{\pi}_{n+1}$ (ele deixa de significar $\wedge_{1\leq i\leq n}\varphi x_i$ mas passa a significar $\wedge_{1\leq i\leq n+1}\varphi x_i$); os enunciado não quantificados, pelo contrário, guardam a mesma significação. Deseja-se pois que c_n e c_{n+1} sejam *ajustados*, quer dizer que eles atribuam o mesmo valor a todo o par de enunciados não quantificados. É ainda uma restrição sobre as escolhas das c-funções : c_n sendo dado, certas escolhas de c_{n+1} serão interditas.

Finalmente, vai-se referir a condição de *adequação à aprendizagem pela experiência.* Trata-se aqui de definir as condições àsquais deverão satisfazer as c-funções para representar fielmente os procedimentos de aprendizagem indutiva. Algumas definições são necessárias para a sua formulação. designa-se por Q-predicado todo o predicado da forma $\wedge_{1\leq i\leq \pi}\varepsilon_i\varphi$, em que ε_i é o símbolo vazio ou a negação (existem pois $q = 2\pi$ $Q-$predicados numa linguagem $\mathbb{L}^{\pi}$). É evidente que pelo menos um indivíduo satisfaz a um $Q-$predicado dado, e que dois indivíduos que satisfaçam a um mesmo $Q-$predicado seriam indiscerníveis em $\mathbb{L}^{\pi}$: os Q-predicados são, no fundo, o análogo das *infimae species* aristotélicas, quer dizer dessas espécies que é impossível de dividir logicamente de novo. Todos os predicados ψ definíveis em $\mathbb{L}^{\pi}$ podem escrever-se *normalmente* sob a forma de uma disjunção de $Q-$predicados,

por uma equivalência do tipo $\psi x \Leftrightarrow \vee_{1 \leq i \leq l_\psi} Q_i x l_\psi$ (respectivamente l_ψ / q) chama-se a extensão lógica absoluta (respectivamente, relativa) de ψ. Pode-se então formular duas condições, uma sobre a pertinência ou "relevância" dos casos particulares, e a outra sobre a suficiência. Na primeira, atende-se a que o grau de confirmação deve crescer quando um caso particular vem juntar-se à antiga evidência : para cada $Q-$predicado $Q_k, c\left(Q_k x_i,\ E\ \&\ Q_k x_j\right) > c\ \left(Q_k x_i,\ E\right)$ quando x_i e x_j não figuram em E. Na segunda condição, de suficiência, quando se procura determinar a parte da experiência que é pertinente para avaliar $Q_k x_i$ é suficiente que se ligar à evidência que diz respeito aos casos particulares de Q_k. Por outras palavras, se $E\prime$ se obtém simplesmente a partir de E substituindo as ocorrências dum $Q-$predicado Q_p pelas de dum $Q-$predicado $Q_{p'}$ (com p e $p\prime$ diferentes de k) deve-se ter que $c\ \left(Q_k x_i,\ E\right) = c\ \left(Q_k x_i,\ E'\right)$.

7.2.3 Aproximação bayesiana e processos gaussianos

Uma rede neuronal pode ser vista como um modelo semi-paramétrico de estimação da densidade condicional. Em especial as redes neuronais que usam técnicas de pré-acção (i.é., são antero-alimentadas, "*feedforward*" em inglês) podem utilizar-se para obter sistemas de classificação.

O raciocínio bayesiano é capaz de fornecer uma aproximação probabilística à inferência, uma vez que se baseia no pressuposto de que as quantidades com interesse são governadas por distribuições de probabilidade e que as decisões óptimas se podem obter através de raciocínios acerca dessas probabilidades juntamente com os dados observados.

Considerando sucessivamente cada uma das classes C_k pode estimar-se as densidades condicionais correspondentes $p\left(\mathbf{x} \mid C_k\right)$ desde que se atribua a cada ponto do conjunto de dados uma etiqueta, de acordo com a classe a que ele pertence. Conhecidas estas densidades pode em seguida determinar-se a probabilidade *a posteriori* de um novo exemplo usando o teorema de Bayes e obter-se assim sua classificação.

A relação existente entre muitas redes neuronais e os classificadores de bayesianos de erro-mínimo é devida aos facto de que as saídas nem são verosimilhanças nem valores binários aproximadamente zero ou um, mas, efectivamente, são probabilidades *a posteriori*. Para um problema com k classes, as probabilidades bayesianas são estimadas no sentido da minimização do erro quadrático médio, quando a rede tem uma saída para cada classe de exemplos e as saídas desejadas são 1de k (i.é.,uma das unidades de saída correspondendo à classe correcta e as outras todas nulas), e se usa uma função de custo erro quadrático. Estas condições verificam-se frequentemente em redes MLP com unidades sigmoidais, usando rectropropagação e em redes FBR. Quando se estima rigorosamente as probabilidades bayesianas, se minimiza a percentagem dos erros de classificação e se garante que a soma das saídas é unitária,

então, as saídas podem ser encaradas como probabilidades. Adicionalmente esta propriedade torna ainda possível uma compensação das diferenças eventualmente existentes entre as probabilidades condicionais de pertença às classes nos exemplos do conjunto de treino e de teste.

Supondo a densidade de probabilidade $p(\mathbf{x}, y)$, nas variáveis aleatórias contínuas $\mathbf{x}$ e y, esta pode interpretar-se como uma descrição do gerador de dados que se admite existir subjacente a qualquer conjunto de exemplos de treino de uma *rede neuronal*. Pode decompor-se esta densidade de probabilidade num produto $p(\mathbf{x}, y) = p(y \,|\, \mathbf{x}) \cdot p(\mathbf{x})$, em que o factor $p(y \,|\, \mathbf{x})$ representa a densidade condicional dos dados alvo y em relação à ocorrência de um exemplo $\mathbf{x}$, e $p(\mathbf{x})$ representa a densidade não-condicionada de $\mathbf{x}$, dada por $p(\mathbf{x}) = \int p(\mathbf{x}, y)\, dy$. A densidade $p(\mathbf{x})$ desempenha um papel importante em vários aspectos das redes neuronais, incluindo os procedimentos para a determinação dos parâmetros das funções de base numa rede FBR. A densidade condicional $p(y \,|\mathbf{x})$ é importante na estimativa de y para novos exemplos $\mathbf{x}$.

Uma das funções de erro mais usadas, a média da soma dos quadrados, é motivada pelo princípio da máxima verosimilhança. Para uma conjunto de dados de treino $\{\mathbf{x}_i, y_i\}$ a verosimilhança $\mathcal{L} = \prod_i p(\mathbf{x}_i, y_i) = \prod_i p(y_i \,|\, \mathbf{x}_i) \cdot p(\mathbf{x}_i)$ o produto das probabilidades é consequência do facto desse supor que cada um dos padrões é retirado iid a partir da mesma distribuição.

Em vez de maximizar a verosimilhança torna-se em geral mais conveniente minimizar uma função de erro $\widehat{er}_{(\mathbf{x},y)}$ dada pelo negativo do logaritmo da verosimilhança

$$\widehat{er}_{(\mathbf{x},y)} = -\ln \mathcal{L} = -\sum_i \ln p\,(y_i|\mathbf{x}_i) - \sum_i \ln p\,(\mathbf{x}_i) \tag{7.3}$$

Na perspectiva de que uma rede neuronal modeliza a densidade de probabilidade condicional $p(y|\mathbf{x})$, o segundo termo de (7.3) não depende dos parâmetros da rede e portanto representa uma constante aditiva que pode ser omitida da função de erro obtendo-se

$$\widehat{er}_{(\mathbf{x},y)} = -\sum_i \ln p\,(y_i|\mathbf{x}_i)$$

Notar que a função de erro tem a forma de somatório estendido a apenas um termo do erro de cada padrão separadamente. Isto deve-se à suposta independência dos exemplos sujeitos a uma mesma distribuição desconhecida. Escolhas diferentes da função de erro são consequência de pressupostos diferentes sobre o tipo de distribuição condicional $p(y|\mathbf{x})$. Em problemas de interpolação (regressão) pretende-se estimar as variáveis alvo y, que são números reais contínuos, enquanto que em problemas de classificação os alvos y representam entidades discretas, etiquetas, que

definem estimativas de probabilidades de pertença à classe C_k.

7.2.4 Medição indirecta e a matemática da inversão

A grande maioria dos métodos de medida actuais são indirectos. Por exemplo, a medida do peso resulta da observação indirecta da distensão de uma mola, da temperatura resulta da observação do incremento da altura duma coluna de mercúrio no interior dum tubo capilar, do perímetro abdominal de um feto resulta da observação da imagem obtida indirectamente através de um processamento dos sinais originados por reflexão (eco) em certos tecidos das vibrações de frequência determinada proveniente de um gerador externo.

Desde sempre tem existido uma grande preocupação com os métodos que requerem que se estime a distribuição de uma dada grandeza a partir de um conjunto de medidas, cada uma das quais sofrendo uma influencia, maior ou menor, dessa distribuição que é desconhecida. Como diz Twomey (1977) "*o conjunto dos números nos quais está incluída a "resposta" tem de ser "desemaranhado", por assim dizer, entre uma emaranhação de combinações desses números* (p.2)." Este autor ilustra esta ideia com um caso de inferência de medidas num fenómeno físico, o qual se reproduz a seguir, uma vez que ele parece apresentar interessantes analogias com o problema da estimação do peso fetal que se estuda detalhe nesta tese. Nesta perspectiva poder-se-ia dizer que o nosso objectivo consiste em a partir de um conjunto "emaranhado" de medidas de grandezas obtidas indirectamente por ecografia, e introduzidas numa rede neuronal tornar possível obter-se uma fórmula aproximada que, por assim dizer, simule, ou contenha um método de "desemaranhar" esses mesmos dados.

