CÉSAR DUARTE ALVES DA ROCHA

Algoritmo Recursivo dos Mínimos Quadrados para Regressão Linear Local

Departamento de Matemática Aplicada
Junho de 2001

CÉSAR DUARTE ALVES DA ROCHA

Algoritmo Recursivo dos Mínimos Quadrados para Regressão Linear Local

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de mestre em Estatística

Departamento de Matemática Aplicada
Junho de 2001

Agradecimentos

Os meus sinceros agradecimentos ao Prof. Luis Torgo e à Prof. Margarida Brito.

ÍNDICE

Índice de Figuras e tabelas

Capítulo I

INTRODUÇÃO

Nenhuma ferramenta estatística tem tido a atenção dada à análise de regressão nos últimos 25 anos. Tanto os analistas de dados práticos como os estatísticos teóricos têm contribuído para um avanço sem precedentes nesta área de investigação. Muitos volumes têm sido escritos resultando num aumento dos métodos de regressão efectivos.

O termo análise de regressão serve para descrever um conjunto de técnicas estatísticas que são usadas para construir inferências sobre relações entre quantidades num sistema específico.

A metodologia analítica que foi sendo desenvolvida tornou-se prontamente acessível devido ao rápido desenvolvimento de *software* informático. Embora não fosse esse o objectivo inicial, a análise de regressão é agora provavelmente o método de análise de dados mais utilizado.

Em 1885 Sir Francis Galton introduziu pela primeira vez a palavra "regressão" num estudo que demonstrava que a estatura de uma descendência não se aproxima dos progenitores, mas sim para a estatura média de ambos. A descoberta do método de regressão baseado nos mínimos quadrados é atribuída a Carl Friedrich Gauss, que usou o procedimento no início do Século XIX, embora exista alguma controvérsia no que respeita a esta descoberta, uma vez que aparentemente Adrien Marie Legendre publicou o primeiro trabalho sobre o seu uso em 1805.

No final dos anos 60 tornou-se evidente que, em muitas situações não ideais, a regressão linear global obtida pelos mínimos quadrados não é adequada. Como ilustração deste facto, temos por exemplo o caso de relações não lineares entre variáveis e para as quais a regressão linear clássica não é eficaz. Os métodos de regressão local conseguem, em muitas situações, ultrapassar este problema. Esta metodologia de regressão pode ser sucintamente descrita pelos seguintes passos: dada uma nova observação para a qual se pretende fazer uma previsão da variável objectivo, os métodos de regressão local começam por pesquisar observações da amostra de treino mais "próximas" (isto é, mais semelhantes); em seguida, estas observações são usadas para calcular os parâmetros de um modelo de regressão local usando por exemplo o método

dos mínimos quadrados; finalmente este modelo local é usado para obter uma previsão para a observação de teste em causa. Inerente a este método está o problema de determinar qual o comprimento da vizinhança a usar para obter os parâmetros do modelo de regressão local (isto é, quantos casos, de entre os mais próximos do caso de teste, devem ser usados). Uma solução é tentar vários comprimentos de vizinhança (por exemplo 10, 20, 30, etc.), obter um modelo para cada comprimento e ver qual é o melhor. Esta solução acarreta custos computacionais acrescido devido à necessidade de obter vários modelos de cada vez que se pretende uma previsão para uma observação de teste. Como veremos, o algoritmo recursivo dos mínimos quadrados permite actualizar um modelo linear obtido com N casos para $N + 1$ casos, sem a necessidade de efectuar todos os cálculos usuais nos mínimos quadrados clássicos.

A presente tese estuda o problema da obtenção de modelos de regressão linear locais. De entre as diversas problemáticas associadas a estes modelos, este trabalho dedica-se ao estudo do comprimento da vizinhança "ideal" a ser utilizado para obter os modelos lineares locais. A resolução deste problema pode ser conseguida através de dois passos: obtendo diferentes modelos com vários tamanhos da vizinhança e escolhendo em seguida um desses modelos como o "melhor" que será usado para obter a previsão para o caso de teste. Estes dois passos podem ser levados a cabo de modo eficiente em termos computacionais graças ao algoritmo recursivo dos mínimos quadrados e à estatística PRESS. O algoritmo recursivo dos mínimos quadrados resolve o problema da reconstrução do modelo linear para diferentes comprimentos de vizinhança, enquanto a estatística PRESS dá a estimativa do erro *leave-one-out* de validação cruzada recorrendo a uma só identificação do modelo oferecendo assim um critério de selecção do comprimento de vizinhança.

O trabalho levado a cabo nesta tese consistiu em implementar o método de regressão linear local brevemente descrito acima; aplicar o *software* desenvolvido a vários conjuntos de dados e comparar o programa desenvolvido com outras possíveis abordagens à regressão múltipla que sejam relacionadas (concretamente os métodos de regressão clássica e a regressão linear local usando um comprimento de vizinhança fixo). Não se pretendeu nesta tese, mostrar que o método de regressão linear local é "superior" a qualquer outro método de regressão. Conforme aferimos em alguns conjuntos de dados usados, ele pode não ser o mais adequado. Este facto, entre outros, é referido no último capítulo da tese, onde se apresentam possíveis caminhos a seguir futuramente com a ideia de melhorar os aspectos menos positivos do método.

Esta tese está organizada da seguinte forma: o capítulo 2 introduz a metodologia genérica da regressão local bem como as suas principais variantes; segue-se o capítulo 3 onde é abordada a questão central da tese que é o problema da selecção do tamanho da vizinhança no contexto da regressão linear local; no capítulo 4 descrevemos algumas das questões principais surgidas na implementação do método estudado, bem como a sua aplicação e comparação com outros métodos de regressão; finalmente o capítulo 5 apresenta as principais conclusões deste trabalho bem como algumas possíveis direcções para trabalho futuro.

Capítulo 2

A REGRESSÃO LOCAL

O problema de regressão múltipla consiste em obter a previsão do valor de uma variável dependente tendo em consideração um conjunto de observações de outras variáveis independentes.

Consideremos um espaço mensurável χ de dimensão J, um vector aleatório $X=(X_1,X_2...,X_J)$ que toma valores em χ e seja Y a variável dependente associada a X.

O nosso objectivo é encontrar uma função $f(x)=y, y\in\Re$, de maneira que y é o valor real de Y, quando X toma o valor x. Em muitas situações práticas, devido a erros de observação e desconhecimento do conjunto total de factores que interferem no comportamento de Y, ficamos impossibilitados de determinar essa mesma função. Daí que somos limitados a tentar uma aproximação de f só com base no conhecimento disponível a partir das observações. Para representar a função de predição ou aproximação utilizaremos o símbolo $\hat{f}$, sendo esta mesma função definida em χ e tomando valores em $\Re$. Dado um valor x observado do vector aleatório, $\hat{f}(x)$ é a resposta de previsão correspondente

$\hat{f}:\chi\rightarrow\Re$

$x\rightarrow\hat{f}(x)$

A função de predição é aquela que minimiza o custo dos eventuais erros nas respostas dadas. Denotamos por $L(y,\hat{f}(x))$ o custo resultante da resposta $\hat{f}(x)$ em que y é o valor real de Y para o caso x. O valor esperado do custo da utilização da função de predição designa-se por $R^*(\hat{f}(x))$ e representa o risco na utilização de $\hat{f}(x)$ para previsão do valor de Y.

$R^*(\hat{f}(x))=E(L(y,\hat{f}(x)))$.

Neste contexto podemos considerar a regressão dos menores desvios absolutos (LAD) para a qual $R^*(\hat{f}(x))=E(|Y-\hat{f}(x)|)$ ou seja, o valor do risco é igual ao erro

absoluto médio. Uma outra hipótese é considerar a regressão dos menores desvios quadrados (LSD) em que $R^*\left(\hat{f}(x)\right) = E\left(\left[Y - \hat{f}(x)\right]^2\right)$ o que equivale a tomar para valor do risco o erro quadrático médio.

2.1 REGRESSÃO LINEAR GLOBAL

Num modelo de regressão linear múltipla, assume-se uma relação entre um conjunto de variáveis independentes, $X_j (j = 1,2,...,J)$, e uma variável dependente também quantitativa, Y, com a forma da expressão $Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + ... + \beta_j X_{jn} + W_n$ onde,

n: índice representando as observações das variáveis $X_1, X_2, ..., X_J$ e $Y (n = 1,...N)$

$(X_{1n}, X_{2n} ... X_{jn}, Y_n)$: n-ésima observação das variáveis $X_1, X_2, ..., X_J$ e Y

$\beta_0, \beta_1, ... \beta_J$: parâmetros fixos (desconhecidos, a estimar) da relação linear entre $X_1, X_2, ..., X_j$ e Y

W_n : erro aleatório associado ao valor observado Y_n

As hipóteses subjacentes a este modelo são as seguintes:

(i) Os valores X_{jn} são constantes predeterminadas, sem erro;

(ii) Os erros W_n são mutuamente independentes, têm valor esperado nulo, variância constante, σ^2, e são normalmente distribuídos, isto é, $W_n \rightarrow N(0, \sigma^2)$

Assumindo como valor do risco os menores desvios quadrados, os parâmetros $\beta_0, \beta_1, ..., \beta_j$ do modelo considerado podem ser estimados a partir de um conjunto de observações na forma $(X_{1n}, X_{2n}, ..., X_{jn}, Y_n)$ $(n = 1, ..., N \quad com \quad N > J + 1)$ recorrendo ao método dos mínimos quadrados. Em notação matricial, podemos escrever o modelo de regressão linear múltipla na forma $Y = X\beta + W$ onde Y é o vector objectivo de dimensão N (número de observações), X é a matriz de N linhas e $P = J + 1$ colunas, β é o vector de dimensão $P = J + 1$ (vector dos parâmetros) e W é o vector de erros cuja dimensão é igual ao número de observações:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{J1} \\ 1 & x_{12} & x_{22} & \dots & x_{J2} \\ . & . & . & & . \\ & . & . & & . \\ & . & . & & . \\ 1 & x_{1N} & x_{2N} & \dots & x_{JN} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_J \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_N \end{bmatrix}$$

Neste modelo de regressão quer-se minimizar o critério $C = \sum_{i=1}^{n} L(y_i, \hat{y}_i)$ em que L é a função custo, y_i é um valor objectivo, $\hat{y}_i = x_i^T \beta$, sendo β o vector de parâmetros, x_i um vector correspondente à i-ésima observação das variáveis independentes onde se acrescenta a constante 1 de modo a considerar um termo constante. No caso da estimação pelo método dos mínimos quadrados a função custo é $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ e isso significa que vamos minimizar $C = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2$. Mais uma vez, em notação matricial, temos que minimizar a expressão $(Y - X\hat{\beta})^T (Y - X\hat{\beta})$ o que dá origem às equações normais $(X^T X)\hat{\beta} = X^T Y$ obtendo-se $\hat{\beta} = (X^T X)^{-1} X^T Y$ para estimador dos mínimos quadrados.

