General Multivariate Spline Models

What if we don't believe that an additive spline model is appropriate?

We can instead use a multivariate spline model, with basis functions that depend on *all* the input variables.

One option is create a *tensor product* spline from splines for each variable. For two variables, we would use basis functions of the form

$$h_{1,j}(X_1)\, h_{2,k}(X_2)$$

where $h_{1,j}$ and $h_{2,k}$ are basis functions from the splines for the first and second input variables. The multivariate spline model is then

$$f(X) \;=\; \sum_{j=1}^{M_1} \sum_{k=1}^{M_2} \theta_{j,k}\, h_{1,j}(X_1)\, h_{2,k}(X_2)$$

where M_1 and M_2 are the numbers of basis functions for the two univariate splines. The multivariate spline has $M_1 M_2$ basis functions.

Exercise: Show that the model is the same regardless of *which* sets of basis functions are chosen for the univariate splines.

Characteristics of Tensor Product Splines

Suppose that the univariate splines for two input variables are unconstrained piecewise polynomials of degrees D_1 and D_2, with K_1 and K_2 knots.

Then the tensor product spline will be composed of rectangular pieces, with each piece a polynomial of degree D_1 in X_1 and D_2 in X_2.

Simplest example: Piecewise constant univariate basis functions ($D_1 = D_2 = 0$) produce a multivariate spline that is piecewise constant on rectangular regions.

The number of basis functions grows exponentially with dimension — for this example, with 10 input variables and 5 knots for each input, there are $6^{10} = 60,466,176$ basis functions! This isn't practical.

We might try to select a subset of basis functions to use — sort of like selecting a subset of input variables. The MARS (Multivariate Adaptive Regression Splines) procedure (See Section 9.4 in the text) works this way, with tensor products of linear splines.

Multivariate Smoothing Splines

We can bypass this "curse of dimensionality" problem by using multivariate smoothing splines, with a penalty based on partial derivatives.

The *thin plate splines* are one such possibility. In two dimensions, they use a penalty for a function $f(X)$ of

$$\lambda \int \int f_{11}(X_1, X_2)^2 + 2 f_{12}(X_1, X_2)^2 + f_{22}(X_1, X_2)^2 \, dX_1 \, dX_2$$

where $f_{ab}(X_1, X_2)$ is the partial derivative of $f(X_1, X_2)$ with respect to X_a and X_b.

Unfortunately, the tricks for fast computation don't work for multivariate smoothing splines. Computations with N training cases take time proportional to N^3 — a lot better than exponential, but not very good if $N = 10000$.

A Bayesian Approach — Gaussian Process Models

If we want to learn a function $f(x)$ from data using Bayesian methods, we need to have a prior distribution over such functions. *Gaussian processes* are one interesting class of prior distributions.

A Gaussian process is a distribution over functions, $f(x)$, for which the joint distribution of $f(x_1), \ldots, f(x_N)$, for any $x_1, \ldots, x_N$, is multivariate Gaussian.

We specify the Gaussian process by giving the mean of $f(x_i)$ as a function of x_i and the covariance of $f(x_{i_1})$ and $f(x_{i_2})$ as a function of x_{i_1} and x_{i_2}.

These finite-dimensional specifications fix the distribution over the infinite dimensional space of functions. Note that the covariance function must be such that the covariance matrices are always positive semi-definite.

Note: I'll also use the phrase "Gaussian process" to refer to "functions" in which there is some noise — ie, for which the values at x_{i_1} and x_{i_2} can differ even when $x_{i_1} = x_{i_2}$.

Bayesian Linear Regression as a Gaussian Process Model

Consider a linear regression model:

$$y_i \;=\; \beta_0 \;+\; \sum_{j=1}^{p} \beta_j\, x_{ij} \;+\; \epsilon_i$$

where ϵ_i is Gaussian noise with variance σ_ϵ^2.

We give the unknown parameters β_0 and β_j (for $j = 1, \ldots, p$) independent Gaussian priors with means of 0 and with variances of σ_0^2 and σ_j^2 (for $j = 1, \ldots, p$).

Considering the y_i, ϵ_i, and β_j as variables, but the x_{ij} as fixed, the y_i have a multivariate Gaussian distribution, since they are linear combinations of the Gaussian-distributed β_j and ϵ_i. So we can view this as a Gaussian process model.

The mean function for this Gaussian process is just $E(y_i) = 0$.

The covariance function is found from

$$\mathrm{Cov}(y_{i_1}, y_{i_2}) \;=\; \sigma_0^2 \;+ \sum_{j=1}^{p} x_{i_1 j}\, x_{i_2 j}\, \sigma_j^2 \;+\; \delta_{i_1 i_2}\, \sigma_\epsilon^2$$

where $\delta_{ab} = 1$ if $a = b$ and 0 if $a \neq b$.