Suponha-se que a energia de radiação infravermelha numa dada altitude x da camada gasosa atmosférica depende de uma certa forma da temperatura, e suponha-se também que é uma característica dessa mesma altitude o comprimento de onda irradiada λ. A primeira suposição é verdadeira, a segunda só parcialmente. É claro que um instrumento de medida apropriado colocado num satélite artificial acima da atmosfera poderia registar a variação de temperatura com a altitude a partir da medida da intensidade da radiação infravermelha em comprimentos de onda discretos. Porém surge um problema, a energia total de radiação infravermelha para um dado comprimento de onda resulta não só da contribuição da camada atmosférica a uma dada altitude, mas também da contribuição das camadas com altitudes vizinhas, em grau variável, e que se poderá supor obedecerem a leis de distribuição gaussianas. Surge assim uma certa "nebulosidade" à volta do valor da altitude a que corresponde uma dado comprimento de onda.

Para se descrever matematicamente este fenómeno da "nebulosidade" que se acaba de referir, considere-se $f(x)$ a função de distribuição que se pretende determinar e seja a função $K(\lambda, x)$ a curva de contribuição relativa em valores discretos

de $\lambda = \lambda_1, \lambda_2, \ldots, \lambda_n$. Uma medida do comprimento de onda λ_i envolve radiações proveniente da altitude x_{λ_i} na qual se verifica um máximo da função $K(\lambda_i, x)$, assim como das regiões vizinhas desta altitude em que a função K não se anula. Se o intervalo entre x e $x + \Delta x$ contribuir para a i-ésima medida na quantidade $K(\lambda_i, x)\ f(x)\ \Delta x$, então a radiação total medida é obviamente

$$g(\lambda_i) = \int K(\lambda_i, x)\ f(x)\ dx, \tag{7.4}$$

em que os limites de integração exactos dependem de circunstâncias que de momento não são relevantes. É suficiente admitir-se que em todas as experiências o valor da altitude x nunca é negativo, nem excede um determinado altitude máxima X . Obtém-se assim a especificação de um problema indirecto de um sensor para o qual o integral relevante é $\int_0^X K(\lambda_i, x)\ f(x)\ dx$, o qual por uma simples mudança de variável de x para x/X é sempre possível transformar num integral $\int_0^1 K(\lambda_i, x)\ f(x)\ dx$.

Se as funções de distribuição das contribuições em várias altitudes, responsáveis pela "nebulosidade", correspondentes aos comprimentos de onda λ_i, tiverem todas a mesma forma e diferirem apenas na sua posição ao longo do eixo dos xx, então, se se definir $K_0(x)$ como sendo a função centrada na origem, pode escrever-se a expressão (7.4) de uma forma diferente, como o integral dependente de x_i, e das formas das funções $K_0(x)$ e $f(x)$, i.é., tem-se o chamado *produto de convolução*, ou simplesmente *convolução* destas duas funções:

$$g(\lambda_i) = \int K_0[\lambda_i, (x - x_i)]\ f(x)\ dx$$

Nas situações em que é necessário retirar a "nebulosidade", ou tornar as medidas espectrais mais nítidas, pode fazer-se a operação inversa, a chamada *desconvolução*. Vem a propósito referir, primeiramente, que a desconvolução está intimamente relacionada com a inversão de medidas indirectas, sendo apenas um caso especial em que a forma das funções $K(\lambda_i, x)$ são iguais e sofreram translações segundo o eixo dos xx . Em segundo lugar, assinalamos que a determinação da convolução é equivalente ao processo de geração do valor médio da distribuição de uma variável causal, i.é., a convolução permite uma *suavização* de um conjunto de dados, aspecto muito importante no reconhecimento de padrões e que aparece referido muitas vezes. Mas para se fazer a conversão de dados de uma tabela através de um integral precisa-se ainda de uma "função", chamada *delta* (de Dirac), representa-se por $\delta(x)$. Aplicando esta função, por exemplo, ao caso ideal sem "nebulosidade" obtém-se:

$$g(\lambda_i) = \int \delta[\lambda_i, (x - x_i)]\ f(x)\ dx$$

Em termos matemáticos pode dizer-se que o problema da inversão consiste em determinar-se uma função $f(x)$ resolvendo uma equação integral:

$$g(y) = \int K(y,x)\ f(x)\ dx \tag{7.5}$$

que se chama *equação integral de Fredholm* (porque os limites de integração não são variáveis como acontece no caso das equações de Volterra), *de primeira espécie* (porque $f(x)$ aparece apenas no integrando), e em que a função $K(y,x)$ é conhecida como "*função kernel*" ou simplesmente "*kernel*".

Situações deste tipo e equações semelhantes a equação (7.5) apareceram recorrentemente na teoria das redes neuronais, relacionadas, por exemplo, com um enfraquecimento da performance numa rede Bishop (1995), ou com o método de regularização Haykin (1999).

Por outro lado, notar que um conceito básico de um método de aprendizagem referido nesta tese, e que será desenvolvido e contextualizado para a explicação do "truque" da criação do "espaço de características" utilizado pelas máquinas de vectores de suporte para fazer a separação dos exemplos, é precisamente o conceito de "*kernel*". Relembremo-lo disfarçado noutras expressões nossas familiares, como por exemplo, $K(y,x) = e^{-yx}$ na transformada de Laplace, e $K(\omega,x) = e^{i\omega x}$ na transformada de Fourier.

Em Bishop (1995) aparece uma significativa referência à resolução de *problemas inversos*. Por exemplo, no caso que se acaba de referir a análise espectral, a reconstrução tomográfica, o controlo de instalações industriais, e a cinemática de robots constituem aplicações potenciais das redes neuronais. Nestes casos existe um problema *directo* ("*forward*") bem definido que é caracterizado por uma correspondência funcional (i.é., um valor único). Muitas vezes isto corresponde a uma causalidade num sistema físico. No caso da reconstrução espectral, por exemplo, o problema directo corresponde à avaliação do espectro quando se sabem os parâmetros (localizações, largura e amplitude) das linhas espectrais. Em aplicações práticas resolve-se, em geral, o problema inverso correspondente no qual estão trocados os papéis das variáveis de entrada e saída. No caso da análise espectral isto corresponde a determinar-se os parâmetros da linha espectral a partir do espectro observado. No problema inverso a correspondência pode muitas vezes ser plurívoca, existindo vários valores de entrada aos quais correspondem mais do que um valor de saída válido. Por exemplo, pode haver várias escolhas para os parâmetros de linha espectral que originam o mesmo espectro observado. Aplicando-se a este tipo de problemas inversos uma aproximação dos mínimos quadrados obtém-se uma aproximação da média condicional dos dados alvo a qual, frequentemente, originará uma performance muito fraca, uma vez que a média de várias soluções não é necessariamente a solução.

Em Haykin (1999) aparece também uma referência à dualidade problema directo-inverso. O fenómeno físico responsável pela geração dos casos de treino, as imagens ecográficas, por exemplo, parece ser um problema directo bem definido ("*well-posed*", em inglês). Contudo, a aprendizagem a partir de características extraídas destas imagens, vista como um problema de reconstrução duma hipersuperfície é um problema de inversão mal definido ("*ill-posed*", em inglês). Segundo Haykin (1999) (p.266) podem considerar-se as seguintes razões. Primeiro, o critério da *existência* pode ser violado na medida em que podem não existir saídas diferentes para todas as entradas. Segundo, pode não existir informação suficiente nos exemplos de treino tal como seriam precisas para reconstruir univocamente o isomorfismo (mapping) entre os conjunto das entradas e o das saídas, pelo que o critério da *unicidade* pode ser violado. Terceiro, a inevitável presença de ruído ou imprecisão em conjuntos de treino que aparecem na prática acrescentam incerteza à reconstrução dos isomorfismo entradas-saídas. Em particular , se o nível de ruído das entradas for muito elevado, é possível que a rede neuronal produza uma saída exterior ao espaço imagem correspondente ao espaço de entradas., por outras palavras é provável que o critério da *continuidade* seja violado. Se um problema de aprendizagem falha a propriedade da continuidade então o isomorfismo entradas saídas que for estabelecido pela rede neuronal não tem muito a ver com a solução do problema. Não existe uma maneira de ultrapassar esta dificuldade a não ser que se disponha alguma informação a priori sobre o referido isomorfismo entrada-saída. Neste contexto, relembra Haykin (1999) uma frase célebre devida a Cornelius Lanczos (1964), "Uma falta de informação nunca poderá ser remediada por um qualquer truque matemático" e conclui, "o aspecto importante de como fazer com que um problema mal-definido se torne um problema bem-definido reside na regularização." A teoria da regularização é referida em mais em detalhe no capítulo terceiro.

7.3 Discussão da evidência de benefícios provenientes da utilização de modelos de rede neuronal

7.3.1 Tomada de decisão assistida por computador e por inferência estatística em medicina

A evidência de benefícios para a saúde provenientes das redes neuronais na intervenção médica parece ser indiscutível Lisboa (2002). Poder-se-á colocar a questão de saber se o raciocínio médico se enriquece quando é apoiado em sistemas computacionais. A resposta poderá ser afirmativa desde que seja possível os próprios médicos interessarem-se e terem um papel activo nas fases mais importantes da sua concepção, prototipagem, teste, validação e disseminação. Por exemplo, um

dado sistema médico como um sistema de rim artificial, de coração artificial, de monitorização de cuidados intensivos, e muitos outros parece ter beneficiado desse envolvimento dos próprios interessados. A resposta poderá não ser afirmativa tudo dependendo da cultura que existir e das particularidades duma dada situação, mas, de um modo geral, é de esperar que uma maioria qualificada de médicos tenda a responder negativamente.

Neste secção vai abordar-se alguns tópicos relevantes sobre a tomada de decisão biomédica com recurso a técnicas assistidas por computador, com o objectivo de esclarecer as razões destas posições paradoxais perante a tecnologia da informação.