A estimativa obtida por regressão linear global pode ser muito pobre quando se consideram relações não lineares. Uma alternativa consiste em aplicar uma transformação não linear, mantendo ainda um modelo linear nos parâmetros. A regressão polinomial $P_m(x)$ de grau m é um exemplo dessa transformação, a qual se costuma utilizar quando *a priori* se dispõe de informação sobre os dados. Outra alternativa para resolver este problema é o ajustamento linear local. Dado um ponto q, podemos considerar somente um conjunto de pontos vizinhos para calcular uma estimativa em q.

2.2 O Método de regressão local

A regressão local é uma extensão natural da regressão global de tal modo que apareceu independentemente em lugares diferentes no séc XIX. Nesta altura os conjuntos eram univariados com observações igualmente espaçadas: tais conjuntos eram tão simples que era possível efectuar os cálculos da regressão com papel e lápis. Grande parte deste trabalho surgiu em estudos actuariais.

Segundo Hoem (1983) estes métodos foram usados na Dinamarca em 1829, mas não foram publicados durante 50 anos. Posteriormente, Gram (1883) publicou um trabalho a partir da sua tese de doutoramento de 1879 sobre ajustamento polinomial local com uma função de "pesagem" uniforme e com uma função de pesagem que tende para zero. Muito do seu trabalho foi focado sobre o ajustamento local cúbico e foram usados coeficientes binomiais para os pesos. Igualmente Stigler (1978) refere que nos Estados Unidos, De Forest (1873,1874) usou ajustamento polinomial local. Na Grã-Bretanha o trabalho nesta área começou em 1870 quando Woolhouse publicou um método baseado em ajustamento local quadrático. O método recebeu muita discussão mas foi eventualmente eclipsado por um método de Spencer (1904) que se tornou popular devido à sua eficiência de cálculo e bom desempenho. Um trabalho que se tornou bastante conhecido foi o de Henderson (Estados Unidos, 1924) e da comunidade britânica – que incluía matemáticos tais como Whittaker (1923), que, junto com Henderson, também inventaram os *smoothing splines*. Daqui resultou um direccionamento de métodos de ajustamento local para a literatura de séries temporais. Por exemplo, o livro *The Smoothing of Time series* (Macaulay, 1931), mostra como os métodos de ajustamento local podem ser aplicados a séries económicas, o qual por sua vez teve uma influência substancial no que se tornaria um marco importante em métodos de ajustamento local. Tudo começou no *U.S. Bureau of the Census* (1954) em que uma série de programas de computador foi desenvolvida para ajustamento sazonal de séries temporais, culminando com o método X-11 (Shishkin, Young e Musgrave, 1967). O X-11 foi pouco mencionado na literatura estatística nessa altura porque os seus métodos eram empíricos em vez de resultarem de um modelo estatístico completamente especificado. Contudo, o X-11 tornou-se no padrão para ajustamento de séries temporais económicas e é ainda hoje bastante usado.

A visão mais moderna de ajustamento por regressão local tem a sua origem nos anos 50/60, com métodos *kernel* introduzidos no contexto da estimação de funções densidade de probabilidade (Rosenblatt, 1956; Parzen, 1962) e no contexto da regressão (Nadaraya, 1964; Watson, 1964). Esta nova visão ampliou o ajustamento de uma função de uma única variável independente com medidas igualmente espaçadas para o ajustamento de uma função de medidas esparsas de uma ou mais variáveis independentes.

Os métodos *kernel* são um caso especial da regressão local; um método *kernel* restringe-se a escolher a família de funções constantes. Reconhecendo a fraqueza de uma aproximação local constante, a regressão local mais geral renasceu no final dos anos 70 (Stone, 1977; Cleveland, 1979; Katkovnik, 1979). Outra limitação deste trabalho inicial era a suposição de distribuição quase Gaussiana, o que levou Brillinger (1977) a formular uma aproximação geral e Cleveland (1979) e Katkovnik (1979) a desenvolverem aproximadores robustos. Mais tarde, também Tibshirani e Hastie (1987) alargaram substancialmente o domínio para muitas distribuições tal como na regressão logística e desenvolveram algoritmos gerais de ajustamento. O alargamento a novos conjuntos continua hoje (Fan e Gijbels, 1994; Loader, 1995).

O trabalho sobre regressão local continuou ao longo dos anos 80 e 90. Aqui numerosas aproximações podem ser feitas: Cleveland e Devlin (1988) aplicam o ajustamento local linear e quadrático directamente a dados multivariados. Friedman e Stuetzle (1981) usam regressão linear local como base para construir estimativas *projection pursuit.* Hastie e Tibshirani (1990) usam regressão local em modelos aditivos. Estes métodos têm diferenças substanciais em requisitos nos dados e tipos de superfícies que podem ser modeladas com sucesso; o uso de diagnósticos gráficos para ajudar a tomar decisões torna-se crucial nestes casos.

Acompanhando a corrente moderna do trabalho em regressão local está uma nova procura de resultados assimptóticos. Ela começou nos artigos mais antigos (e.g., Rosenblatt, 1956; Stone, 1977) e cresceu em intensidade desde os anos 80 (e.g., Muller, 1987; Hardle, 1990; Fan, 1993; Ruppert e Wand, 1994).

2.3 Principais questões da regressão local

2.3.1 Distância

O método de regressão local é bastante sensível à escolha da função distância. Esta função é usada para aferir a distância entre duas observações no espaço multidimensional definido pelas variáveis independentes do problema de regressão múltipla em estudo. Podemos definir uma função distância para regressão local sem olhar aos formalismos matemáticos de uma métrica. Em seguida apresentam-se três maneiras diferentes de definir e utilizar funções distância:

(i) Função distância global: Escolhemos uma função distância igual para todo o espaço;

(ii) Função distância local baseada no caso de teste: Através da minimização do erro de validação cruzada ou de um critério relacionado são determinados os parâmetros da função distância ou a sua forma para cada caso de teste. Stanfill (1987) designa esta abordagem por métrica uniforme sendo a mesma discutida e Stanfill e Waltz (1986), Hastie e Tibshirani (1994) e Friedman (1994);

(iii)Função distância local baseada nos casos de treino: utiliza-se uma função distância diferente para cada ponto x_i, de maneira que o critério de erro se escreve na forma $C(q)=\sum_{i=1}^{n}\left[\left(y_i - f\left(x_i,\hat{\beta}\right)\right)^2 K\left(d_i(x_i,q)\right)\right]$. A função $d_i(\ )$ é normalmente seleccionada em simultâneo com as observações de treino utilizadas no modelo, partindo da minimização do erro da validação cruzada ou por cálculo directo. Stanfill (1987) chama a esta distância métrica variável. Na área da classificação isto equivale a ter uma função distância que é variável no conjunto das classes (Waltz, 1987; Aha e McNulty, 1989; Aha, 1989,1990).

No trabalho presente a distância utilizada será do tipo global e euclidiana, sendo possível encontrar outros tipos de função distância no apêndice A.

2.3.2 Vizinhança

Uma das questões fundamentais da regressão local é a escolha do tamanho da vizinhança a ser usada para obter cada modelo local. Este tamanho vai determinar quais

os casos da amostra de treino que vão ser usados para obter os parâmetros do modelo local que irá determinar a previsão para o caso de teste em jogo. Como exemplo observemos a figura:

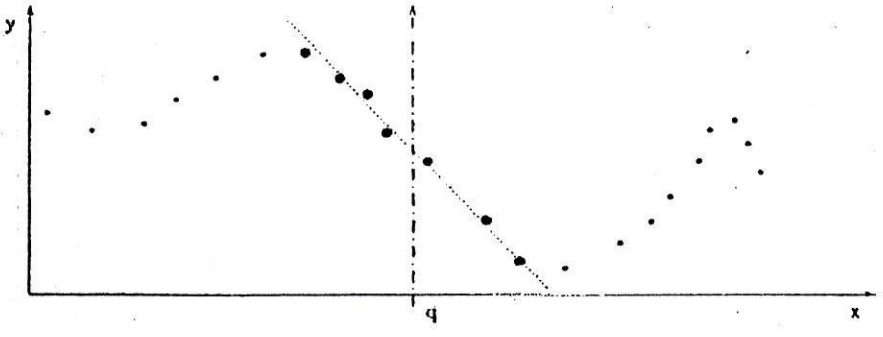

Figura 2.1: Ajustamento linear local com base em 7 vizinhos

Quando consideramos um número de vizinhos superior a 7 para o caso de teste q, o modelo local adequado deixa de ser o linear.

A escolha de uma vizinhança muito pequena faz com que a aproximação obtida pela regressão local seja muito "sensível" a pequenas flutuações nas observações de treino. Isto torna a superfície de regressão aproximada muito pouco "suave", tendo portanto muitos "saltos". Já a escolha de uma vizinhança demasiado larga poderá trazer um alisamento demasiado grande o que faz com que a superfície aproximada não capture algumas das características da função de regressão a aproximar. Isto mostra a importância da escolha da vizinhança no método de regressão local, que é o tópico principal que será discutido nesta tese.

2.3.3 Modelo local usado

Dado um conjunto de vizinhos de uma observação de teste, para a qual pretendemos obter uma previsão, vários tipos de modelos locais podem ser obtidos com estas amostras de treino. O uso de diferentes modelos vai obviamente levar a previsões diferentes. A utilização de modelos demasiado complexos no contexto destas vizinhanças pode levantar problemas de identificação, pelo facto de estarmos a estimar demasiados parâmetros para o tamanho da amostra local em questão. De entre as

diferentes alternativas existentes na literatura, os modelos constantes (tipo *kernel*), e os modelos lineares (polinómios do primeiro grau), estão entre os mais usados. No entanto, existem também trabalhos onde polinómios de grau mais elevado são usados no contexto da regressão local.

