

## 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 supervisionada 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 percepção 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.