Para começar, e como introdução, pode referir-se uma citação extraída de um manual de Medicina Interna, e que todos os médicos conhecem, "*Harrison's Principles of Internal Medicine* Fauci et al. (1998)", em que se diz, correcta ou incorrectamente, que "*nesta era de medicina baseada na evidência é tentador pensar-se que muitos dos problemas de mais difícil decisão que surgem na prática clínica ou já estão, ou estão quase a ser completamente resolvidos e digeridos em 'guidelines'*",ou em auxiliares de memória facilmente acessíveis via os potentes motores de pesquisa existentes em '*sites*' da *internet*.

Muitas aplicações biomédicas parecem ter-se mostrado exigirem um constante desenvolvimento e investigação de novas tecnologias informáticas. Ao contrário de outras áreas os sistemas que processam dados biológicos e médicos têm tido muitas histórias de insucesso. Os primeiros programas de computadores parece que tiveram um bom desempenho em sistemas muito bem especificados, no caso extremo eram modelos físicos que se podiam descrever rigorosamente em termos matemáticos e quanto mais estes se afastavam desta característica pior se comportavam os sistemas. Acontece que a modelação algorítmica matematicamente muito rigorosa nem sempre é fácil de se fazer, particularmente em sistemas biomédicos que muitas vezes apenas parcialmente são compreendidos.

Uma das primeiras abordagens começada nos anos 1960 e 1970 foi a do reconhecimento de padrões, mais propriamente classificação de padrões, quer dizer, algoritmos que permitiam o computador procurar determinados invariantes ou padrões de dados, os quais podiam tratar-se de imagens ou de grupos de parâmetros associados a doenças específicas.

Com o objectivo de se obstar a alguns inconvenientes inerentes à classificação de padrões, como por exemplo a característica de "*caixa-negra*" que alguns algoritmos apresentavam uma vez que os médicos apenas acediam aos resultados e o sistema não lhes dava pistas sobre o raciocínio subjacente, surgiram nos anos 1980 sistemas periciais com uma abordagem baseada na então designada "*inteligência artificial*". O mais famoso destes sistemas foi o *MYCIN* Shortliffe (1976) que, no entanto, apresentava tantas dificuldades com a representação do chamado conhecimento de

senso comum que se verificava uma efectiva incapacidade do sistema, quer de generalização, quer de adaptação à mais simples modificação dos valores de entrada. Para além destas dificuldades surgia ainda um outro problema devido à elevada complexidade do registo de dados médicos. Este registo é inerentemente complicado porque compreende muitas componentes diversificadas, desde imagens obtidas por radiografia, tomografia computacional, ressonância magnética ou ecografia, até notas manuscritas, muitas vezes à pressa, ou sob a pressão dos acontecimentos, passando por resultados de testes quantitativos, sinais impressos no electrocardiógafo, ou no electroencefalógrafo.

Muitas das aproximações tradicionais que recorriam a uma descrição algorítmica de um dado sistema, com vista ao apoio da decisão médica, apresentavam-se como insucessos, mais tarde ou mais cedo, porque em muitos casos os modelos completos que descreviam o funcionamento dos sistemas biológicos ou não eram conhecidos, ou não eram determinísticos, ou ainda, em muitos casos em que os sistemas procuravam classificar os dados, acontecia que a precisão dada pelos sistemas era inferior à requerida, ou então, outras vezes, os algoritmos estocásticos não eram aplicáveis às populações reais em que os sistemas operavam.

Uma outra fonte de dificuldade podia surgir na diversidade, ou insuficiências, do corpo de especialistas em que se baseava a engenharia de extracção de conhecimentos subjacente à concepção dos sistemas ou, então, ao permanente debate entre aproximações simbólicas baseadas em conhecimento de peritos e aproximações comandadas pelos dados (*data-driven*), ou ainda aproximações híbridas.

Passando agora a um outro aspecto, o da utilização de estatísticas em investigação médica, há que referir a existência também de muita controvérsia, uma vez que, frequentemente, se ouve dizer que a informação estatística praticamente pouco contributo dá para o progresso da medicina, porque, como referem Armitage and Berry (1987), "*a verdadeira preocupação do médico é o tratamento de um doente em particular, e todo o doente difere em aspectos muito importantes de outros doentes.*" É claro que a variabilidade de uma doença não deveria ser um argumento em desfavor da informação estatística mas, antes pelo contrário, em favor desta. O problema parece residir em saber-se quando a informação estatística deve ser memorizada informalmente na mente do médico assistente ou se, pelo contrário, quando ela deve ser coligida e relatada de uma maneira sistemática.

Para muita pessoas os métodos matemáticos utilizados pelo estatístico profissional são, frequentemente, encarados como "obscuros e até mesmo repelentes". Uma das formas de minimizar esta tendência assim como encarar as aplicações de métodos numéricos em medicina como uma "*trivialiadade e perda de tempo*" reside, na opinião de Sir Austin Bradford Hill (1983), em fazer com que os médicos, e para-médicos, tenham uma aprendizagem suficiente e adequada em métodos es-

tatísticos. Numa atitude de optimismo (talvez de sabor amargo para quem trabalha em estatística) acrescentava, este reconhecido especialista em estatística médica e professor na Universidade de Londres:

> "Os métodos estatísticos quer sejam encarados como a melhor forma para casos particulares, quer sejam essenciais, ou então, quer sejam vistos como uma elaboração desnecessária, a questão que se coloca é a de saber se, em qualquer dos problemas de estatística médica, os resultados satisfatórios se poderão obter por intermédio apenas de métodos numéricos relativamente simples. Por outra palavras, poder-se-á fazer testes de hipóteses satisfatoriamente, e fazer deduções a partir de um conjunto de dados se estes forem analisados por intermédio de tais métodos simples? A resposta é indubitávelmente que sim, empregando métodos estatísticos relativamente simples muitos números que aparecem citados em trabalhos médicos poderiam conter uma informação de muito maior valor Hill (1983)."

Ao longo dos muitos contactos havidos entre o autor desta tese e a equipa de médicos obstetras que participaram neste projecto, e contribuíram decisivamente para o seu bom andamento, transpareceu uma imagem da problemática da decisão médica em relação à qual se advogava uma necessidade de experiência, associada a capacidades de raciocínio excelentes, complementada por ferramentas quantitativas bem concebidas e, muito em especial, uma penetrante avaliação das particularidades da doente (a parturiente na generalidade das situações em obstetrícia).

A pesquisa para a discussão do ponto de vista médico que se tentou sumariamente aqui apresentar, em relação aos sistemas computacionais, recorreu à longa experiência e maturidade nestas matérias que é relatada em Hudson and Cohern (2000), Goldman (1998) e, para a estatística em medicina, citaram-se algumas opiniões consagradas de especialistas internacionalmente reconhecidos em que se destacam Armitage and Berry (1987) e Hill (1983).

7.3.2 Estimador por rede neuronal baseado no Windows

Com o objectivo de apresentar de uma forma simples os resultados da investigação desta tese concebeu-se um calculador que permite determinar o Peso Fetal Estimado (PFE) a partir de medidas ecográficas. e com as seguintes especificações:

- Entrada de dados em milímetros apresenta os pesos estimados em gramas.
- Aceita dados omissos e rejeita outliers.

- Mostra o valor do erro absoluto e permite explorar a influência da variabilidade das entradas.

Os dados a serem introduzidos são constituídos por medidas em milímetros do perímetro abdominal, perímetro cefálico, comprimento do fémur e diâmetro biparietal, mais o índice de resistência umbilical (ou de Pourcelot).

Para além da estimativa do PFE, em gramas, depois de introduzir o Peso Fetal à Nascença (PFN) o utilitário determina o erro relativo, definido pela razão entre o erro simples ($PFE - PFN$) e o peso fetal à nascença (PFN).

Os dois primeiras valores da estimativa do PFE são determinados segundo as fórmulas de Hadlock e Shepard, referidas em Farmer et al. (1992). Os dois últimos valores do PFE são determinados por modelos de rede neuronal artificial, segundo duas arquitecturas diferentes, uma de perceptrão multicamada e a outra em rede de funções de base radiais (Figura 7.1).

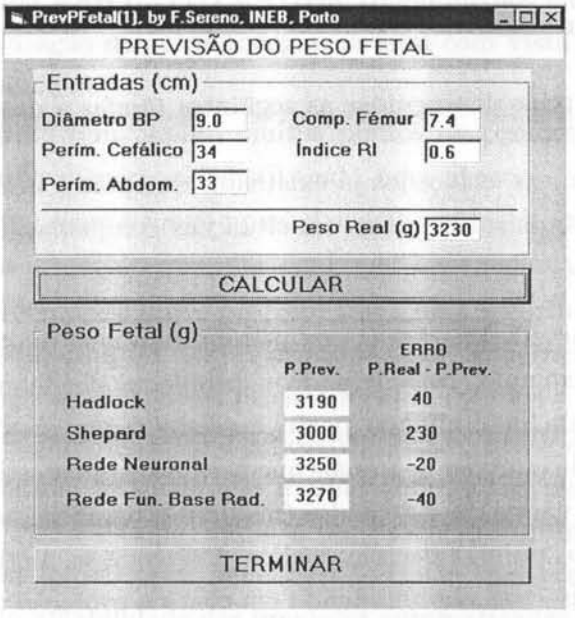


Figura 7.1: Estimador de rede neuronal em Windows

7.3.3 Perspectivas de desenvolvimento no futuro: pesquisa por conteúdo e *data mining*

A colheita de *dados* parece nalguns casos apresentar algumas inconsistências nos valores apresentados as quais se poderiam tentar reduzir introduzindo melhoramentos nas especificações do protocolo. Este poderia conter mais informações, como por exemplo, as medidas ecográficas às 12, 22, 32 semanas, sexo do feto, dados sobre a

BIBLIOTECA F.E. PORTO

mãe respeitantes a diabetes se é fumadora. Em particular salienta-se as vantagens de se estabelecer um índice de qualidade Sereno and Marques_de_Sá (2000)

Um *projecto de extracção de informação obstétrica*, pesquisa por conteúdo, e data mining em bases de dados multimédia de obstetrícia poderia ter, entre outros a serem introduzidos por sugestão dos interessados, os seguintes *objectivos*:

- Concepção e desenvolvimento de uma base de dados multimédia de obstetrícia para registos de grávidas com exames suplementares de imagens ecográficas, vídeo clipes, electrocardiogramas, etc.