2.3.4 Relevância das variáveis independentes

Como referido anteriormente uma das questões cruciais na regressão local é a procura das amostras de treino mais semelhantes à observação de teste em análise. Esta procura é feita com recurso à definição de uma métrica no espaço multidimensional definido pelas variáveis independentes do problema. Neste contexto, é definida uma função de distância que vai quantificar a semelhança entre quaisquer duas observações. A presença de variáveis independentes irrelevantes para o problema de regressão em estudo, ou mesmo a presença de variáveis com escalas de valores bastante diversos, pode colocar sérios problemas aos métodos de regressão local, ao "distorcerem" a noção de semelhança entre duas observações. No sentido de minimizar este problema é comum o uso de técnicas de normalização para aliviar os problemas de diferenças de escala, e também o uso de técnicas de diferenciação do "peso" (importância) das variáveis no contexto do cálculo da distância entre duas observações.

2.4 Alguns exemplos de regressão local

Nesta secção descrevemos os dois métodos de regressão local mais usados. De entre estes, a regressão linear local será o objecto de estudo desta tese.

2.4.1 Regressão *kernel*

Neste método a ideia é usar uma função constante para a aproximação local a obter para cada caso de teste. Se escolhermos o critério de erro dos mínimos quadrados, $C = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0\right)^2$, então temos

$$\frac{\partial C}{\partial \hat{\beta}_0} = 0 \Leftrightarrow$$

$$-2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0\right) = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\beta}_0 \Leftrightarrow$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 \Leftrightarrow$$

$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$ e $\frac{\partial^2 C}{\partial\left(\hat{\beta}_0\right)^2} = 2n > 0$ o que prova que $\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$ minimiza o erro quadrático, sendo então suficiente calcular a média aritmética de n valores de treino $\{y_i, y_2, ...y_n\}$. No entanto, convém dar mais importância aos valores próximos do caso de teste e desprezar aqueles que estão distantes. Existem duas formas equivalentes de fazer isto: podemos "pesar" os dados ou o critério definido atrás. Quando "pesamos" os dados temos que definir uma distância $d(x_i, q)$ entre o caso de teste q e o caso de treino x_i. Uma função distância que pode ser utilizada é a euclidiana. Além disso, necessitamos de uma função de pesagem (também designada por função *kernel*). A função de pesagem uniforme é um exemplo comum e pode ser encontrada no apêndice B além de outras funções *kernel.*

A média pesada correspondente à previsão é neste caso $\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i K(d(x_i, q))}{\sum_{i=1}^{n} K(d(x_i, q))}$ sendo K a função *kernel.* Resulta da última expressão que $\hat{\beta}_0$ é dependente da localização do caso de teste q. No caso em que ponderamos o critério de erro temos $C(q) = \sum_{i=1}^{n}\left[\left(y_i - \hat{\beta}_0\right)^2 K(d(x_i, q))\right]$.

Pretendemos saber qual é a expressão de $\hat{\beta}_0$ que minimiza o custo C(q). Então temos que

$$\frac{\partial C}{\partial \hat{\beta}_0} = 0 \Leftrightarrow$$

$$-2\sum_{i=1}^{n}\left[\left(y_i - \hat{\beta}_0\right)K\left(d\left(x_i, q\right)\right)\right] = 0 \Leftrightarrow$$

$$\sum_{i=1}^{n} y_i K\left(d\left(x_i, q\right)\right) = \sum_{i=1}^{n} \hat{\beta}_0 K\left(d\left(x_i, q\right)\right) \Leftrightarrow$$

$$\hat{\beta}\sum_{i=1}^{n} K\left(d\left(x_i, q\right)\right) = \sum_{i=1}^{n} y_i K\left(d\left(x_i, q\right)\right) \Leftrightarrow$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i K\left(d\left(x_i, q\right)\right)}{\sum_{i=1}^{n} K\left(d\left(x_i, q\right)\right)} \quad e \quad \frac{\partial^2 C}{\partial\left(\hat{\beta}_0\right)^2} = 2\sum_{i-1}^{n} K\left(d\left(x_i, q\right)\right) > 0$$ porque a função *kernel* é

não negativa e isto significa que a estimativa obtida quando se pesam os dados directamente é a mesma que se obtém pesando o critério de erro.

Uma metodologia bastante conhecida que pode ser vista como um caso particular da regressão *kernel* é o modelo dos *k* vizinhos mais próximos. Este método corresponde a um modelo *kernel* no qual escolhemos uma função de pesagem uniforme e uma vizinhança definida como a distância ao *k*-ésimo vizinho mais próximo.

2.4.2 Regressão linear local

Na regressão linear local em vez de usarmos uma aproximação local constante, como na regressão *kernel,* vamos obter um modelo linear com a vizinhança de cada caso de teste. O ajustamento é feito tal como na regressão *kernel,* usando uma função de pesagem. Além disso, podemos também pesar os dados directamente ou pesar o critério de erro. As duas formas são equivalentes nos modelos lineares locais. Se pesarmos o critério, temos $C(q) = \sum_{i=1}^{n}\left[\left(y_i - x_i^T \hat{\beta}\right)^2 K\left(d\left(x_i, q\right)\right)\right]$.

Por outro lado, quando pesamos os dados, consideramos uma matriz W cujos elementos diagonais são $W_{ii} = w_i = \sqrt{K\left(d\left(x_i, q\right)\right)}$ sendo os restantes elementos nulos. Em notação matricial temos que $Z = WX$ e $v = Wy$ em que X é a matriz de dados e y é o vector objectivo. Isto dá origem a novas variáveis no modelo de regressão e analogamente ao modelo de regressão linear global, usando o critério dos mínimos

quadrados podemos escrever $\left(Z^T Z\right)\hat{\beta} = Z^T v \Leftrightarrow \hat{\beta} = \left(Z^T Z\right)^{-1} Z^T v$. O estimador resultante para o caso de teste q é $\hat{y} = q^T \left(Z^T Z\right)^{-1} Z^T v$.

A regressão linear local não apresenta grandes diferenças em relação à regressão *kernel* quando se consideram conjuntos de dados regularmente distribuídos. No entanto, a regressão *kernel* tem muitas desvantagens quando a distribuição dos dados é irregular.

Capítulo 3

A SELECÇÃO DO TAMANHO DA VIZINHANÇA

Um problema importante que se levanta com os métodos de regressão local é o de saber qual o tamanho da vizinhança a considerar no modelo a obter para cada caso de teste. Se tomamos uma vizinhança demasiado grande para um caso particular de teste podemos afastar-nos da região em que a distribuição das observações é adequada ao modelo em questão. Isto pode distorcer a nossa previsão pelo facto de incluirmos observações que não deveriam ser utilizadas na construção da estimativa.

Temos várias formas de escolher o tamanho h da vizinhança (Scott, 1992; Cleveland e Loader, 1994c):

I) Vizinhança de tamanho fixo para a qual o valor de h é constante (Fan e Marron, 1993) e a função *kernel* se pode escrever como $K\left(\frac{d(x_i,q)}{h}\right)$;

II) Selecção com base no vizinho mais próximo em que se toma como valor de h a distância à k-ésima observação (Stone, 1977; Cleveland, 1979; Farmer e Sidorowich, 1988a,b; Townshend, 1992; Hastie e Loader, 1993; Fan e Gijbels, 1994; Ge et al., 1994; Naes et al., 1990; Naes e Isakson, 1992; Wang et al., 1994; Cleveland e Loader, 1994b). O tamanho da vizinhança vai variar em função da concentração de vizinhos do caso de teste;

III) Selecção feita globalmente, ou seja, h é escolhido tendo em conta o erro mínimo de validação cruzada feita no conjunto total de dados;

IV) Selecção feita localmente para cada caso de teste, que equivale a determinar o valor de h por minimização do erro de validação cruzada ou critério semelhante, relativamente a cada caso de teste. (Vapnik, 1992)

V) Selecção feita localmente para cada caso de treino: o tamanho da vizinhança depende de cada observação do conjunto de treino. O critério de erro resultante é $C(q)=\sum_{i=1}^{n}\left[\left(y_i - f(x_i,\hat{\beta})\right)^2 K\left(\frac{d(x_i,q)}{h_i}\right)\right]$. Os valores de h_i são obtidos por minimização do erro de validação cruzada ou por cálculo directo. Estes valores são guardados inicialmente com o conjunto de treino.

Segundo Fan e Marron (1994b), quando consideramos uma vizinhança de tamanho fixo, estamos limitados nas aplicações embora esta seja de fácil interpretação. Cleveland e Loader (1994a) consideram a selecção com base no vizinho mais próximo mais adequada do que aquela que é fixa. Quando usamos uma vizinhança de tamanho fixo associada com uma função *kernel* que tende para zero numa distância finita esta tem o inconveniente de poder originar grande variância para zonas de dados esparsos. Esta situação tende a agravar-se quando a dimensão dos dados aumenta. Além disso, em certos casos não dispomos de observações na região limitada pelo tamanho de vizinhança fixo, sendo a estimativa obtida indefinida (Cleveland e Loader, 1994b). Fan e Marron (1994b) argumentam sobre a vantagem em usar vizinhanças locais variáveis. Elas são flexíveis em relação à distribuição dos dados, condições de heteroscedasticidade e suavidade da função que se quer estimar. Fan e Gijbels (1992) preferem a escolha da vizinhança local para cada caso de teste pois assim se conseguem acomodar variações rápidas ou assimétricas nos dados.

Na descrição que se segue vamos considerar um tipo de vizinhança no contexto da regressão linear local que será escolhida para cada caso de teste usando como critério a minimização do erro *leave-one-out* de validação cruzada.

3.1. Mínimos quadrados recursivos

A selecção da vizinhança feita para cada caso de teste implica o cálculo sucessivo de estimativas. Na regressão linear local o excesso de computação resultante deste método pode ser evitado com o algoritmo dos mínimos quadrados recursivos. Este permite a actualização da estimativa dos parâmetros do modelo linear obtida com N vizinhos para $N+1$ vizinhos de uma maneira eficiente, isto é, sem que seja necessário recorrer ao método usual da resolução das equações normais.

Vamos supor que para um determinado caso de teste q os vizinhos x_i estão ordenados de acordo com a distância $d(x_i, q)$. Suponhamos ainda que dispomos de uma função *kernel* e que pretendemos considerar um comprimento de vizinhança, limitado por N_M (valor máximo) e N_m (valor mínimo). Denotemos por $\hat{\beta}_{(N)}$ o vector de

parâmetros obtido com N vizinhos e por $V_{(N)}$ a matriz $\left(X^T X\right)^{-1}$ também relativa a N vizinhos.

