Modelos Aditivos generalizados com o R

Henrique Silva Dallazuanna
Wagner Hugo Bonat *

2 de outubro de 2007

1 Introdução

Um modelo aditivo generalizado (Hastie and Tibishirani, 1986, 1990) é um modelo linear generalizado com um preditor linear envolvendo a soma de funções suavizadas das covariáveis.O modelo pode ser escrito da seguinte forma:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \quad (1)$$

Onde $\mu_i \equiv E(Y_i)$ sendo que $Y_i \sim$ alguma distribuição da familía exponencial.
Y_i é a variável resposta, X_i^* é uma coluna da matriz do modelo para os componentes parámetricos, θ é o vetor de parâmetros, e f_j são funções suaves das covariáveis x_k.O modelo permite flexibilidade na especificação da dependence da resposta com as covariáveis, usar modelos com especificações suaves das covariáveis pode trazer mais informação do que apenas o modelo paramétrico. Esta facilidade e conveniência vem acompanhada de dois problemas teóricos. Primeiro é necessário representar esta função suave de alguma maneira, segunda avaliar o ajuste desta função. Neste trabalho vai-se mostrar como pode ser representado um GAM usando regressão por splines, metodos de estimação por splines, e verificar como o grau de suavização para as funções f_j pode ser estimado por validação cruzada. O caso mais complicado que vamos tratar aqui é o modelo considerando duas componentes suavizadas univariadas. Além disso, os metódos presentes aqui não são os

*Graduandos Estatística-UFPR

mais apropriados para casos reais, porém são os mais simples para se explicar os métodos básicos.Vai-se apresentar também comandos do R para explicar as idéias básicas.

2 Funções Suaves Univariadas

A representação de uma função suave univariada é melhor introduzida considerando um modelo contendo apenas uma função suave e uma covariada.

$$y_i = f(x_i) + \epsilon_i \tag{2}$$

onde y_i é a variável resposta, x_i uma covariável, f uma função suave e ϵ_i é uma variável aleatória i.i.d $N(0, \sigma^2)$. Para simplificar mais ainda vamos considerar que x_i pertence ao intervalo [0,1].

2.1 Representando uma função suave por Regressão por Splines

Para estimar f, usando os métodos cobertos pela teoria de regressão linear e de modelos lineares generalizados, necessita que f seja representada na forma em (2) tornando-se um modelo linear. Isto pode ser feito escolhendo-se uma "base", definindo o espaço das funções para f (ou uma aproximação para ela).Ao escolher uma base deve-se escolher de tal forma que esta tenha uma função de base que possa ser assumida como totalmente conhecida. Se $b_j(x)$ é da j-ésima ordem tal função de base para f tem supostamente uma representação:

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j \tag{3}$$

para se obter os valores dos parametros β_j desconhecidos, substitui-se (3) em (2) e se obtem facilmente um modelo linear.

2.2 Um simples exemplo de polinômio de base

um simples exemplo, suponha que f é um polinômio de quarta ordem, contendo o espaço dos polinomios de quarta ordem. Uma base para este espaço

é $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$, $b_4(x) = x^3$ e $b_5(x)^4$, de modo que (3) torna-se:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_2 + x^3\beta_3 + x^4\beta_5$$

De modo que (2) torna-se:

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_2 + x^3\beta_3 + x^4\beta_5 + \epsilon_i$$

As bases polinomiais tendem a ser muito utéis para situações em que se tem interesse na propriedades de f na vizinhança de um único ponto especifico, mais quando a questão de interesse é f sobre os demais inteiros, as bases polinomiais tem alguns problemas. As bases splines tem boa performance na maior parte dos problemas de aproximação pode ser demonstrado suas boas propriedades teoricas.

2.3 Spline de base cúbica

Uma função univarida pode ser representada usando um spline cubico. Um spline cubico é uma curva composta de seções polinomiais também, só que permitem primeira e segunda derivadas. Os pontos de que juntam as seções são os Knots do spline. Para um spline convencional, os knots ocorrem onde tenha alguma referencia, mais para regressão por splines que é o interesse aqui, a localização dos knots deve ser escolhida. Tipicamente os knot são espalhados uniformemente através da escala dos valores observados x, ou pelos quantis da distribuição de x. Independente do método a ser usada a notação para descrever a localização dos Knot será denotada por $x_i^* : i = 1, \ldots, q-2$. Dada a localização dos knots, existem muitas maneiras, equivalentes, para escrever uma bases para um spline cúbico. Bases simples para serem usadas podem ser encontradas nos livros de Wahba (1990) e Gu (2002), embora a função de base são ligeiramente dificeis de escrever como explicado atraz. Para isto baes como: $b_1(x) = 1$, $b_2(x) = x$ e $b_{i+2} = R(x, x_i^*$ para $i = 1, \ldots q-2$ onde

$$R(x,z) = \frac{[(z-\frac{1}{2})^2 - \frac{1}{12}][(x-\frac{1}{2})^2 - \frac{1}{12}]}{4} - \frac{[(|x-z| - \frac{1}{2})^4 - \frac{1}{2}(|x-z| - \frac{1}{2}\ )^2 + \frac{7}{240}]}{24}$$

Para mais detalhes (ver Gu, 2002,p.37). Usar esta base para o spline cúbico significa transformar a equação (2) em um modelo linear $y = X\beta + \epsilon$, onde a i-ésima coluna da matrix do modelo é

$$X_i = [1, x_i, R(x_i, x_i^*), R(x_i, x_2^*), \ldots, R(x_i, x_{q-2}^*)] \tag{4}$$

Daqui o modelo pode ser estimado por minímos quadrados.

2.4 Exemplo ilustrativo

Agora consideraremos um exemplo ilustrativo. É dito frequentemente que um carro com maior capacidade de motor se desgastará mais lentamente que um carro com menor capacidade.A figura (1) mostra alguns dados para 19 motores volvo. O modelo em (2) parece ser adequado.

Colocar o gráfico do modelo