- Desenvolvimento e utilização de ferramentas para pesquisa por conteúdo da base de dados multimédia de obstetrícia para ajuda ao diagnóstico assim como para apoio à investigação e ensino.

- Extracção de dados (data mining) com o objectivo da descoberta de correlações (regras) na base de dados. Como ferramenta de data mining pretende-se utilizar redes neuronais.

Para este efeito pode distinguir-se as seguintes *tarefas* a desenvolver

- Identificar técnicas avançadas já existentes de apoio à decisão em obstetrícia, procurando minimizar os riscos em situações complexas de cuidados médicos em obstetrícia.

- Descobrir características novas que possibilitem uma maior precisão das previsões. Por exemplo, considerando o problema da determinação das complicações de um determinado parto baseado numa previsão do peso fetal, pretende-se utilizar redes neuronais que permitam obter estimativas mais precisas a partir de vários tipos de dados (medidas ecográficas, observações clínicas, etc.), apoiando, assim, a decisão médica. Pretende-se, ainda, que a maior precisão das previsões se obtenha tendo em conta a evolução temporal dos dados

- Optimizar as decisões e não apenas previsões. Por exemplo, mesmo sabendo que é uma grande ajuda poder-se prever quais são as mulheres que estão em condição de risco de complicações de parto, seria muito mais útil aprender quais as acções que se poderiam tomar com antecedência para reduzir esse risco.

Notar que o verdadeiro ponto de partida na análise de um *projecto de extracção de informação (obstétrica)*, em lugar de começar pela recolha dos dados, deveria começar pela identificação de uma necessidade: (1) Seleccionar claramente os problemas obstétricos com benefícios tangíveis; (2) Especificar quais as soluções que se

BIBLIOTECA F.E. PORTO

pretende; (3) Definir como as soluções devem ser implementadas; (4). Ter uma boa compreensão do domínio obstétrico; (5) Estipular claramente quais são os pressupostos e suposições elementares; (6) Deixar que seja o problema a conduzir a modelação; (7) Refinar o modelo iterativamente; (8) Tornar o modelo tão simples quanto possível (mas não o "*simplificar*" demais); (9) Procurar áreas de instabilidade; (10) Procurar áreas de incerteza

O objectivo central do projecto é o desenvolvimento de técnicas avançadas de apoio à decisão em obstetrícia, procurando minimizar os riscos em situações complexas de cuidados médicos em obstetrícia, com reflexos nos índices de morbilidade e mortalidade infantil.

O projecto podia integrar e beneficiar da experiência adquirida com o desenvolvimento de ferramentas de apoio à decisão médica em obstetrícia e ginecologia, nomeadamente o Sistema SisPorto de análise automática de cardiotocogramas fetais o projecto $INTERCARE$ de telemedicina envolvendo também a cardiotocografia fetal; projecto europeu $LAHYSTOTRAIN$de treino médico em laparoscopia e histeroscopia e a modelização do cardiotocograma fetal com vista à sua integração no "*Human Patient Simulator*" .

O presente projecto tem, assim, muitos pontos de contacto com a actividade do grupos de trabalho dispersos, constituindo uma extensão natural e um passo importante para o progresso a longo prazo das actividades de investigação do grupo em Obstetrícia e sua projecção quer nacional quer internacional.

A longo prazo pretende-se que o projecto traga uma contribuição importante na organização e gestão informática dos serviços de Obstetrícia, bem como na utilização de ferramentas avançadas de apoio à decisão médica.

Referindo agora alguns aspectos das possíveis *repercussões éticas , sociais e ambientais,* é de salientar o facto de que mesmo abordagens muito preliminares de extracção de dados (*data mining*) estão a ser usadas por rotina em muitas organizações e produzem importantes benefícios. O principal problema ético que se coloca tem a ver com a fiabilidade das previsões automatizadas, com o grau de confiança que o médico (o decisor) deverá atribuir às " recomendações" da máquina. O problema da confidencialidade e das suas relações com os aspectos administrativos e jurídicos da responsabilidade médica ultrapassam o âmbito do presente projecto, mas a existência das ferramentas que o projecto pode gerar pode ser um contributo para permitir mais segurança e uma diminuição quer de erros, quer de negligências no acto médico.

No que diz respeito à *relação com o estado da arte* este projecto tem a limitação de apenas se ter limitado a uma pesquisa das possíveis sinergias que se poderiam estabelecer com uma iniciativa surgida em 2000 no *Center for Automated Learning and Discovery (CALD)*, na Universidade de Carnegie Mellon, em Pittsburg , U.S.A.

pretende; (3) Definir como as soluções devem ser implementadas; (4). Ter uma boa compreensão do domínio obstétrico; (5) Estipular claramente quais são os pressupostos e suposições elementares; (6) Deixar que seja o problema a conduzir a modelação; (7) Refinar o modelo iterativamente; (8) Tornar o modelo tão simples quanto possível (mas não o "*simplificar*" demais); (9) Procurar áreas de instabilidade; (10) Procurar áreas de incerteza

O objectivo central do projecto é o desenvolvimento de técnicas avançadas de apoio à decisão em obstetrícia, procurando minimizar os riscos em situações complexas de cuidados médicos em obstetrícia, com reflexos nos índices de morbilidade e mortalidade infantil.

O projecto podia integrar e beneficiar da experiência adquirida com o desenvolvimento de ferramentas de apoio à decisão médica em obstetrícia e ginecologia, nomeadamente o Sistema SisPorto de análise automática de cardiotocogramas fetais o projecto $INTERCARE$ de telemedicina envolvendo também a cardiotocografia fetal; projecto europeu $LAHYSTOTRAIN$de treino médico em laparoscopia e histeroscopia e a modelização do cardiotocograma fetal com vista à sua integração no "*Human Patient Simulator*" .

O presente projecto tem, assim, muitos pontos de contacto com a actividade do grupos de trabalho dispersos, constituindo uma extensão natural e um passo importante para o progresso a longo prazo das actividades de investigação do grupo em Obstetrícia e sua projecção quer nacional quer internacional.

A longo prazo pretende-se que o projecto traga uma contribuição importante na organização e gestão informática dos serviços de Obstetrícia, bem como na utilização de ferramentas avançadas de apoio à decisão médica.

Referindo agora alguns aspectos das possíveis *repercussões éticas , sociais e ambientais,* é de salientar o facto de que mesmo abordagens muito preliminares de extracção de dados (*data mining*) estão a ser usadas por rotina em muitas organizações e produzem importantes benefícios. O principal problema ético que se coloca tem a ver com a fiabilidade das previsões automatizadas, com o grau de confiança que o médico (o decisor) deverá atribuir às " recomendações" da máquina. O problema da confidencialidade e das suas relações com os aspectos administrativos e jurídicos da responsabilidade médica ultrapassam o âmbito do presente projecto, mas a existência das ferramentas que o projecto pode gerar pode ser um contributo para permitir mais segurança e uma diminuição quer de erros, quer de negligências no acto médico.

No que diz respeito à *relação com o estado da arte* este projecto tem a limitação de apenas se ter limitado a uma pesquisa das possíveis sinergias que se poderiam estabelecer com uma iniciativa surgida em 2000 no *Center for Automated Learning and Discovery (CALD)*, na Universidade de Carnegie Mellon, em Pittsburg , U.S.A.

Este centro inclui uma área de "*Descoberta e Extracção de Conhecimento*" (*KDD - Knowledge Discovery and Data Mining*) que oferecia uma educação rica e interdisciplinar, através da combinação de tópicos tais como algoritmos de aprendizagem automatizada avançada e os princípios estatísticos que lhes estão subjacentes, as metodologias das bases de dados e "*data warehousing*", a análise da complexidade, as aproximações à visualisação de dados, as questões da privacidade e da segurança, em aplicações específicas tais como cuidados de saúde, negócios, marketing e gestão pública.

Alguns projectos a serem desenvolvidos no *CALD* envolvem a aprendizagem da previsão de riscos que os doentes recém chegados correm com base na extracção de dados nos registos médicos do passado. Na opinião de Tom Mitchell, Director do *CALD*, a aprendizagem automatizada, a extracção de dados e a estatística computacional estão a atrair um grande interesse comercial, na medida em que cada vez é maior o volume de dados disponível, é mais acessível recolher dados graças às redes, e a aprendizagem automatizada como investigação básica está mais próxima da sua maturidade. O presente projecto (para mais detalhes Sereno and Marques_de_Sá (2000)) situava-se na linha de investigação "*Searching by Content and Data Mining in Multimedia Databases*", no campo da Aprendizagem Automatizada e Descoberta.