Na suposição de uma função *kernel* uniforme podemos resumidamente apresentar o algoritmo recursivo dos mínimos quadrados com o seguinte sistema:

$$\begin{cases} V_{(N+1)} = V_{(N)} - \dfrac{V_{(N)} x_{(N+1)} x_{(N-1)}^T V_{(N)}}{1 + x_{(N+1)}^T V_{(N)} x_{(N+1)}} \\ \gamma_{(N+1)} = V_{(N+1)} x_{(N+1)} \\ e = y_{(N+1)} - x_{(N+1)} \hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} = \hat{\beta}_{(N)} + \gamma_{(N+1)} e \end{cases}$$

onde $x_{(N+1)}$ é a N + 1-ésima linha da matriz X e $x_{(N+1)}^T$ a linha transposta correspondente.

Antes de justificar as equações anteriores, é necessário enunciar o seguinte resultado:

Lemma 1 (*Fórmula de Inversão Matricial*)

Consideremos quatro matrizes F, G, H e K e a matriz (F + GHK). Suponhamos que as inversas das matrizes F, G e (F+GHK) existem.

Então

$$\left(F + GHK\right)^{-1} = F^{-1} - F^{-1} G \left(H^{-1} + KF^{-1} G\right)^{-1} KF^{-1}$$

Consideremos agora o caso em que F é uma matriz quadrada, não singular de ordem n, G = z onde z é um vector de dimensão $n, K = z^T$ e H= I. Então, a fórmula anterior simplifica-se, ficando:

$$\left(F + zz^T\right)^{-1} = F^{-1} - \frac{\left(F^{-1} zz^T F^{-1}\right)}{1 + z^T F^{-1} z}$$

onde o denominador do lado direito da equação é um escalar.

Demonstração:

Multiplicando $F + zz^T$ *por* $F^{-1} - \dfrac{\left(F^{-1} zz^T F^{-1}\right)}{1 + z^T F^{-1} z}$ *à direita temos*

$$\left(F + zz^T\right)\left(F^{-1} - \frac{\left(F^{-1} zz^T F^{-1}\right)}{1 + z^T F^{-1} z}\right) = I - \frac{zz^T F^{-1}}{1 + z^T F^{-1} z} + zz^T F^{-1} - \frac{zz^T F^{-1} zz^T F^{-1}}{1 + z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right) - zz^T F^{-1} + zz^T F^{-1} + zz^T F^{-1} z^T F^{-1} z - zz^T F^{-1} zz^T F^{-1}}{1+z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right) + zz^T F^{-1} z^T F^{-1} z - zz^T F^{-1} z^T F^{-1} z}{1+z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right)}{1+z^T F^{-1} z} = I$$

Do mesmo modo fazendo a multiplicação à esquerda obtém-se

$$\left(F^{-1} - \frac{\left(F^{-1} zz^T F^{-1}\right)}{1+z^T F^{-1} z}\right)\left(F + zz^T\right) = I - \frac{F^{-1} zz^T}{1+z^T F^{-1} z} + F^{-1} zz^T - \frac{F^{-1} zz^T F^{-1} zz^T}{1+z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right) - F^{-1} zz^T + F^{-1} zz^T + F^{-1} zz^T z^T F^{-1} z - F^{-1} zz^T F^{-1} zz^T}{1+z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right) + F^{-1} zz^T z^T F^{-1} z - F^{-1} zz^T z^T F^{-1} z}{1+z^T F^{-1} z} =$$

$$= \frac{I\left(1+z^T F^{-1} z\right)}{1+z^T F^{-1} z} =$$

$= I$, *o que completa a demonstração*

O vector de parâmetros $\hat{\beta}_{(N)}$ tem a forma $\hat{\beta}_{(N)} = \left(X_{(N)}^T X_{(N)}\right)^{-1} X_{(N)}^T y_{(N)}$. Dispondo de uma nova observação $\left(x_{N+1}, y_{N+1}\right)$, pode derivar-se $\hat{\beta}_{(N+1)}$ por actualização de $\hat{\beta}_{(N)}$ partindo das $N+1$ observações. É esta a solução dos mínimos quadrados recursivos (Goodwin & Sin, 1984; Ljung, 1978). Quando adicionamos uma nova observação $\left(x_{N+1}, y_{N+1}\right)$ ao conjunto de dados, a matriz X tem uma nova linha, o vector y uma nova componente e podemos escrever

$$\hat{\beta}_{(N+1)} = \left(\begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}\right)^{-1} \begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} y_{(N)} \\ y_{N+1} \end{bmatrix} \qquad 3.1$$

Fazendo

$$S_{(N)} = \left(X_{(N)}^T X_{(N)}\right) \qquad 3.2$$

temos

$$S_{(N+1)} = \left(X^T_{(N+1)} X_{(N+1)}\right) =$$

$$= \left(\left[X^T_{(N)} x^T_{N+1} \right] \begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix} \right) =$$

$$= \left(X^T_{(N)} X_{(N)} + x^T_{(N+1)} x_{(N+1)} \right) \qquad 3.3$$

Por

$$\begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} y_{(N)} \\ y_{N+1} \end{bmatrix} = X^T_{(N)} y_{(N)} + x^T_{N+1} y_{N+1} \qquad 3.4$$

e

$$S_{(N)} \hat{\beta}_{(N)} = X^T_{(N)} y_{(N)} \qquad 3.5$$

temos

$$\begin{bmatrix} X_{(N)} \\ x_{N+1} \end{bmatrix}^T \begin{bmatrix} y_{(N)} \\ y_{N+1} \end{bmatrix} = S_{(N)} \hat{\beta}_{(N)} + x^T_{N+1} y_{N+1} = \left(S_{(N+1)} - x^T_{N+1} x_{N+1} \right) \hat{\beta}_{(N)} + x^T_{N+1} y_{N+1}$$

Tendo em conta as igualdades 3.1, 3.2, 3.3, 3.4 e 3.5 podemos escrever

$$\hat{\beta}_{(N+1)} = S^{-1}_{(N+1)} \left(X^T_{(N)} y_{(N)} + x^T_{N+1} y_{N+1} \right) = S^{-1}_{(N+1)} \left(S_{(N)} \hat{\beta}_{(N)} + x^T_{N+1} y_{N+1} \right) =$$

$$= S^{-1}_{(N+1)} \left[\left(S_{N+1} - x^T_{N+1} x_{N+1} \right) \hat{\beta}_{(N)} + x^T_{N+1} y_{N+1} \right] =$$

$$= \hat{\beta}_{(N)} - S^{-1}_{(N+1)} x^T_{N+1} x_{N+1} \hat{\beta}_{(N)} + S^{-1}_{(N+1)} x^T_{N+1} y_{N+1} =$$

$$= \hat{\beta}_{(N)} + S^{-1}_{(N+1)} x^T_{N+1} \left(y_{N+1} - x_{N+1} \hat{\beta}_{(N)} \right)$$

o que conduz à seguinte formulação recursiva:

$$\begin{cases} S_{(N+1)} = S_{(N)} + x^T_{N+1} x_{N+1} \\ \gamma_{(N+1)} = S^{-1}_{(N+1)} x_{N+1} \\ e = y_{N+1} - x_{N+1} \hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} = \hat{\beta}_{(N)} + \gamma_{(N+1)} e \end{cases}$$

Conseguimos deste modo escrever $\hat{\beta}_{(N+1)}$ em função de $\hat{\beta}_{(N)}$ e da nova observação $\langle x_{N+1}, y_{N+1} \rangle$. Sendo a inversão da matriz $S_{(N+1)}$ uma operação custosa computacionalmente podemos utilizar o lema dado anteriormente para solucionar este problema. Designando $S^{-1}_{(N)} = \left(X^T_{(N)} X_{(N)} \right)^{-1}$ por $V_{(N)}$ e através do lema de inversão matricial obtemos

$$V_{(N+1)} = V_{(N)} - V_{(N)} x_{(N+1)} \left(1 + x^T_{(N+1)} V_{(N)} x_{(N+1)} \right)^{-1} x^T_{(N+1)} V_{(N)} =$$

$$= V_{(N)} - \frac{V_{(N)} x_{(N+1)} x_{(N+1)}^T V_{(N)}}{1 + x_{(N+1)}^T V_{(N)} x_{(N+1)}}$$

e

$$\begin{cases} V_{(N+1)} = V_{(N)} - \dfrac{V_{(N)} x_{(N+1)} x_{(N+1)}^T V_{(N)}}{1 + x_{N+1}^T V_{(N)} x_{(N+1)}} \\ \gamma_{(N+1)} = V_{(N+1)} x_{(N+1)} \\ e = y_{N+1} - x_{N+1} \hat{\beta}_{(N)} \\ \hat{\beta}_{(N+1)} = \hat{\beta}_{(N)} + \gamma_{(N+1)} e \end{cases}$$

Fixado um valor inicial $N = N_m$ podemos calcular recursivamente estimativas para valores sucessivos de N, só com base nas últimas equações.

Tendo vários modelos obtidos recursivamente para diferentes tamanhos de vizinhança, põe-se a questão de determinar qual deles vamos usar. O critério que vamos usar nesta tese tem como base as estimativas de erro *leave-one-out* de validação cruzada.

3.2. A estatística PRESS

Consideremos um conjunto de N dados para o qual deixamos de lado a primeira observação e usamos as restantes $N-1$ observações para estimar o valor da variável objectivo nessa mesma observação. Num segundo passo, deixamos de lado a segunda observação e usamos as restantes $N-1$ observações para estimar o valor da variável objectivo correspondente à segunda observação. Procedendo deste modo, para o i-ésimo passo, deixamos de lado a i-ésima observação e usamos as restantes $N-1$ observações para estimação do valor da variável objectivo nessa mesma observação posta de lado. Obtemos assim, para $i = 1,...,N$ um conjunto de N erros de previsão ou resíduos *leave-one-out* $y_i - \hat{y}^{-i} = e_{i,-i}$. Estes resíduos são erros de previsão com $\hat{y}^{-i}$ independente de y_i. Então, deste modo, a observação y_i não foi usada simultaneamente para ajustamento e previsão, sendo este o verdadeiro teste de validação cruzada.

