1 Spline

1.1 Extended linear model

Regression analysis often estimates the conditional expectation of the dependent variable Y given the independent variables X: $E(Y|X) = f(X)$. The form of the function f must be specified in order to carry out regression analysis. One convenient and most common way to represent $f(X)$ is to use a linear model which assumes a linear relationship between a dependent variable and one or more independent variables: $Y_i = \beta_0 + \beta_i X_i + \epsilon_i$. However, the assumption of linearity sometimes may not be adequate for describing the relationship. One way to extend this model is to replace the vector of the independent variable X with transformations of X, and then use linear models in this new space of transformed X (Friedman, Hastie, & Tibshirani, 2001). Let $h_m(X)$ denote the m_{th} transformation of X, $m = 1, ..., M$. The extended linear model is formulated as

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X), \tag{1}$$

where $\{h_m\}$ are called basis functions. If $h_m(x) = x, m = 1, ..., M$, then it recovers the original linear model. If $h_m(x)$ include powers of x, then the model becomes a polynomial regression model, which is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_p X_i^p + \epsilon_i$.

1.2 Truncated polynomials

Polynomial terms of high degree are hard to interpret and are highly unstable. The polynomial model suffers from its global nature that each observation affects the entire curve. A slight change in coefficients to achieve a functional form in one region may cause the function to change dramatically in remote regions (Friedman et al., 2001). An alternative to use high order polynomial terms to account for curvature is to divide the domain of X into smaller intervals and fit simpler polynomial models to each pieces. The points where the division occurs are called knots. A spline can be constructed by joining polynomials. The term spline comes from a flexible tools used by shipbuilders and draftsmen to draw smooth curves (Wegman & Wright, 1983). In general, we impose some constraints to a spline of degree D. The constraints include that the function is continuous, the function has $D - 1$ continuous derivatives, and the D_{th} derivative is constant between knots (Friedman et al., 2001). Mathematically, a spline of degree D with K knots can be expressed as

$$f(X) = \beta_0 + \beta_1 X + ... + \beta_D X^D + \sum_{k=1}^{K} b_k (x - \xi_k)_+^D, \tag{2}$$

where

$$(X-\xi)_+^D = \begin{cases} X & X < \xi \\ (X-\xi)^D & X \geq \xi \end{cases} \tag{3}$$

is called a truncated polynomial function. Its value only affects the spline to the right of the knot ξ. These splines are also known as regression splines. The value of parameters $b_1, ..., b_k, \beta_0, ..., \beta_D$ can be estimated using least squares criterion.

1.3 Natural cubic spline

The polynomials fit to data can be erratic near the boundary knots. (Friedman et al., 2001) summarized this through pointwise variance of spline fit by least squares and showed a clear explosion of variance near the boundaries in cubic splines. One solution is to add additional constraints that the function is linear beyond the boundary knots. Such splines are called natural splines. Assuming linearity near boundaries may introduce extra bias, but this approach is often considered to be reasonable since we have less information anyway (Friedman et al., 2001). A natural cubic spline frees up 2 degrees of freedom in both boundary points. We can represent a natural cubic spline with K knots using K basis functions. The basis functions are

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_{k+2}(X) = -d_{(K-1)}(X), \tag{4}$$

where

$$d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}. \tag{5}$$

1.4 B-spline

While splines constructed from the truncated polynomials are conceptually simple, they have some numerical flaws. If the knots are close then the polynomials terms will be similar for all observations and thus cause nearly singular matrix problems. Also, values of polynomial terms may be very large, which makes the approach unstable. One more attractive approach is to use B-spline basis. The term "B-spline" is short for basis spline (De Boor, De Boor, Mathématicien, De Boor, & De Boor, 1978). B-splines are defined by their order and number of interior knots. B-splines can be defined via recursion formula (De Boor, 1986). Consider a B-spline of order M with K interior knots. Let ξ_0 and ξ_{K+1} be two boundary knots. We first define an augmented knot sequence (Friedman et al., 2001) τ so that

- $\tau_1 \leq \tau_2 \leq ... \leq \tau_M \leq \xi_0$

- $\tau_{j+M} = \xi_j, \quad j = 1, ..., K$
- $\xi_{K+1} \leq \tau_{K+M+1} \leq \tau_{K+M+2} \leq ... \leq \tau_{K+2M}$.

For each of the augmented knots τ_i, $i = 1, ..., K + 2M$, a set of B-spline basis function of order m ($m \leq M$) $B_{i,m}$ can be defined recursively as follows: (Friedman et al., 2001)

$$B_{i,1}(x) = \begin{cases} 1 & \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

for $i = 1, ..., K + 2M - 1$.

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{m+i-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x) \tag{7}$$

for $i = 1, ..., k + 2M - m$. The B-spline basis are not colinear and the values are always between 0 and 1 so this approach is more stable than splines constructed from the truncated polynomials. B-spline is also more computationally efficient even when the number of knots K is large (Friedman et al., 2001).

1.5 Choose knots (cross validation)?

References

De Boor, C. (1986). *B (asic)-spline basics.* (Tech. Rep.). WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., & De Boor, C. (1978). *A practical guide to splines* (Vol. 27). Springer-Verlag New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York, NY, USA:.

Wegman, E. J., & Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association*, *78*(382), 351–365.