7.4 Conclusões

O objectivo principal desta dissertação relaciona-se com o desenvolvimento de *um método de estimação* do peso do recém-nascido antes do nascimento, baseado num modelo de rede neuronal que processasse medidas biométricas extraídas pela técnica ecográfica fetal, e que permitisse obter estimativas do peso fetal com mais precisão do que era possível com os *modelos de regressão linear generalizados.*

O problema foi investigado através de estudos teóricos e experimentais. Primeiramente, fez-se uma *análise exploratória aos dados* recolhidos em quatro hospitais portugueses e procurou-se obter um vasto conhecimento a priori sobre o problema. Procurou-se ainda identificar as causas pelas quais os vários modelos convencionais de regressão, multilineares, generalizados ou não-lineares, apresentavam mais dificuldades em fazer uma estimação dentro dos níveis precisão desejáveis. Em segundo lugar procedeu-se a estudos teóricos sobre *a teoria da aprendizagem estatística* aplicada às redes neuronais, e adquiriu-se experiência prática de modelização com as ferramentas de redes neuronais do *MATLAB*. Estabeleceram-se, nesta fase, os primeiros modelos baseados no perceptrão multicamada e em funções de base radiais, os quais permitiram concluir que as redes neuronais permitiam uma melhoria dos erros relativos comparativamente com as fórmulas de regressão. Estes modelos de rede neuronal utilizam um método de optimização de gradiente, ou outras varian-

tes deste mais eficazes, mas que não permitiam saber se a convergência tinha sido para uma solução óptima local, ou para a solução óptima global (caso exista). Os modelos de *máquina de vectores de suporte* apresentavam, por seu lado, a vantagem teórica de utilizarem uma optimização por técnicas de programação matemática (nomeadamente a programação quadrática). Iniciou-se assim uma nova fase de estudos teóricos destas técnicas, seguindo-se uma extensa experimentação neste novo paradigma de máquina de *kernel*, o qual do sempre que os dados permitissem um bom condicionamento do problema, o algoritmo convergia sempre para uma e uma só solução, permitindo assim conhecer com firmeza uma resposta definitiva. Os resultados das experiências com um modelo de máquina de vector de suporte numa implementação do *algoritmo da decomposição*, em linguagem C, as quais permitem fazer o treino com um número razoável de exemplos em tempos computacionalmente aceitáveis, mostraram que este paradigma permite uma diminuição dos erros relativos. Finalmente, fizeram-se estudos teóricos sobre a modelação de soluções cooperativas inspirada na *teoria da informação*. Nesta última fase pretendia-se estabelecer as bases para uma combinação de modelos que levasse em linha de conta as especificidades dos dados e uma cooperação entre modelos de estimação alternativos. Nas experiências com modelos neuronais , como se referiu, estavam a ser desprezados casos porque se pensava estarem contaminados com imprecisões de origem diversa nas medidas ecográficas. Mas, de facto, muitos desses casos poderiam conter *informações* com graus de incerteza variável e poderia ser utilizados na aprendizagem. Se esta utilizasse uma técnica baseada na informação mútua e entropia dos erros e, adicionalmente, se introduzisse um postprocessamento com lógicas não clássicas, como por exemplo a lógica *fuzzy*, deveria ser possível ainda uma redução dos erros relativos.

Obtiveram-se os seguintes resultados:

- Contribuições para a medicina
  - Redução do erro relativo de estimação do peso fetal para cerca de $\pm 7\%$ quando os valores admissivéis na prática clínica para este erro são, actualmente, da ordem de grandeza $\pm 15\%$.
  - Estabelecimento de um modelo de rede neuronal de perceptrão de multicamada, PMC, que apresentou um erro relativo médio de estimação do peso fetal de $\pm 7.17\%$.
  - Estabelecimento de um modelo de rede neuronal de funções de base radiais, FBR, que apresentou um erro relativo médio de estimação do peso fetal de $\pm 6.95\%$.
  - Estabelecimento de um modelo de máquina de vectores de suporte, MVS,

que apresentou um erro relativo médio de estimação do peso fetal de ±6.71%.

- Concepção e realização de um calculador do peso fetal em $Visual-Basic$ que permite testar na prática clínica os modelos estabelecidos.

- Contribuições para a ciência da informação
  - Concepção e implementação de um método de edição e pré-processamento de dados por sobre-ajustamento de uma máquina de kernel aos exemplos e detecção dos casos desviantes quando estes mesmos exemplos são re-processados pela máquina.
  - Interpretação na perspectiva da teoria da aprendizagem, incluindo tradução original para língua portuguesa, dos princípios de funcionamento, propriedades e algoritmos dos modelos baseados na máquina de vectores de suporte e de rede neuronal.

- Contribuições para o desenvolvimento profissional em didáctica da matemática
  - Interpretação e tradução original para língua portuguesa dos principais conceitos da teoria estatística da aprendizagem aplicados à aprendizagem supervisada de funções reais.

Os resultados estão distribuídos ao longo da dissertação da seguinte maneira.

O presente capítulo sétimo contém referências às perspectivas de desenvolvimento, trabalho a fazer-se no futuro e às tentativas de disseminação do presente trabalho como por exemplo uma aplicação para *Windows* realizada em $Visual-Basic$.

O esboço do projecto de um estimador combinatório e cooperativo de três modelos estimadores encontra-se referido no Capítulo 6.

As experiências conducentes à concepção e experiências com os modelos estimadores de rede neuronal PMC ou FBR e máquina kernel, MVS, encontram-se descritos no Capítulo 5.

A descrição, concepção e implementação do método de pré-processamento e edição de dados, assim como os resultados práticos obtidos estão descritos na subsecção 5.3.3 do Capítulo 5.

A análise exploratória dos dados colhidos em quatro serviços de obstetrícia de hospitais portugueses numa perspectiva clássica e com base numa modelação com regressores multilineares encontra-se descrita no Capítulo 4.

Uma narrativa sobre o funcionamento, propriedades e algoritmos da máquina de vectores de suporte e uma interpretação e tradução original para língua portuguesa

sobre aspectos teóricos do modelo de rede neuronal, na perspectiva da teoria da aprendizagem, encontra-se desenvolvida no Capítulo 3.

Uma revisão, interpretação e tradução original para língua portuguesa dos principais conceitos da teoria estatística da aprendizagem aplicados à aprendizagem supervisada de funções reais encontram-se descritos no Capítulo 2.

O trabalho descrito nesta dissertação pode ser continuado no futuro em várias direcções diferentes.

- Desenvolvimento de um novo método de extracção automática de medidas de estruturas anatómicas fetais a partir de imagens ecográficas, actualmente em curso no Instituto Superior Técnico de Lisboa Jardim and Figueiredo (2002).

- Concepção e desenvolvimento de uma base de dados multimédia de obstetrícia, que permita o registo de exames suplementares contendo imagens ecográficas, vídeo clipes, electrocardiogramas, e muitos outros dados temporais, com a possibilidade de utilizar ferramentas para pesquisa por conteúdo, e extracção de dados (*data mining*), tendo por objectivos a descoberta de correlações (regras) que possam ajudar o diagnóstico, assim como dar apoio à investigação e ao ensino, recentemente sugerido em Guan et al. (2002).

- Modelação da estimação por critério entrópico, baseando-se na teoria da informação e em lógicas não clássicas, tendo em vista a obtenção de um único estimador combinatório e cooperativo de vários modelos estimadores, baseados em máquinas de *kernel*, tendo em consideração os recentes desenvolvimentos neste paradigma apresentados em Scholkopf and Smola (2002) e os estimadores de entropia de Renyi referidos em Erdogmus and Príncipe (2002).

A utilização de algoritmos de aprendizagem em problemas de regressão e classificação que se relacionem com tarefas concretas da prática clínica real, tal como a estimação do peso fetal em obstetrícia, tem de ter em linha de conta a existência de uma multiplicidade de soluções candidatas a serem ponderadas, devendo sempre procurar-se a solução mais adequada em termos de previsão da sua capacidade de generalização relativamente à referida tarefa.

Apresentou-se nesta dissertação, tendo-se procurado uma inspiração nos mesmos princípios matemáticos e estatísticos que permitiram no passado o estabelecimento de melhores estimadores de regressão multilinear generalizada, um estudo do problema da generalização de modelos computacionais de estimação e uma investigação de técnicas de rede neuronal com perceptrão multicamada. ou funções de base radiais, assim como de técnicas de kernel com máquinas de vectores de suporte, que se acredita permitirem responder às necessidades da prática clínica obstétrica de

BIBLIOTECA FE PORTO

diminuição das situações de risco, particularmente, no caso de a tarefa a resolver ser uma estimação mais rigorosa do peso fetal.

BIBLIOTECA F.E. PORTO

Bibliografia

Abrantes, S. A. (1998). Apontamentos de teoria da informação. Technical report, Faculdade de Engenharia da Universidade do Porto.

Abrantes, S. A. (2000). *Processamento Adaptativo de Sinais.* Fundação Calouste Gulbenkian, Lisboa.

Adler, M. (1997). An introduction to pattern recognition statistical, neural net and syntactic methods of getting robots to see and hear. Technical report, Centre for Intelligent Information Processing Systems. Department of Electrical and Electronic Engineering. The University of Western Australia.

Almeida, L. B. (1987). A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *First International Conference on Neural Networks*, pages 609–618, San Diego CA.

Almeida, L. B., Langlois, T., and Amaral, J. D. (1997). On-line step size adaptation. Technical Report RT07/97, Instituto Superior Técnico.

Anthony, M. (1997). Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. *Neural Computing Surveys*, 1:1–47.

Anthony, M. (1998). Probabilistic 'generalization' of functions and dimension-based uniform convergence results. *Statistics and Computing*, 8(1):5–14.

Anthony, M. and Bartlett, P. (1999). *Neural Network Learning: Theoretical Foundations.* Cambridge University Press.

Anthony, M. and Holden, S. B. (1998). Cross-validation for binary classification by real-valued functions. In *Proceedings of the 1998 Annual Conference on Computational Learning Theory.(COLT98), Madison*, pages 1–29, Madison. ACM Press.

Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research.* Blackwell Scientific Publications, Oxford.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404.

Attneave, F. (1953). Symmetry, information and memory patterns. *Amer. J. Psych*, 68:209–222.

Bartlett, P. L., Long, P., and Williamson, R. C. (1996). Fat–Shattering and the Learnability of Real–Valued Functions. *Journal of Computer and System Sciences*, 52(3):434–452.

Baum, E. B. (1991). Neural net algorithms that learn in polynomial time from exxamples and queries. *IEEE Trans. On Neural Networks*, 2(1):5–19.

Bennett, K. P. and Bredensteiner, E. J. (1998). Geometry in learning. In C. Gorini, E. Hart, W. Meyer and Phillips, T., editors, *Geometry at Work*, New York. Mathematical Association of America.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis (Second Edition)*. Springer, New York.

Billingsley, P. (1995). *Probability and Measure*. Wiley, New York.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Haussler, D., editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA. ACM Press.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. In Scholkopf, B., Smola, A., and Burges, C., editors, *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers.

Burges, C. J. C. (1999). Geometry and invariance in kernel based methods. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods: Support Vector Learning*. MIT Press.