A previsão $\hat{y}^{-i}$ é a função de regressão calculada em $x = x_i^T$, mas onde y_i foi posto de lado e não foi utilizado para obter os coeficientes. Temos $\hat{y}^{-i} = x_i^T \hat{\beta}^{-i}$ onde $\hat{\beta}^{-i}$ é o vector dos coeficientes calculado sem o uso da i-ésima observação. Para cada

conjunto dos pontos vizinhos de um caso de teste utilizado vamos ter N resíduos *leave-one-out* associados. Estes resíduos são importantes na medida em que se tem informação na forma de N validações para as quais a amostra de ajustamento tem dimensão $N-1$ em cada caso. Eles dão medidas separadas da estabilidade da regressão e podem ajudar o analista a isolar os dados ou observações que têm uma grande influência no resultado da regressão. Contudo, é computacionalmente pesado obter os resíduos *leave-one-out* pois isso implica a construção de N modelos para cada caso de teste. Existe um procedimento estatístico para os modelos lineares que permite calcular a estimativa de erro *leave-one-out* de validação cruzada de uma forma vantajosa. Trata-se da estatística PRESS (Allen, 1974). Com a estatística PRESS podem calcular-se os erros *leave-one-out* evitando a validação cruzada lenta para a qual o procedimento *leave-one-out* é repetido N vezes. Esta estatística consiste nos seguintes passos:

1. Paralelamente ao cálculo do vector de parâmetros $\hat{\beta}$ obtemos a matriz $H = X\left(X^T X\right)^{-1} X^T$. Esta matriz é simétrica, idempotente e permite calcular as estimativas de regressão $\hat{y}$ a partir dos valores y de treino: $\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T y = Hy$.

Portanto, podemos escrever o vector de resíduos como $e = y - \hat{y} = y - X\hat{\beta} = y - X\left(X^T X\right)^{-1} X^T y = \left[I - H\right]y$

e a soma dos quadrados dos resíduos como $e^T e = y^T \left[I - H\right]^2 y = y^T \left[I - H\right]y = y^T Py$ uma vez que I - H é também simétrica e idempotente e em que P é designada por matriz projecção.

2. Em seguida, determinamos o vector e de resíduos para o qual o i-ésimo termo é $e_i = y_i - x_i^T \hat{\beta}$.

3. Finalmente calculamos o resíduo *leave-one-out* $e_{i,-i} = \dfrac{e_i}{1 - H_{ii}}$ sendo H_{ii} o i-ésimo termo diagonal da matriz H.

A matriz H tem duas propriedades importantes:

$I - tr(H) = p$, o número de parâmetros do modelo;

II – Para um modelo com um termo constante $\dfrac{1}{N} \leq H_{ii} \leq 1$.

A propriedade *I* implica que $\sum_{i=1}^{N} \frac{Var(\hat{y})}{\sigma^2} = p$, um resultado que indica que independentemente de σ^2, a variância da previsão, somada sobre todos os pontos, iguala o número de parâmetros do modelo.

A propriedade *II* implica que $\frac{1}{N} \leq \frac{Var(\hat{y}(x_i))}{\sigma^2} \leq 1$, o que sugere que a precisão na previsão num ponto não é pior que a variância do erro numa observação, isto é, $Var(\hat{y}(x_i)) \leq \sigma^2$.

Uma observação muito interessante deve ser feita. A quantidade de H_{ii} é claro, a diagonal da matriz H e, independentemente de σ^2, representa a variância da previsão. Os pontos cuja previsão é pobre (H_{ii} próximo de 1) são aqueles em que o resíduos *leave-one-out* é um exagero do resíduo ordinário.

Em seguida vamos descrever como se pode deduzir a fórmula da estatística PRESS. Cálculos matriciais mostram que $X^T X - x_j x_j^T = X_{-j}^T X_{-j}$ onde $X_{-j}^T X_{-j}$ é obtida por eliminação da j-ésima linha da matriz $X^T X$.

Usando o lema de inversão matricial temos:

$$\left(X_{-j}^T X_{-j}\right)^{-1} = \left(X^T X - x_j x_j^T\right)^{-1} = \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_j x_j^T \left(X^T X\right)^{-1}}{1 - H_{jj}}$$

e

$$\hat{\beta}^{-j} = \left(X_{-j}^T X_{-j}\right)^{-1} X_{-j}^T y_{-j} =$$

$$= \left[\left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} x_j x_j^T \left(X^T X\right)^{-1}}{1 - H_{jj}}\right] X_{-j}^T y_{-j}$$ onde y_{-j} é obtido por eliminação da j-ésima observação do vector *y*.

Tendo em conta a última igualdade podemos escrever:

$$e_j^{loo} = y_j - x_j^T \hat{\beta}^{-j} =$$

$$= y_j - x_j^T \left[\left(X^T X\right)^{-1} + \frac{(X^T X)^{-1} x_j x_j^T \left(X^T X\right)^{-1}}{1 - H_{jj}}\right] X_{-j}^T y_{-j} =$$

$$= y_j - x_j^T \left(X^T X\right)^{-1} X_{-j}^T y_{-j} - \frac{H_{jj} x_j^T \left(X^T X\right)^{-1} X_{-j}^T y_{-j}}{1 - H_{jj}} =$$

$$= \frac{(1-H_{jj})y_j - x_j^T (X^T X)^{-1} X_{-j}^T y_{-j} + H_{jj} x_j^T (X^T X)^{-1} X_{-j}^T y_{-j} - H_{jj} x_j^T (X^T X)^{-1} X_{-j}^T y_{-j}}{1-H_{jj}} =$$

$$= \frac{(1-H_{jj})y_j - x_j^T (X^T X)^{-1} X_{-j}^T y_{-j}}{1-H_{jj}} =$$

$$= \frac{(1-H_{jj})y_j - x_j^T (X^T X)^{-1} (X^T y - x_j y_j)}{1-H_{jj}} =$$

$$= \frac{(1-H_{jj})y_j - x_j^T (X^T X)^{-1} X^T y + x_j^T (X^T X)^{-1} x_j y_j}{1-H_{jj}} =$$

$$= \frac{(1-H_{jj})y_j - \hat{y}_j + H_{jj} y_j}{1-H_{jj}} =$$

$$= \frac{y_j - H_{jj} y_j - \hat{y}_j + H_{jj} y_j}{1-H_{jj}} =$$

$$= \frac{y_j - \hat{y}_j}{1-H_{jj}} =$$

$$= \frac{e_j}{1-H_{jj}}$$

onde $X_{-j}^T y_{-j} + x_j y_j = X^T y$ e $x_j^T (X^T X)^{-1} X^T y = \hat{y}_j$.

O sistema de equações de mínimos quadrados recursivos actualiza a matriz $V_{(N)}$ e o vector $\hat{\beta}_{(N)}$ que tornam possível, pela formulação anterior, o cálculo de erros de validação cruzada *leave-one-out* evitando a identificação de um novo modelo. Sendo $e^{loo}(N) = \{e_j^{loo}(N)\}$ com $j = 1,...,N$, o vector que contém todos os erros *leave-one-out* associados ao modelo identificado com N vizinhos, o erro quadrático médio de validação cruzada é igual a $\frac{\sum_{j=1}^{N} (e_j^{loo}(N))^2}{N}$.

Assim, munidos desta estimativa de erro de um modelo, poderemos usar como critério de escolha da vizinhança o valor mínimo do erro quadrático médio de validação cruzada. Esta avaliação é feita no conjunto limitado das vizinhanças consideradas para cada caso de teste.

Em resumo, o uso dos mínimos quadrados recursivos em conjunto com a estatística PRESS permitem-nos escolher de forma computacionalmente eficiente, qual o tamanho de vizinhança "ideal" para construir um modelo linear local, que melhor se adeque a cada caso de teste.

Capítulo 4

APLICAÇÃO

4.1 Descrição do método implementado

Nesta secção iremos descrever o método de regressão linear local implementado no contexto da tese, e que foi posteriormente avaliado num conjunto de domínios de regressão. De entre as várias alternativas existentes no contexto da regressão local, que foram mencionadas anteriormente, foi necessário fazer algumas escolhas. O método implementado tem as seguintes características principais:

— Os conjuntos de dados aos quais é aplicado são multidimensionais e com observações desigualmente espaçadas no domínio

— Não é conhecida *a priori* informação extra sobre a função que se quer prever

— A função distância usada é a euclidiana e a função *kernel* é a uniforme

— As variâncias dos erros são heterogéneas (condição de heteroscedasticidade)

Este método foi implementado num programa escrito na linguagem C no ambiente Linux. A sintaxe do comando que permite executar este programa é a seguinte:

Ir<Nome do conjunto de dados><Limite inferior de vizinhos><Limite superior de vizinhos>

Para que os parâmetros dos modelos lineares locais possam ser estimados com alguma fiabilidade o segundo argumento do comando anterior deve ser superior ou igual a $J + 1$ em que J é o número de variáveis de regressão para o conjunto de dados que é utilizado.

O algoritmo recursivo dos mínimos quadrados, descrito anteriormente, parte de um valor inicial dos parâmetros do modelo. Assim, um dos problemas que inicialmente se colocou foi a inicialização do vector de parâmetros. Esta inicialização foi conseguida obtendo um modelo linear com o número inicial de vizinhos a ser tentado. Para obter um modelo linear com N casos temos que resolver as equações normais, $\hat{\beta}_{(N)} = \left(X_{(N)}^T X_{(N)}\right)^{-1} X_{(N)}^T y_{(N)}$. A resolução destas equações obriga à inversão da matriz $X_{(N)}^T X_{(N)}$ o que pode levantar problemas em situações de singularidade desta matriz.

No entanto, usando a metodologia da decomposição em valores singulares (SVD) (ver por exemplo Press et al., 1992) podemos ultrapassar estas situações. O método de decomposição em valores singulares é baseado no seguinte teorema da Álgebra Linear:

Teorema 4.1 *Qualquer matriz* $A_{[mxn]}$ *cujo número de linhas m é maior ou igual que o seu número de colunas n, pode ser escrita como o produto de uma matriz ortogonal* $U_{[mxn]}$, *uma matriz diagonal* $W_{[nxn]}$ *com elementos nulos ou positivos (os valores singulares) e a transporta de uma matriz ortogonal* $V_{[nxn]}$:

$A = U.W.V^T$ em que $U^T.U = V^T.V = I$.

