Data with Nonlinear Relationships

The inputs, X, are often related to the response, Y, in a *nonlinear* way.

For a **regression** problem, we predict using $E(Y|X)$, and this may be a nonlinear function of X.

For a **classification** problem, the discriminant functions, $\delta_k(X)$ may be nonlinear. If we are producing probabilistic predictions, $\text{logit}(P(Y = 1|X))$ — maybe any other monotonic function of the probability — may be nonlinear.

We can handle such problems with k-NN, or other "nonparametric" methods. But we can also use a parametric model that can express functions $f(X)$ that are nonlinear in X.

Nonlinear Relationships From Linear Models

We can re-use the apparatus of linear models to model nonlinear relationships.

We define M *basis functions*, $h_m(X)$, and then model $f(X)$ as

$$f(X) \;=\; \sum_{m=1}^{M} \beta_m h_m(X)$$

where the β_m are parameters to be estimated from the training data — eg, by least squares or maximum likelihood, perhaps with a penalty.

If we let $M = p$ and define $h_m(X) = X_m$, this is just the usual linear model. (With another basis function defined as $h_m(X) = 1$, we can include an intercept.)

But we can also include basis functions such as

$$h_m(X) = X_2^3, \quad h_m(X) = X_1 X_2, \quad h_m(X) = \sin(X_1), \quad \text{etc.}$$

These models are *linear in the parameters,* β, but *nonlinear in the inputs,* X.

Because the models are linear in the parameters, we can use the same methods for estimating parameters as before. For each case, we replace our original input vector, $X = (X_1, \ldots, X_p)$ by a new vector, $(h_1(X), \ldots, h_M(X))$.

Why They're Called "Basis Functions"

As $\beta_1, \ldots, \beta_M$ vary over R^M, the function

$$f_\beta(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

also varies. These functions form a vector space under the operations of multiplying the function by a scalar or adding functions pointwise.

To see this, note that if a is a scalar,

$$a f_\beta(X) \ = \ \sum_{m=1}^{M} a\beta_m h_m(X) \ = \ f_{a\beta}(X)$$

and if β and β' are in R^M,

$$f_\beta(X) + f_{\beta'}(X) \ = \ \sum_{m=1}^{M} (\beta_m + \beta'_m) h_m(X) \ = \ f_{\beta+\beta'}$$

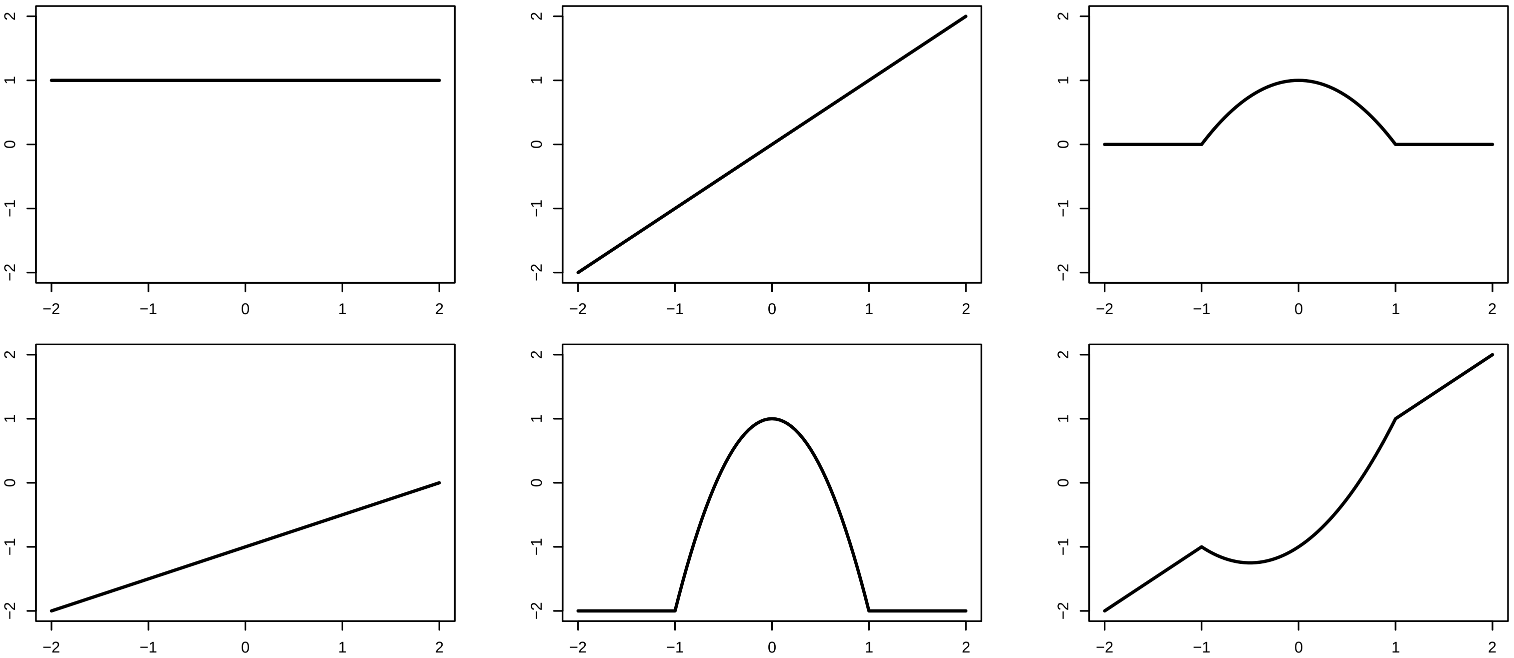
The functions $h_m(X)$ for $m = 1, \ldots, M$ for a *basis* for this vector space. But note that other bases also exist.