Campani, C. A. P. and Menezes, P. B. (2002). Introdução À complexidade de kolmogorov. Technical report, Universidade Federal de Pelotas, Campinas, Brasil.

Campbell, C. (2000). An introducton to kernel methods. In Howlett, R. J. and Jain, L. C., editors, *Radial Basis Function Networks: Design and Applications*, Berlin. Springer Verlag.

Cardoso, J. F. (2001). *Predição da qualidade na indústria de fundição injectada*. PhD thesis, Faculdade de Engenharia da Universidade do Porto.

Chaitin, G. J. (1974). Information theoretic computational complexity. *IEEE Trans. on Information Theory*, 20:10–15.

Chaitin, G. J. (1990). *Information Randomness and Incompleteness, Papers on Algorithmic Information Theory (Second Edition)*. World Scientific, New Jersey.

Chatfield, C. and Collins, A. (1989). *Introduction to Multivariate Analysis*. Chapman & Hall, London.

Chauhan, S. P., Charania, S. F., McLaren, R. A., Devoe, L. D., Ross, E. L., Hendrix, N. W., and Morrison, J. C. (1998). Ultrasonographic estimate of birth weight at 24 to 34 weeks: A multicenter study. *Am. J. Obstetrics and Gynecology*, 179:909–916.

Cherkassky, V. and Mulier, F. (1998). *Learning From Data - Concepts, Theory, and Methods*. John Wiley & Sons Inc, New York.

Cherkassky, V., Xuhui, S., Mulie, F. M., and Vapnik, V. N. (1999). Model complexity control for regression using VC generalization bounds. *Neural Networks, IEEE Transactions on*, 10(5):1075–1089.

Cibas, T. (1996). *Contrôle de la Complexité Dans Les Réseux de Neuronnes: Régularisation et Sélection de Caractéristiques*. PhD thesis, Université de Paris-Sud, U.F.R. Scientifique D'Orsay.

Cibas, T., Gallinari, P., and Gascuel, O. (1995). Experimental investigation on the complexity-performance relations in multilayer perceptrons. In *International Conference ICANN-95*, Paris.

Clemons, T. and Pagano, M. (1999). Are babies normal? *The American Statistician*, 53(4):298–302.

Collobert, R. and Bengio, S. (2001). SVMTorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160.

Collobert, R., Bengio, S., and Mariethoz, J. (2002). A small user guide for SVMTorch II. Technical report, IDIAP, Switzerland, Université de Montréal, Québec.

Copland, B. J. and Proudfoot, D. (1999). Alan turing's forgotten ideas in computer science. *Scientific American*, 280(4):77–81.

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20:273 – 297.

Courant, R. and Hilbert, D. (1953). *Methods of Mathematical Physics*, volume 1. Interscience Publishers, Inc, New York.

Cover, J. A. T. T. M. (1991). *Elements of Information Theory.* John Wiley & Sons, Inc., New York.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comp.*, 10:326–334.

Cox, D. R. (1992). *Applied Statistics Principles and Examples.* Chapman and Hall, London.

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, 14:1–13.

Cristianini, N. (2001). Support vector and kernel machines (A tutorial). In *NIPS 2001 Conference.*

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge.

Cube, F. V. (1960). Uber informationstheoretische problem in lerntheorie und didaktik. *GrKG*, 1:105–112.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39(1):1–49.

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

David, R. J. and Collins, J. W. (1997). Differing birth weight among infants of U.S.-born blacks, african-born blacks, and U.S. - born whites. *The New England Journal of Medicine*, 337(17):1209–1214.

Deco, G. and Obradovic, D. (1996). *An Information-Theoretic Approach to Neural Computing.* Springer, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1976). Maximum likelihood from incomplete data via de EM algorithm. *Journal of the Royal Statistical Society*, B39(1):1–38.

D'Ottavio, G. and Mandruzzato, G. P. (1998). Screening for intrauterine growth retardation: Principles and techniques of fetal ultrasound biometry. In Kurjak, A., editor, *Textbook of Perinatal Medicine - a Comprehensivce Guide to Modern Clinical Perinatology (Vol.2)*, pages 1212–16, London. The Parthenon Publishing Group Ltd.

Dubucs, J.-P. (1998). Logiques non classiques. In *Dictionnaire Des Mathématiques, Fondements, Probabilités, Applications*, pages 336–343, Paris. Albin Michel.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification (Second Edition)*. John Wiley & Sons Inc, New York.

Erdogmus, D. and Principe, J. C. (2001). Entropy minimization algorithm for multilayer perceptrons. In IEEE, editor, *Proceedings of IJCNN'01 International Joint Conference on Neural Networks*, pages 0–0, Washington, DC. IEEE.

Erdogmus, D. and Príncipe, J. C. (2002). Generalized information potential criterion for adaptive system training. *IEEE Trans. On Neural Networks*, 13(5):1035–1044.

Evgeniou, T. (2000). *Learning with Kernel Machine Architectures*. PhD thesis, Massachusetts Institute of Technology.

Evgeniou, T., Pontil, M., and Poggio, T. (2000a). Regularization networks and support vector machines. In Smola, A., Bartlett, P., Scholkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, Cambridge, MA. MIT Press.

Evgeniou, T., Pontil, M., and Poggio, T. (2000b). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50.

Fahlman, S. E. (1988). Faster-learning variations of back propagation. In Touretzky, D. S., editor, *Proceedings of the 1988 Connectionist Models Summer School*, San Mateo. Morgan Kaufman.

Farmer, R. M., Medearis, A. L., Hirata, G. I., and Platt, L. D. (1992). The use of a neural network for the ultrasonographic estimation of foetal weight in macrosomic fetus. *Am. J Obstet. Gynecol.*

Fauci, A., Braunwald, E., Inelbacher, K., Wilson, J., Martin, J., Kasper, D., Hauser, S., and Longo, D. (1998). *Harrison's Principles of Internal Medicine (14th Edition)*. McGraw Hill, New York.

Fisher, J. W., Ihler, A. T., and Viola, P. A. (1999). Learning informative statistics: A nonparametric approach. *Advances in Neural Information Processing Systems*.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Society*, 22:700–725.

Fisher, R. A. (1973). *Statistitical Methods for Research Workers (14th Edition)*. Hafner Publishing Company, New York.

Fletcher, R. (2000). *Practical Methods of Optimization (Second Edition)*. John Wiley, New York.

Frank, H. (1960). Uber eine informationspsychologische massbestimmung der semantischen und pragmatischen information. *GrKG*, 1:25–32.

Freund, Y. and Schapire, R. E. (1998). Large margin classification using the perceptron algorithm. In Shavlik, J., editor, *Machine Learning: Proceedings of the Fifteenth International Conference*, San Francisco. Morgan Kaufmann.

Friedman, J. H. (1997). On bias , variance, 0/1-loss and curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

Gallinari, P. (1997). Heuristiques pour la Généralisation. In Thiria, S., Lechevalier, Y., Gascuel, O., and Canu, S., editors, *Statistiqus e Réseaux Neuronales*, pages 230–243, Paris. Dunod.

Gascuel, O. (1997). La dimension de vapnik-chervonenkis. In Thiria, S., Lechevalier, Y., Gascuel, O., and Canu, S., editors, *Statistiques e Réseaux Neuronales*, pages 244–274, Paris. Dunod.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks. *Neural Computation*, 7(2):219–269.

Goldman, L. (1998). Quantitative aspects of clinical reasoning. In Fauci, A., Braunwald, E., Inelbacher, K., Wilson, J., J.Martin, Kasper, D., Hauser, S., and Longo, D., editors, *Harrison's Principles of Internal Medicine*, pages 9–14, New York. McGraw Hill.

Gray, R. M. (1990). *Entrropy and Information Theory*. Springer-Verlag, New York.

Guan, L., Adali, T., Katagiri, S., Larsen, J., and Principe, J. (2002). Guest editorial – special issue on intelligent multimedia processing. *IEEE Transactions on Neural Networks*, 13(4):789–90.

Gunn, S. (1998). Support vector machines for classification and regression. Technical report, Image Speech and Intelligent Systems Group, University of Southampton.

Guyon, I., Boser, B., and Vapnik, V. (1993). Automatic capacity tuning of very large VC-dimension classifiers. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems*, volume 5, pages 147–155. Morgan Kaufmann, San Mateo, CA.

Hadlock, F. P., Harrist, R. B., and Sharman, R. S. (1985). Estimation of fetal weight with the use of head, body, and femur measurements- a prospective study. *AM J Obstet Gynecol*, 151:333–7.

Hagan, M. T., Demuth, H. B., and Beale, M. (1996). *Neural Network Design.* PWS Publishing Company, New York.

Hagan, M. T. and Menhaj, M. (1994). Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993.

Hartley, R. V. (1928). Transmission of information. *Bell Sys. Tech. Journal*, 7:535.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Transactions on Information Theory*, 43(4):1276–1280.

Haussler, D., Kearns, M., and Schapire. R. (1994a). Bounds on the sample complexity of bayesian learning using information theory and the VC dimension. *Machine Learning*, 14(1):83–114.

Haussler, D., Kearns, M., Seung, H. S., and Tishbi, N. (1994b). Rigorous learning curve bounds from statistical mechanics. In *Proceedings of the Seventh ACM Conference on Computational Learning Theory, (COLT)*, New York. ACM Press.

Haussler, D. and Opper, M. (1997a). Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*, 25(6).

Haussler, D. and Opper, M. (1997b). Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*, 25(6).

Haykin, S. (1999). *Neural Networks - A Comprehensive Foundation (Second Edition).* Prentice Hall, New York.

Hebb, D. (1949). *The Organization of Behaviour.* Wiley, New York.

Herrero, R. L. (1999). Estimated fetal weight. maternal vs physician estimate. MEDLINE, NLM Identifier 99413034.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation.* Addison-Wesley, Redwood City–California.

Hill, A. B. (1983). *A Short Textbook of Medical Statistics.* Hodder and Stoughton, London.