Se a matriz A é quadrada de dimensão n, por exemplo, então U, V e W são matrizes quadradas com a mesma dimensão. As suas inversas são fáceis de calcular: U e V são ortogonais, e portanto as suas inversas são iguais às transpostas correspondentes; W é diagonal e a sua inversa é igual à matriz diagonal cujos elementos diagonais são os inversos dos elementos diagonais de W. Então segue-se que $A^{-1} = V.\left[diag\left(\frac{1}{w_j}\right)\right].U^T$

O único problema que pode surgir com esta construção é quando algum w_j é nulo ou bastante próximo de zero. A condição de uma matriz é definida como a razão de grandeza entre o maior e o menor valor dos w_j s. Uma matriz é singular se esta razão é infinita e é quase singular se esta razão é muito grande, isto é, próxima da precisão de vírgula flutuante da máquina. A melhor solução prática para este problema com esta técnica é substituir o inverso de w_j por zero quando w_j é muito próximo de zero.

Nesta tese usamos a rotina apresentada no livro *Numerical Recipes in C* (Press et al., 1992), que permite obter um modelo linear sem a exigência de não singularidade mencionada acima.

O algoritmo que serviu como base para a implementação do programa descrito é o seguinte:

$i = 1$;

Enquanto (i < Número de casos de teste){

Calcular a distância do i-ésimo caso de teste a cada caso de treino;

Ordenar o conjunto de treino de acordo com as distâncias anteriores;

m = número inicial de vizinhos;

Efectuar a regressão com base nos m primeiros casos do conjunto ordenado anterior, de modo a obter a inicialização do vector de parâmetros;

Calcular o erro *leave-one-out* correspondente e guardá-lo em memória;

M = número máximo de vizinhos;

m=m+1;

SE=0;

Enquanto (m<=M){

Efectuar a regressão com base nos m primeiros casos do conjunto ordenado anterior, usando as fórmulas do algoritmo recursivo dos mínimos quadrados;

Calcular o erro *leave-one-out* correspondente e guardá-lo em memória;

Calcular e guardar em memória o erro quadrático médio de validação cruzada *leave-one-out* correspondente à regressão no i-ésimo caso de teste com base em m casos de treino;

m=m+1;

}

Seleccionar o valor MIN de vizinhos que minimiza o erro quadrático médio de validação cruzada *leave-one-out* para o i-ésimo caso de teste;

Usar o modelo correspondente a MIN vizinhos para obter a previsão para o i-ésimo caso de teste;

Calcular e guardar em memória o erro quadrático para o i-ésimo caso de teste com base em MIN casos de treino e incrementar a variável SE;

}

Calcular o erro quadrático médio MSE para o conjunto total de validação;

4.2 Metodologia experimental

Com o objectivo de avaliar o desempenho do método foram utilizados diferentes conjuntos de treino e correspondentes conjuntos de teste. Fixado o conjunto de teste e com o mesmo conjunto de treino as previsões obtidas por dois métodos diferentes foram comparadas através de um teste de amostras emparelhadas. Para cada observação $\langle x_i, y_i \rangle$ do conjunto de validação $\{\langle X_i, Y_i \rangle\}_{i=1}^{N}$ calcularam-se $EQ_A = \left(Y_i - \hat{f}_A(X_i)\right)^2$ e

$EQ_B = \left(Y_i - \hat{f}_B(X_i)\right)^2$ onde $\hat{f}_A(X_i)$ e $\hat{f}_B(X_i)$ representam, respectivamente, a resposta obtida no elemento X_i pelo método A e B.

Em seguida calcularam-se os valores das variáveis $D_{AB_i} = EQ_{A_i} - EQ_{B_i}$, obtendo a realização $\{d_{AB_i}\}_{i=1}^{N}$ da amostra $\{D_{AB_i}\}_{i=1}^{N}$ da variável $D_{AB} = EQ_A - EQ_B$

Testou-se então $H_0 : \mu_{DAB} = 0$ contra $H_1 : \mu_{DAB} \neq 0$

Pelo teorema do limite central a variável $\overline{D_{AB}}$ segue aproximadamente uma lei normal de média μ_{DAB} e variância $\frac{1}{N}\sigma^2_{D_{AB}}$.

Sob a hipótese H_0 e porque a amostra é grande podemos assumir que

$$Z = \frac{\overline{D_{AB}}}{\frac{1}{\sqrt{N}} S_{D_{AB}}}$$

segue aproximadamente uma lei normal e reduzida.

Rejeitamos H_0 para um nível de significância α se

$$\left| \frac{\overline{d_{ABi}}}{\frac{1}{\sqrt{N}} S_{D_{AB}}} \right| \geq z_\alpha$$

onde z_α é tal que $P(Z \leq z_\alpha) = 1 - \frac{\alpha}{2}$

4.3 Descrição dos conjuntos de dados

Na tabela abaixo estão os conjuntos de dados e as suas principais características:

Conjunto de dados	Variáveis	Dimensão de treino	Dimensão de teste
P_abalone	8	3133	1044
Housing	14	300	206
Stock	10	600	350
Gate	11	300	150
D2a	3	1000	500
I	15	1000	555

Tabela 3.1: Descrição dos conjuntos de dados

Abalone é o nome comum para certos moluscos marinhos encontrados em rochas de mares quentes. Pode prever-se a idade do *abalone* a partir de medições físicas. A idade do *abalone* é determinada, cortando a concha através do cone e contando o número de anéis através de um microscópio – uma tarefa aborrecida e demorada. No conjunto de dados *P_abalone* são usadas outras medições que são mais fáceis de obter, com o objectivo de obter um modelo de regressão que possa ser usado para prever a idade dos *abalones*. Informação adicional tal como os padrões de clima e localização (disponibilidade de alimentos) são necessários para resolver o problema. Os dados resultaram de um estudo original (Warwick Nash & al., 94). São dados atributos nome e tipo, unidade de medida e uma breve descrição. A variável nominal correspondente ao sexo foi ignorada dado que na regressão linear só fazem sentido variáveis contínuas. O número de anéis é o valor objectivo a prever.

Nome	Tipo	Medida	Descrição
Comprimento	Contínuo	mm	Comprimento máximo da concha
Diâmetro	Contínuo	mm	Diâmetro da concha
Height	Contínuo	mm	Altura do abalone incluindo a concha
Whole	Contínuo	gr	Peso do abalone incluindo a concha
Shucked	Contínuo	gr	Peso do abalone sem concha
Víscera	Contínuo	gr	Peso da tripa após sangramento
Concha	Contínuo	gr	Peso da concha depois de seca
Anéis	Contínuo		+ 1,5 dá-nos a idade do abalone

Tabela 3.2: Variáveis do primeiro conjunto de dados

O conjunto de dados *housing* tem como origem o trabalho de Harrison e Rubinfeld (1978). Estes autores descreveram um problema de regressão em que se tenta prever o valor monetário das casas em Boston com base numa série de outras variáveis relativas a cada casa. O seu objectivo era o de entender se a poluição devido à concentração de óxido de nitrogénio tinha algum efeito nestes valores. Com esse objectivo eles recolheram informação respeitante a 506 tipos de casas em diferentes áreas.

As variáveis medidas foram:

MV (variável objectivo): valor médio das casas em milhares de dólares
CRIM: índice de criminalidade
ZN: percentagem de área reservada para loteamentos
INDUS: percentagem de actividade económica não retalhista
CHAS: 1 se perto do rio Charles, 0 em caso contrário
NOX: concentração de óxido de nitrogénio (NOX) em pphm

RM :número médio de quartos
AGE: percentagem de casas anteriores a 1940
DIS: distância pesada aos principais centros de emprego
RAD: acessibilidade às vias de cintura da cidade
TAX: imposto municipal da área
P/T: rácio aluno/professor
B: percentagem de população negra
LSTAT: percentagem de população de extracto social mais baixo

Tabela 3.3: Variáveis do segundo conjunto de dados

Quanto ao conjunto *stock*, os dados contêm informação sobre as cotações de acções de 10 companhias aéreas diferentes. O objectivo é obter um modelo que permita prever o valor das acções de uma companhia em função do valor das outras.

Quanto ao problema *gate* ele diz respeito a dados sobre falhas observadas em barragem hidroeléctricas.

O problema *d2a* é um conjunto de dados artificial com duas variáveis e com um alto grau de não linearidade.

Finalmente, o problema i diz respeito a dados sobre uma simulação de um ambiente com várias empresas em concorrência directa num determinado mercado. Os dados provêm de uma simulação usando aproximações multi-agente.

4.4 Comparação com o método de regressão global

Nesta secção apresentamos os resultados de uma comparação entre o método de regressão local implementado com a alternativa de obter um modelo de regressão linear global através do método dos mínimos quadrados. O objectivo desta comparação é o de verificar qual a diferença entre considerar o conjunto total de treino para ajustamento (regressão linear global) e o caso em que se utiliza o método de regressão linear local, no qual só uma vizinhança variável para cada caso de teste é usada a partir do conjunto de treino. Como já foi descrito antes, o método local deverá ser considerado como uma alternativa eficiente nos casos em que o método global não é praticável.