A Simple Example

Suppose that X is scalar (ie, $p = 1$) and we use $M = 3$ basis functions, as follows:

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_3(X) = \max(0,\, 1 - X^2)$$

Here are these basis functions and some linear combinations of them:



$\beta_1 = -1,\ \beta_2 = 1/2,\ \beta_3 = 0$ $\beta_1 = -2,\ \beta_2 = 0,\ \beta_3 = 3$ $\beta_1 = 0,\ \beta_2 = 1,\ \beta_3 = -1$

We can add one basis function to another to get an alternative basis:

$$h_1(X) = 1, \quad h_2(X) = X, \quad h_3(X) = \max(1,\, 2 - X^2)$$

Some Restricted Families of Basis Functions

A basis function can be anything. This may be too much to think about, so we might look more narrowly.

In *additive models*, each basis function depends on only one of the inputs. Equivalently, we can write

$$f(X) \;\; = \;\; \sum_{j=1}^{p} f_j(X_j)$$

and then model each univariate function $f_j(X_j)$ using basis functions of X_j only.

In *piecewise polynomial models*, each basis function is a polynomial, except that it is zero outside some region. Using piecewise polynomial functions rather than a single polynomial helps avoid extreme sensitivity to the data.

We'll first look at piecewise polynomial models when X is univariate. When we impose some smoothness constraints, these are known as *spline* models.

Univariate Splines

In one dimension, we can divide the input space into regions by specifying the locations of K *knots*, ξ_j. We use these knots when specifying a set of basis functions for a piecewise polynomial of order M (having degree $M-1$).

If $K = 2$ and $M = 3$, we could use the following basis functions:

$$h_1(X) = I(X < \xi_1), \quad h_2(X) = I(\xi_1 \leq X < \xi_2), \quad h_3(X) = I(\xi_2 \leq X)$$

$$h_4(X) = h_1(X)\,X, \quad h_5(X) = h_2(X)\,X, \quad h_6(X) = h_3(X)\,X$$

$$h_7(X) = h_1(X)\,X^2, \quad h_8(X) = h_2(X)\,X^2, \quad h_9(X) = h_3(X)\,X^2$$

By combining $h_1(X)$, $h_4(X)$, and $h_7(X)$ with coefficients β_1, β_4 and β_7, we can get any second-order polynomial we wish in the first piece (where X is below ξ_1), without affecting the other pieces, and similarly for the other two pieces.

In the text, Figures 5.1 (top two plots) and 5.2 (top left plot) show fits of such models to data for $K = 2$ and $M = 1$, 2, and 4.

Imposing Linear Constraints

We can see the set of piecewise, degree $M-1$ polynomial functions with K knots as a vector space, of dimension $(K+1)M$. This vector space is really the same as $R^{(K+1)M}$, since a vector $\beta \in R^{(K+1)M}$ specifies one of these functions.

As with any vector space, imposing linear constraints leads to a subspace of lower dimension. If we like these constraints, we can find basis functions that span this subspace, and use them rather than the original basis functions.

For many problems, we think the function should be *continuous* at the knots (it will obviously be continuous elsewhere).

For the $K = 2$, $M = 3$ example, this leads to two constraints:

$$\begin{aligned} \beta_1 h_1(\xi_1) + \beta_4 h_4(\xi_1) + \beta_7 h_7(\xi_1) &= \beta_2 h_2(\xi_1) + \beta_5 h_5(\xi_1) + \beta_8 h_8(\xi_1) \\ \beta_2 h_2(\xi_2) + \beta_5 h_5(\xi_2) + \beta_8 h_8(\xi_2) &= \beta_3 h_3(\xi_2) + \beta_6 h_6(\xi_2) + \beta_9 h_9(\xi_2) \end{aligned}$$

These two constraints reduce the dimensionality of the space from 9 to 7

Figure 5.1 (bottom left) and Figure 5.2 (top right) show the effect of requiring continuity for $M = 2$ and $M = 4$.