Hinton, G. and Ghahramani, Z. (2000). Unsupervised learning (fall 2000). Technical report, Gatsby Computational Neuroscience Unit – University College London.

Hudson, D. L. and Cohern, M. E. (2000). *Neural Networks and Artificial Intelligence for Biomedical Engineering.* IEEE Press in Biomedical Engineering, New York.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1(4):295–307.

Jardim, A. V. B. and Figueiredo, M. A. T. (2002). Automatic analysis of fetal echographic images. In Vaz, F., Santos, B. S., Pinho, A., Vieira, J., and Bastos, C., editors, *Proceedings of 12th Portuguese Conference on Pattern Recognition*, Aveiro. APRP.

Jaynes, E. T. (1978). Where we stand on maximum entropy? In Levine, R. D. and Tribus, M., editors, *The Maximum Entropy Formalism*, Cambridge. The MIT Press.

Jaynes, E. T. (1994). Probability theory : The logic of science. Technical report, Washington University, St. Louis.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press, Oxford.

Joachims, T. (1999). Making large–scale SVM learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods — Support Vector Learning*, pages 169–184, Cambridge, MA. MIT Press.

Kapur, J. N. (1993). *Maximum Entropy in Engineering and Science*. John Wiley and Sons, New York.

Kecman, V. (2001). *Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Models*. MIT Press, Boston.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information, problems of information transmission. *Problems of Information Transmission*, 1:4–7.

Kolmogorov, A. (1968). Logical basis for information theory and probability theory. *IEEE Trans. on Information Theory*, 14(5):662–664.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability (Second English Edition)*. Chelsea Publishing Company, New York.

Lengellé, R. and Schaltenbraud, N. (1997). Détermination de L'architecture des Réseaux multicouches. In Thiria, S., Lechevalier, Y., Gascuel, O., and Canu, S., editors, *Statistique et Méthodes Neuronales*, pages 53–74, Paris. Dunod.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27:986–1005.

Linsker, R. (1988). Self-organization in perceptual network. *Computer*, March:105–117.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Mag.*, 4:4–22.

Lisboa, P. J. G. (2002). A review of evidence of health benifit from artificial neural networks in medical intervevntion. *Neural Networks*, 15:11–39.

Luenberger, D. G. (1989). *Linear and Nonlinear Programming (Second Edition)*. Addison-Wesley Publishing Company, Menlo Park.

MacKay, D. J. C. (1992a). *Bayesian Methods for Adaptative Models*. PhD thesis, California Institute of Techology.

MacKay, D. J. C. (1992b). The evidence framework applied to classification networks. *Neural Computation*, 4:720–736.

MacKay, D. J. C. (1992c). A practical Bayesian framework for backprop networks. *Neural Computation*, 4:448–472.

Mackay, D. J. C. (2002). *Information Theory, Inference, and Learning Algorithms*. Author webpage, Cambridge UK.

Mandelbrot, B. (1999). A multifractal walk down wall street. *Scientific American*, 280(2).

Mandruzzato, G. P., Meir, Y. J., and Tamaro, L. F. (1994). Monitoring fetal hypoxemia: Doppler flow measurement and computarized cardiotocography. In Kurjak, A. and (Editor), F. A. C., editors, *The Fetus as a Patient: Advances in Diagnosis and Therapy*, pages 383–404, New York. CRC Press-Parthenon Publishers.

Mangasarian, O. L. (1994). *Nonlinear Programming*. SIAM, Philadelphia, PA.

Marquardt, D. W. (1963). An algorithm for least-squares estimation pf non-linear parameters. *Journal of the SIAM*, 11(2):431–441.

Marques, J. S. (1999). *Reconhecimento de Padrões, Métodos Estatísticos e Neuronais*. IST Press, Lisboa.

Marques_de_Sá, J. P. (2001a). *Pattern Recognition — Concepts, Methods and Applications*. Springer-Verlag, Berlin.

Marques_de_Sá, J. P. (2001b). The vapnik-chervonenkis dimension and the learning capability of neural nets. Technical Report INEB/FEUP-2001, INEB/Faculdade Engenharia do Porto.

Martinez, W. L. and Martinez, A. R. (2002). *Computational Statistics Handbook with MATLAB*. Chapman & Hall CRC, Boca Raton.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. Morgan Kaufman, San Diego.

McConnell, C. (1995). Minimal model complexity search. In Springer, editor, *Portuguese Conference on Artificial Intelligence*, pages 177–187, Lisbon.

McCullagh, P. and Nelder, J. A. (1997). *Generalized Linear Models.* Chapman and Hall, London.

McLachlan, J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons Inc, New York.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, A 209:415–446.

Micchelli, C. A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22.

Miller, G. A. (1958). The magical number seven plus or minus two. some limits on capacity for processing information. In *Reading in Perception*, New York.

Minsky, M. and Papert, S. (1969). *Perceptrons.* The MIT Press, Boston.

Mirchandi, G. and Cao, W. (1989). On hidden nodes for neural nets. *IEEE Tr. Circuits and Systems*, 36.

Mitchell, T. M. (1997). *Machine Learning.* McGraw Hill Book Co., New York.

Moore, D. S. and McCabe, G. P. (1999). *Introduction to the Practice of Statistics (Third Edition).* W. H. Freeman and Company, New York.

Murteira, B. (1993). *Análise Exploratória de Dados - Estatística Descritiva.* McGraw-Hill, Lisboa.

Murteira, B., Ribeiro, C. S., Silva, J. A., and Pimenta, C. (2001). *Introdução à Estatística.* McGraw-Hill, Lisboa.

Murteira, B. J. F. (1988). *Estatística: Inferência e Decisão.* Imprensa Nacional Casa da Moeda, Lisboa.

Nabney, I. (2002). *Netlab: Algorithms for Pattern Recognition.* Springer, London.

Neter, J., Kutner, M. K., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Regression Models (3d Edition).* IRWIN, Chicago.

Niyogi, P. and Girosi, F. (1996). On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8:819–842.

Novikoff, A. B. (1962). On convergence proofs on perceptrons. *Symposium on the Mathematical Theory of Automata*, 12:615–622.

Nyquis, H. (1924). Certain factors affecting telegraph speed. *Bell Sys. Tech. Journal*, 3:324.

Odone, F., Trucco, E., and Verri, A. (1998). Visual learning of weight from shape using support vector machines. In *British Machine Vision Conference.*

O'Sullivan, J. A., Blahut, R. E., and Snyder, D. L. (1998). Information-theoretic image formation. *IEEE Transactions on Information Theory*, 44:2094–2123.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: An application to face detection. In *Proc. Computer Vision and Pattern Recognition '97*, pages 130–136.

Pedroso, J. P. and Murata, N. (1999). Support vector machines for linear programming: Motivation and formulations. Technical Report 99-XXX, Riken Brain Science Institute.

Pereira-Leite, L. (2000). Reunião INEB-obstectras (2000-06-21). Notas pessoais de F.S.

Pereira-Leite, L. and Bernardes, J. (1998). Macrosomia: Antenatal diagnosis. In Kurjak, A., editor, *Textbook of Perinatal Medicine - a Comprehensivce Guide to Modern Clinical Perinatology (Vol.2)*, pages 1290–3, London. The Parthenon Publishing Group Ltd.

Pestana, D. D. and Velosa, S. F. (2002). *Introdução À Probabilidade e À Estatística (Volume I).* Fundação Calouste Gulbenkian, Lisboa.

Plaut, D. C., Nowlan, S. J., and Hinton, G. E. (1986). Experiments on learning by back propagation. Technical Report CMU-CS-86-126, Carnegie-Mellon University, Pittsburg.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9).

Poggio, T., Mukherjee, S., and Rifkin, R. (2002). Statistical learning theory and applications. MIT Course 9520 Spring 2002.

Popper, K. R. (1972). *A Lógica Da Pesquisa Científica.* Editora Cultrix, São Paulo.

Powell, M. J. D. (1987). Radial basis functions for multivariate interpolation: A review. In Cox, M. G., editor, *Algorithms for Approximation.* Clarendon Press and Oxford University Press.

Principe, J. C., Lefebvre, W. C., and Euliano, N. R. (2000a). *Neural and Adaptive Systems - Fundamentals Through Simulations.* John Wiley & Sons, Inc., New York.

Principe, J. C., Xu, D., Zhao, Q., and Fisher, J. W. (2000b). Learning from examples with information theoretic criteria. *Journal of VLSI Signal Processing-Systems*, 26(1–2):61–77.

Quastler, H. (1955). Approximate estimation of information measures. In Quastler., H., editor, *Information Theory in Psychology*, Illinois. Free Press.

Reed, R. D. and Marks, R. J. (2000). *Neural Smithing : Supervised Learning in Feedforward Artificial Neural Networks.* The MIT Press, Boston.

Renyi, A. (1970). *Probability Theory.* American Elsevier Publishing Company Inc., New York.

Richard, M. D. and Lippmann, R. P. (1991). Neural networks classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.

Rosenblatt, F. (1959). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 318–362. MIT Press.

Saporta, G. (1990). *Probabilités, Analyse Des Données et Statistique.* Technip, Paris.

Sarramon, M.-F. and Grandjean, H. (1998). *Le Foetus et Son Image.* Flammarion, Paris.

Schölkopf, B. (1997). *Support Vector Learning.* R. Oldenbourg Verlag, Munich.

Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning.* MIT Press, Cambridge, MA.

Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning.* Cambridge, MA.

Scholkopf, B. and Smola, A. (2002). *Learning with Kernels.* MIT Press, Cambridge MA.

Sereno, F. (2002). Notas sobre conversa pessoal na conferência RECPAD 2002 aveiro junho 27 2002. Technical report, Faculdade de Engenharia da Universidade do Porto.