Para cada conjunto de dados foi medida uma estatística de erro, o erro quadrado médio (*MSE*):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}(X_i)\right)^2$$

Os valores obtidos para *MSE* foram os seguintes:

Erro	*P_abalone*	*Housing*	*Stock*	*Gate*	*D2a*	*I*
MSE (RG)	4.62	366.15	26.89	0.00018	166.30	415
MSE (RL)	4.42	6626.15	16.53	0.00013	35.64	318.18

RG – regressão linear global

RL – regressão linear local

Tabela 3.4: Regressão global vs. Regressão local

Quando se consideram os limites de vizinhança correspondentes dados pela tabela:

Limite de vizinhança	*P_abalone*	*Housing*	*Stock*	*Gate*	*D2a*	*I*
Valor mínimo	60	20	220	20	5	20
Valor máximo	80	40	240	40	25	40

Tabela 3.5: Intervalos das vizinhanças tentadas

Fez-se então o seguinte teste de hipóteses para um nível de 5%, ou seja, $z_a \cong 1.96$:

$H_0 : \mu_{DAB} = 0$ contra $H_1 : \mu_{DAB} \neq 0$ sendo A o método de regressão linear global e B o método de regressão linear local e obtiveram-se os seguintes resultados:

Conjunto de dados	Valor de z	Decisão
P_abalone	2.01	Rejeita-se H_0
Housing	9.16	Rejeita-se H_0
Stock	-6.94	Rejeita-se H_0
Gate	5.17	Rejeita-se H_0
D2a	-13.42	Rejeita-se H_0
I	-14.8	Rejeita-se H_0

Tabela 3.6: Decisão final na comparação com o método de regressão global

Como seria de esperar, o método de regressão clássica é acentuadamente pior que o método de regressão local. No entanto, com o conjunto de dados H*ousing* os resultados foram diferentes do geral. Uma possível explicação para esta situação é o facto de no conjunto *Housing* existirem variáveis independentes com escalas de valores bastante diferentes, o que poderá causar uma distorção da medida de distância entre duas observações. No sentido de testar esta hipótese procedeu-se à normalização dos valores das variáveis no conjunto de dados *Housing*, criando um novo conjunto que chamamos *Housing_normalizado*. A normalização de cada variável foi levada a cabo subtraindo cada valor nos dados originais, pela média da variável e dividindo o resultado pelo respectivo desvio padrão. Os resultados obtidos neste novo conjunto de dados foram os seguintes:

Erro	*Housing normalizado*
MSE (RG)	366.15
MSE (RL)	2575.45

Tabela 3.7: Erros relativos ao conjunto *housing* normalizado

Estes resultados confirmam que uma das explicações para os maus resultados da regressão linear local é a diferente escala de valores das variáveis. No entanto, mesmo usando os dados normalizados, os resultados da regressão linear local são desapontadores. Isto quer dizer que a nossa hipótese não explica completamente a má *performance* no conjunto *Housing*. Analisando mais atentamente o resultado da regressão linear global neste problema, verificamos que este modelo explica cerca de

90% da variância dos dados. Isto quer dizer que este problema de regressão é praticamente "linear" o que constitui a explicação mais plausível para a diferença de resultados entre a regressão linear local e global. Este problema serve também para chamar a atenção de que quando os dados são praticamente lineares, não faz sentido tentar usar modelos mais complexos, como é o caso da regressão linear local.

4.5 Comparação com a alternativa de vizinhança fixa

Nesta secção descrevemos uma comparação entre o método de regressão linear local implementado e a alternativa de usar uma vizinhança fixa (em vez de escolher o tamanho da vizinhança para cada caso de teste). Nesta comparação foram considerados três tipos de vizinhança fixa:

I) Vizinhança fixa de tamanho igual ao limite mínimo de vizinhos usado para o conjunto de dados correspondente no método de regressão linear local

II) Vizinhança fixa de tamanho igual à média aritmética dos limites mínimo e máximo usados para o conjunto de dados no método de regressão linear local

III) Vizinhança de tamanho igual ao limite máximo de vizinhos usado para o conjunto de dados no método de regressão linear local

Relativamente aos três tipos de vizinhanças descritos antes obtiveram-se os resultados dos seis quadros abaixo, igualmente para um nível de significância de 5%:

Erro	*P_abalone*	*Housing*	*Stock*	*Gate*	*D2a*	*I*	*Housing normalizado*
MSE (RG)	4.42	6626.15	16.53	0.00013	35.64	318.18	2575.45
MSE (RL)	4.55	6616.60	16.99	0.00013	34.70	318.53	2947.42

Tabela 3.8: Regressão local vs. Vizinhança fixa I

Conjunto de dados	Valor de z	Decisão
P_abalone	-2.95	Rejeita-se H_0
Housing	1.98	Rejeita-se H_0
Stock	5.63	Rejeita-se H_0
Gate	0.49	Não se rejeita H_0
D2a	1.52	Não se rejeita H_0
I	1.45	Não se rejeita H_0
Housing normalizado	-6.71	Rejeita-se H_0

Tabela 3.9: Decisão final na comparação com a vizinhança fixa I

Erro	*P_abalone*	*Housing*	*Stock*	*Gate*	*D2a*	*I*	*Housing normalizado*
MSE (RG)	4.42	6626.15	16.53	0.00013	35.64	318.18	2575.45
MSE (RL)	4.52	7196.17	16.31	0.00013	70.61	940.35	722.84

Tabela 3.10: Regressão local vs. Vizinhança fixa II

Conjunto de dados	Valor de z	Decisão
P_abalone	-1.85	Rejeita-se H_0
Housing	-0.17	Não se rejeita H_0
Stock	-0.30	Não se rejeita H_0
Gate	-0.35	Não se rejeita H_0
D2a	-6.33	Rejeita-se H_0
I	-0.49	Não se rejeita H_0
Housing normalizado	7.86	Rejeita-se H_0

Tabela 3.11: Decisão final na comparação com a vizinhança fixa II

Erro	*P_abalone*	*Housing*	*Stock*	*Gate*	*D2a*	*I*	*Housing normalizado*
MSE (RG)	4.42	6626.15	16.53	0.00013	35.64	318.18	2575.45
MSE (RL)	4.43	4136.43	16.72	0.00012	94.59	441.99	208.47

Tabela 3.12: Regressão local vs. Vizinhança fixa III

Conjunto de dados	Valor de z	Decisão
P_abalone	-0.12	Não se rejeita H_0
Housing	1.50	Não se rejeita H_0
Stock	-2.87	Rejeita-se H_0
Gate	-0.33	Não se rejeita H_0
D2a	-8.05	Rejeita-se H_0
I	-5.66	Rejeita-se H_0
Housing normalizado	10.72	Rejeita-se H_0

Tabela 3.13: Decisão final na comparação com a vizinhança fixa III

Destas experiências comparativas pode deduzir-se que o método de regressão linear local não é em geral muito diferente dos 3 métodos de vizinhança fixa, mas em certos casos é mais eficiente. Isto parece mostrar que é preferível considerar um ajustamento linear local, em vez de usar uma vizinhança fixa. A vizinhança fixa tem o inconveniente de escolher observações de treino que devem ser excluídas pelo facto de distorcerem em muitas situações a nossa previsão. Isto torna-se evidente em conjuntos

de dados para os quais a distribuição das observações é muito irregular. No entanto, também observamos que em alguns conjuntos de dados as alternativas fixas foram superiores. Tendo em conta que os valores fixos que foram considerados, estão incluídos no intervalo de vizinhanças que é tentado pelo nosso método, isto pode ser uma indicação que o método de escolha através da estimativa de erro LOOCV poderá não estar a fornecer estimativas do erro tão fiáveis quanto o desejável.

Capítulo 5

CONCLUSÃO

Como foi verificado é possível usar eficientemente a regressão linear num contexto local. Este método tem a vantagem de fácil compreensão e interpretação e é bastante flexível em relação ao tipo de superfícies que podem ser modeladas. Foi verificada na prática a importância dos conjuntos de dados utilizados. Também é claro que o tamanho das amostras é importante. Quando o tamanho da amostra é muito pequeno, não é possível calcular medidas adequadas do erro nos resultados de regressão. Um bom teste estatístico de comparação também deve ter em conta este facto.

Em muitas situações, as dificuldades com a análise de regressão são resultado da falha de uma ou mais suposições. Em particular, o modelo de regressão linear múltipla é analisado sob a hipótese de que as variáveis de regressão são medidas sem erro. Se um erro excessivo na medição das variáveis existe, as estimativas dos coeficientes de regressão podem ser bastante afectadas. Provavelmente a limitação mais séria num conjunto de dados é a impossibilidade de juntar dados sobre todos os regressores potencialmente importantes. Isto pode acontecer porque o analista não sabe quais são os regressores relevantes.

Como trabalho futuro uma alternativa que podia ser tentada era a de escolher diferentes funções distância e funções de pesagem assim como utilizar outros critérios para a escolha da vizinhança.

Bibliografia

Aha, D. (1997): Lazy Learning, edited by D. Aha. Kluwer Academic Publishers.

Aha, D. W. (1989). Incremental, instance-based learning of independent and graded concept descriptions. In Sixth International Machine Learning Workshop, pages 387–391. Morgan Kaufmann, San Mateo, CA.

Aha, D. W. (1990). A Study of Instance-Based Algorithms for Supervised Learning Tasks: Mathematical, Empirical, amahwah, Nj.nd Psychological Observations. PhD dissertation, University of California, Irvine, Department of Information and Computer Science.

Aha, D. W. and McNulty, D. M. (1989). Learning relative attribute weights for instance-based concept descriptions. In 11th Annual Conference of the Cognitive Science Society, pages 530-537. Lawrence Erlbaum Associates, Mahwah, NJ.

Allen (1974): The relationship between variable and data augmentation and a method of prediction. Technometrics, 16, 125-127.

Atkeson, C. G. Moore, A. W., Schaal, S. (1997): Locally Weighted Learning, Artificial Intelligence Review, 11, 11-73. Special Issue on lazy learning, Aha, D. (Ed.).

Brillinger. D. (1977). Discussion of a paper of Stone. Ann. Statist. 5, 622-623.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. J. Amer. Statis. Assn. 83, 596-610.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, Journal of the American Statistical Association, 74, 829-836.

Cleveland, W. S. and Loader, C. (1994a). Computational methods for local regression. Technical Report 11, AT&T Bell Laboratories, Statistics Department, Murray Hill, NJ. http://netlib.att.com/netlib/att/stat/doc/.

Cleveland, W. S. and Loader, C. (1994b). Local fitting for semiparametric (nonparemetric) regression: Comments on a paper of Fan and Marron. Technical Report 8, AT&T Bell Laboratories. Statistics Department, Murray Hill, NJ.
http://netlib.att.com/netlib/att/stat/doc/ , 94.8. ps, earlier version is 94.3. ps.

Cleveland, W. S. and Loader; C. (1994 c). Smoothing by local regression: Principles and methods. Technical Report 95.3, AT&T Bell Laboratories, Statistics Department, Murray Hill, NJ. http://netlib.att.com/netlib/att/stat/doc/.

Cleveland, W., Loader, C. (1995): Smoothing by local Regression: Principles and Methods (with discussion), Computational Statistics.

De Forest, E.L. (1873). On some methods of interpolation applicable to the graduation of irregular series. Annual Report of the Board of Regents of the Smithsonian Institution for 1871, 275-339.

De Forest, E. L. (1874). Additions to a memoir on methods of interpolation applicable to the graduation of irregular series. Annual Report of the Board of Regents of the Smithsonian Institution for 1873, 319-353.

Fan. J. (1993). Local linear regression smoothers and their minimax efficiencies. Ann. Statist. 21, 196-216.

Fan. J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. The Annals of Statistics, 20(4): 2008-2036.