Sereno, F. and Marques_de_Sá, J. P. (2000). Discussão com os obstetras sobre a estimação do peso fetal e um projecto de extracção de informação em bases de dados obstétricas. Technical Report 2000-05-06, INEB – FEUP.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (1999). Foetal weight estimation using neural networks. *Journal of the International Federation for Medical & Biological Engineering*, 37(2):1140–1141.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2000a). The application of radial basis functions and support vector machines to foetal weight prediction. In Dagli, C. H., Buczak, A. L., Ghosh, J., Embrechts, M., Ersoy, O., and Kercel, S., editors, *Proceedings of ANNIE 2000 (St Louis)*, pages 801–806, New York. Asme Press.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2000b). A comparative study of MLP and RBF neural nets in the estimation of the foetal weight and length. In Campilho, A. and Mendonça, A., editors, *Proceedings of RECPAD 2000 - 11th Portuguese Conference on Pattern Recognition*, pages 265–270, Porto. APRP - Associação Portuguesa de Reconhecimento de Padrões.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2001a). Foetal weight estimation by support vector regression. In Ruano, M. G., Ruano, A. E., Leiria, A. I., Moura, M. M., and Cacodcar, S., editors, *Sixth Portuguese Conference on Biomedical Engineering*, pages 1–4, Faro, Portugal. Faculdade de Ciências e Tecnologia, Universidade Do Algarve.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2001b). Support vector regression applied to foetal weight estimation. In Marko, K., Werbos, P., and Prokhorov, D., editors, *Proceedings of IJCNN'01 International Joint Conference on Neural Networks*, pages 1455–1458, Washington, DC. IEEE.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2002a). Comparison of two neural networks models to estimate foetal weight. In Pinho, A., Santos, B. S., Bastos, C., and Vieira, J., editors, *RECPAD 2002 - 12th Portuguese Conference on Pattern Recognition*, page 15, Portugal. Universidade de Aveiro.

Sereno, F., Marques_de_Sá, J. P., Matos, A., and Bernardes, J. (2002b). Information theoretic training of neural networks applied to foetal weight estimation. In IEEE, editor, *Proceedings Biosignal'2002 EURASIP Conference*, pages 174–176, BRN. IEEE.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.

Shepard, M. J. (1982). An evaluation of two equations for predicting fetal weight by ultrasound. *AM J Obstet Gynecol*, 142:47–54.

Shortliffe, E. H. (1976). *Computer-Based Medical Consultations - MYCIN.* Elsevier and North Holland, New York.

Silva, F. H. C. M. (1998). *Descorrelação de Dados e Aceleração Da Aprendizagem Em Redes Neuronais.* PhD thesis, Universidade Técnica de Lisboa IST.

Silva, F. M. and Almeida, L. B. (1990). Acceleration techniques for the backpropagation algorithm. In Almeida, L. B. and Wellekens, C. J., editors, *Neural Networks - Lecture Notes in Computer Science*, pages 110–119. Springer-Verlag.

Simon, H. U. (1997). Bounds on the number of examples needed for learning functions. *SIAM Journal of Computing*, 26(3):751–763.

Sivia, D. S. (1997). *Data Analysis, A Bayesian Tutorial.* Oxford University Press, Oxford.

Smola, A. and Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211 - 231.

Smola, A. J. (1996). Regression estimation with support vector learning machines. Master's thesis, Physik Department - Technische Universitat Munchen.

Smola, A. J. (1998). *Learning with Kernels.* PhD thesis, Technische Universitat Berlin.

Solomonoff, R. (1964). A formal theory of inductive inference part. *Information and control*, 7:1–22.

Soulie, F. F. (1997). Reseaux de neuronnes et statistiques. une introduction. In Thiria, S., Lechevalier, Y., Gascuel, O., and Canu, S., editors, *Statistique et Méthodes Neuronales*, pages 1–19, Paris. Dunod.

Spanos, A. (1999). *Probability Theory and Statistical Inference.* Cambridge University Press, Cambridge UK.

Tavares, L. V. and Correia, F. N. (1986). *Optimização Linear e Não Linear.* Fundação Calouste Gulbenkian, Lisboa.

Thimm, G. and Fiesler, E. (1995). Neural network initialization. In *International Workshop on Artificial Neural Networks*, pages 553–542, New York. Springer-Verlag.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems.* W. H. Winston, Washington,D.C.

Tribus, M. and McIrvine, E. (1971). Energy and information. *Scientific American*, 225:179–188.

Tuckey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, New York.

Twomey, S. (1977). *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements.* Dover Publications, INC, New York.

Valiant, L. G. (1984). A theory of learnable. *Comm. of the ACM*, 27:1134–142.

Vapnik, V., Golowich, S., and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In Mozer, M., Jordan, M., and Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, pages 281–287, Cambridge, MA. MIT Press.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Tehory.* Springer, New York.

Vapnik, V. N. (1998). *Statistical Learning Theory.* John Wiley & Sons Inc, New York.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Trans. on Neural Networks*, 10(5):988–999.

Veropoulos, K., Cristianini, N., and Campbell, C. (1999). The application of support vector machines to medical decision support: A case study. Technical report, Department of Engineering Mathematics, Bristol University.

Verri, A. and Poggio, T. (2001). Networks for learning: Regression and classification. Course MIT 9520 Spring 2001.

Vogl, T. P., Mangis, J. K., Rigler, A. K., Zink, W. T., and Alkon, D. L. (1988). Accelerating the convergence of the back propagation method. *Biological Cybernetics*, 59:257–263.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

Wahba, G. (1999). Support vector machines, reproducing kernel hilbert spaces and the randomized GACV. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods — Support Vector Learning*, pages 69–88, Cambridge, MA. MIT Press.

Warsof, S. L. and A, A. (1977). The estimation of fetal weight by computer-assisted analysis. *AM J Obstet Gynecol*, 128(8):881–92.

Weisstein, E. (2002). World of mathematics (MathWorld). Technical report, http://mathworld.wolfram.com.

Weltner, K. (1973). *The Measurement of Verbal Information in Psychology and Education*. Springer-Verlag, Berlin.

Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioural Sciences*. PhD thesis, Harvard University.

Widrow, B. and Hoff, M. E. (1960). Adaptative switching circuits. *IRE WESCON Convention Record*, pages 96–104.

Wiener, N. (1948). *Cybernetics, or Control and Communication in the Animal and the Machine*. John Wiley and Sons Inc., New York.

Woo, J. S. K. and Wan, C. W. (1986). Weight estimation in the small for dates fetus - an evaluation of a simple equation containg the fetal femur length. *Journal of Ultrasound in Medicine*, 5:453.

Xu, D. (1999). *Energy, Entropy and Information Potential for Neural Computation*. PhD thesis, University of Florida at Gainsville.

ERRATA

Pág	Onde está escrito:	Deve escrever-se:
4	$r(w) = \sum U_i^2 = \ldots$	$r(\mathbf{w}) = \sum U_i^2 = \ldots$
4	os modelos lineares, mas também outros modelos mais complexos	os modelos lineares nos parametros que afectam as variáveis preditoras, pelo que se aplica também a modelos mais complexos
22	o erro verdadeiro $\widehat{er}_P(h)$	o erro verdadeiro $er_P(h)$
22	$opt_P(\mathbb{H}) = \inf \widehat{er}_P(h)$	$opt_P(\mathbb{H}) = \inf er_P(h)$, (por definição)
24	Por definição, o *tamanho suficiente da amostra*, $m_0(\varepsilon, \delta)$,não depende da distribuição de probabilidade P	Por razões práticas deseja-se um $m_0(\varepsilon, \delta)$ que não depende da distribuição de de probabilidade P.
25	Pode caitar-se que se uma função h tiver um erro amostral pequeno, então o seu erro verdadeiro também será pequeno	Pode aceitar-se que se uma função h tiver um erro amostral que decresce com m, então, o seu erro verdadeiro também decrescerá com m.
26	Este resultado mostra que... o *erro amostral* de h está muito próximo do *erro verdadeiro*	Este resultado mostra que ... o *erro amostral* de h tende em probabilidade para o *erro verdadeiro*
27	Por exemplo, as classes de funções geradas... nem sempre são conjuntos *finitos.*	Por exemplo, os conjuntos de funções gerados... em geral não são . conjuntos *finitos*
30	Definição 15 Exemplo 16	Exemplo 16
30	Seja a seguinte classe de hipóteses definida por:	Sejam os números reais $r_1, r_2, r_3 \in [0, 1]$ e a seguinte classe de hipóteses definida por:
33	Recorde-se que se afirmou atrás que um algoritmo $L\ldots$ o erro amostral $\widehat{er}(L(\mathbf{Z}))$ for arbitrariamente pequeno.	Recorde-se que se afirmou atrás que um algoritmo $L\ldots$ fôr possível escolher m_0 tal que para $m > m_0$ o erro amostral $\widehat{er}(L(\mathbf{Z}))$ convergir em probabilidade para um valor arbitrariamente pequeno.
37	$Pdim(\mathbb{F}) = \dim(\mathbb{F})$	$Pdim(\mathbb{F}) = \lvert\mathbb{F}\rvert$.(cardinalidade de $\mathbb{F}$)
44	e, por conveniência de notação	e, por conveniência de simplificação matemática
45	Definição 35 (Risco Esperado) É o mínimo erro esperado $\mathbf{E}[c(x, y, f(x))]$ para todos os m exemplos de treino, e que pode ...	É o valor esperado $\mathbf{E}[c(x, y, f(x))]$ para todos os possíveis exemplos, e que pode ...
45	Exemplo 37 ... $g_i(x_j)$, pode ver-se que ...	$g_i(x_j)$, pode provar-se que ...
81	*kernel* polinomial de ordem dois, que como se sabe satisfaz as condições de Mercer	se sabe satisfaz as condições de Mercer referidas mais à frente na pág. 96 (secção 3.2.5).
90	$(K + \lambda)\mathbf{c} = \mathbf{y}$	$(K + \lambda\mathbf{I})\mathbf{c} = \mathbf{y}$
247	A secção 6.3 faz uma revisão ... ...entropia relativa e informação mútua	A secção 6.3 faz uma revisão ... ... entropia relativa e informação mútua baseando-se largamente nos trabalhos de Príncipe J., Xu D., Zhao Q. Fisher, 2000b.