Fan. J. na Gijbels, I. (1994). Censored regression: local linear approximations and their applications. J. Amer. Statist. Assn. 89,560-570.

Fan, J., Marron, (1993): Comment on Hastie & Loader (1993). Statistical Science, 8, 120-143.

Fan. J. and Marron. J.S. (1994): Rejoinder to discussion of Cleveland and Loader.

Farmer, J.D. and Sidorowich, J.J. (1988a). Exploiting chaos to predict the future and reduce noise. In Lee, Y.C., editor, Evolution, Learning and Cognition, pages 277-??? World Scientific Press, NJ. also available as Technical Report LA-UR-88-901, Los Alamos National Laboratory, Los Alamos, New Mexico.

Farmer, J.D. and Sidorowich, J.J. (1988b). Predicting chaotic dynamics. In Kelso, J. A. S., Mandell, A.J., and Schlesinger, M. F., editors, Dynamic Patterns in Complex Systems, pages 265-292. World Scientific, NJ.

Friedman, J.H. and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assn. 76, 817-823.

Friedman, J. H. (1994). Flexible metric nearest neighbor classification.
http://playfair.stanford.edu/reports/friedman/.

Ge., Z., Cavinato, A.G., and Callis, J.B. (1994). Noninvasive spectroscopy for monitoring cell density in a fermentation process. Analytical Chemistry, 66:1354-1362.

Gianluca Bontempi: Local Learning Techniques for Modeling, Prediction and Control.

Goodwin & Sin, 1984. Adaptive Filtering Prediction and Control. Prentice-Hall.

Gram, J.P. (1883). Uber Entwickelung reeller Functionen in Reihen mittelst der Methode der kleinsten Quadrate. J. Math. 94, 41-73.

Hardle, W. (1990). Applied Nonparametric Regression. Oxford University Press, Oxford.

Harrison & Rubinfeld, 1978. Hedonic prices and the demand for clean air. J. Environ. Economics and Management, vol.5: 81-102.

Hastie, T. and Loader, C. (1993). Local regression: Automatic kernel carpentry. Statistical Science, 8(2):120-143.

Hastie, T. J. and Tibshirani, R.J. (1990). Generalized Additive Models. Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R.J. (1994). Discriminant adaptive nearest neighbor classification. ftp://playfair.Stanford.EDU/pub/hastie/dann.ps.Z.

Henderson, R. (1924). A new method of graduation. Actuarial Soc. Amer. 25, 29-39.

Hoem, J.M. (1983). The reticent trio: some little-known early discoveries in life insurance mathematics by L.H.F. Oppermann, T.N. Thiele and J.P. Gram. Inter., Stat. Rev. 51, 213-221.

Katkovnik, V. Ya. (1979). Linear and nonlinear methods of nonparametric regression analysis. Soviet Automatic Control. 5, 25-34.

Ljung, 1978. Convergence Analysis of parametric Identification Methods. IEEE Transactions on Automatic Control, 23(5), 770-783.

Loader, C. (1995). Local likelihood density estimation. Ann. Statist., to appear.

Macaulay, F.R. (1931). The Smoothing of Times Series. National Bureau of Economic Research, New York.

Mauro Birattari, Gianluca Bontempi and Hugues Bersini: Lazy Learning Meets the Recursive Leat Squares Algorithm.

Muller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. J. Amer. Statist. Assn. 82, 231-238.

Nadaraya, E.A. (1964): On estimating regression. Theory of Probability and its Applications, 9: 141-142.

Naes, T., Isaksson, T. (1992). Locally weighted regression in diffuse near-infrared transmittance spectroscopy. Applied Spectroscopy, 46(1).34-43.

Naes, T., Isaksson, T., and Kowalski, B.R. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. Analytical Chemistry, 62(7): 664-673.

Parzen, E. (1962): On estimation of a probability density function and mode. Annals Mathematical Statistics. 33, 1065-1076.

Raymond H. Myers: Classical and Modern Regression with Applications, Second Edition.

Rosenblatt, M. (1956): Remarks on some nonparametric estimates of a density function. Annals Mathematical Statistics, 27, 832-837.

Ruppert, D. and Wand. M.P. (1992). Multivariate locally weighted least squares regression. Ann. Statist. 22, No. 3.

Scott, D.W. (1992). Multivariate Density Estimation. Wiley, New York, NY.

Shishkin, J. Young, A.H., and Musgrave, J.C. (1967). The X-11 variant of the Census Method II seasonal adjustment program. Technical Paper 15, U.S. Bureau of the Census.

Spenser, J. (1904). On the graduation of the rates of sickness and mortality. J. Inst. Act. 38, 334-347.

Stanfill, C. (1987). Memory-based reasoning applied to English pronunciation. In Sixth National Conference on Artificial Intelligence, pages 577-581.

Stanfill, C. and Waltz. D. (1986). Toward memory-based reasoning. Communications of the ACM, 29(12): 1213-1228.

Stigler, S.M. (1978). Mathematical statistics in the early States: Ann. Statist. 6, 239-265.

Stone. C.J. (1977). Consistent nonparametric regression (with discussion). Ann. Statist. 5, 595-620.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. J. Amer. Statist. Assn. 82, 559-567.

Townshend, B. (1992). Nonlinear prediction of speech signals. In Casdagli and Eubank (1992), pages 433-453. Proceedings of a Workshop on Nonlinear Modeling and Forecasting. September 17-21, 1990, Santa Fe, New Mexico.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In Moody, J.E., Hanson, S.J. and Lippmann, R.P., editors, Advances In Neural Information Processing Systems 4, pages 831-838, Morgan Kaufman, San Mateo, CA.

Waltz. D.L. (1987). Applications of the Connection Machine. Computer, 20 (1): 85-97.

Wang. Z., Isaksson, T., and Kowalski, B.R. (1994). New approach for distance measurement in locally weighted regression. Analytical chemistry, 66(2):249-260.

Warwick Nash, TracySellers, Simon Talbot, Andrew Cawthorn & Wes Ford, 1994. The Population Biology of Abalone (Haliotis Species) in Tasmania. I. Blacklip Abalone (H.

Rubra) from the North Coast and Islands of Bass Strait. Fisheries Division, technical Report no 48 (ISSN 1034-3288).

Watson. G.S. (1964): Smooth Regression Analysis. Sankhya: the Indian Journal of Statistics, Series A, 26:359-372.

Whittaker, E.T. (1923). On a new method of graduation. Proc. Edinburgh Math. Soc. 41, 63-75.

William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery: Numerical Recipes in C. The Art of Scientific computing, Second Edition (Cambridge University Press).

Woolhouse, W.S.B. (1870). Explanation of a new method of adjusting mortality tables, with some observations upon Mr. Makeham's modification of Gompertz's theory. J. Inst. Act. 15, 389-410.

Apêndice A

Funções Distância

1) Distância Eucludiana

$$d_E(X,q) = \sqrt{\sum_{j=1}^{n}(X_j - q_j)^2} = \sqrt{(X-q)^T(X-q)}$$

2) Distância Euclidiana Ponderada Diagonalmente

$$d_m(X,q) = \sqrt{\sum_{j=1}^{n}\left[m_j(X_j - q_j)\right]^2} = \sqrt{(X-q)^T M^T M(X-q)} = d_E(M_X, M_q)$$

onde m_j é o factor escala para a j-ésima dimensão e M é uma matriz diagonal com $M_{jj} = m_j$

3) Distância Euclidiana Ponderada Completamente

$$d_M(X,q) = \sqrt{(X-q)^T M^T M(X-q)} = d_E(M_X, M_q)$$

onde M não é uma matriz diagonal, mas sim uma matriz arbitrária.Também é conhecida pela distância de Mahalanobis (Tou e Gonzalez, 1974; Weisberg, 1985)

4) Norma não Ponderada Lp (Métrica de Minkowski)

$$d_p(X,q) = \left[\sum_{j}^{n}\left|X_j - q_j\right|^P\right]^{\frac{1}{P}}$$

Apêndice B

Funções *kernel*

Define-se uma função *kernel* (ou de "pesagem") não negativa $K:\Re^n \times \Re^n \times \Re^+ \rightarrow \Re^+$ em que o primeiro argumento de entrada é n-dimensional, o segundo argumento é chamado o centro e o terceiro argumento é o comprimento da vizinhança. Definindo a função distância com o argumento de entrada e o centro $d:\Re^n \times \Re^n \rightarrow \Re^+$ a função *kernel* pode ser expressa como $K:\Re^+ \times \Re^+ \rightarrow \Re^+$ em que os dois argumentos são a distância d e o comprimento da vizinhança

A função *kernel* satisfaz duas condições:

$0 \leq K(x,q,B) \leq 1$

$K(q,q,B) = 1$

Por exemplo no caso unidimensional mais simples tanto a função rectangular de localidade (também chamada *kernel* uniforme)

$$K(x,q,B) = \begin{cases} 1 \ se \left\| x-q \right\| < \frac{B}{2} \\ 0 \ caso \ contrário \end{cases}$$

como a função suave de localidade:

$$K(x,q,B) = \exp\left\{ -\frac{(x-q)^2}{B^2} \right\}$$

satisfazem as condições anteriores.

Alguns exemplos de funções *kernel*:

1) Distância Inversa

$$K(d,B) = \frac{1}{\left(\frac{d}{B}\right)^P}$$

2) Distância Inversa Corrigida

$$K(d,B) = \frac{1}{1+\left(\frac{d}{B}\right)^{P}}$$

3) *kernel* Guassiana

$$K(d,B) = \exp\left(-\frac{d^2}{B^2}\right)$$

4) *kernel* Exponencial

$$K(d,B) = \exp\left(-\left|\frac{d}{B}\right|\right)$$

5) *kernel* Quadrática ou Epanechnikov

$$K(d,B) = \begin{cases} \left(1-\frac{d^2}{B}\right), & se\ |d| < B \\ 0, & caso\ contrário \end{cases}$$

6) *kernel* Tricubo

$$K(d,B) = \begin{cases} \left(1-\left|\frac{d}{B}\right|^3\right)^3, & se\ |d| < B \\ 0, & caso\ contrário \end{cases}$$

7) *kernel* Uniforme

$$K(d,B) = \begin{cases} 1, & se\ |d| < B \\ 0, & caso\ contrário \end{cases}$$

8) kernel Triangular

$$K(d,B) = \begin{cases} 1-\left|\frac{d}{B}\right|, & se\ |d| < B \\ 0, & caso\ contrário \end{cases}$$