Basis Expansions

With $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ the function

$$f(x) = E(Y|X = x)$$

is typically globally a non-linear function. We discuss situations where p is small or moderate, but where the function is complicated.

A *basis function expansion* of f is an expansion

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

with $h_m : \mathbb{R}^p \to \mathbb{R}$ for $m = 1, \ldots, M$.

The basis functions are chosen and fixed and the parameters β_m for $m = 1, \ldots, M$ are estimated. This is a *linear model* in the *derived variables* $h_1(X), \ldots, h_M(X)$.

Polynomial Bases

Classical basis functions consists of monomials

$$h_m(x) = x_1^{r_1} x_2^{r_2} \ldots x_p^{r_p}$$

with $r_i \in \{0, \ldots, d\}$ and $r_1 + \ldots + r_p \leq d$. This basis spans the polynomials of degree $\leq d$.

- If the linear models provide first order Taylor approximations of the function, expansions in the degree d polynomials provide order d Taylor approximations.
- However, if $p \geq 2$ the number of basis functions grows exponentially in d.

Indicators

A completely different, non-differentiable idea is to approximate f locally as a constant. Box-type basis functions are

$$h_m(x) = 1(l_1 \leq x_1 \leq r_1) \ldots 1(l_p \leq x_p \leq r_p)$$

with $l_i \leq r_i$ and $l_i, r_i \in [-\infty, \infty]$ for $i = 1, \ldots, p$.

If the boxes are disjoint, the columns in the **X**-matrix for the derived variables are orthogonal:

$$\mathbf{X}_{im} = h_m(x_i) \in \{0, 1\}$$

We can think of this as *dummy variables representing the box.* Consequently, with least squares estimation

$$\hat{\beta}_m = \frac{1}{N_m} \sum_{i:h_m(x_i)=1} y_i, \qquad N_m = \sum_{i=1}^{N} 1(h_m(x_i) = 1).$$

Basis Strategies

The size of the typical set of basis functions increase rapidly with p. What are feasible strategies for basis selection?

- *Restriction*: Choose a priori only special basis functions
  - Additivity; $h_{mj} : \mathbb{R} \to \mathbb{R}$
$$h_m(x) = \sum_{j=1}^{p} h_{mj}(x_j)$$
  - Radial basis functions:
$$h_m(x) = D\left(\frac{||x - \xi_j||}{\lambda_m}\right)$$
- *Selection*: As variable selection – implement exhaustive or step-wise inclusions/exclusions of basis functions.
- *Retriction*: As ridge regression – keep the entire set of basis functions but penalize the size of the parameter vector.

Figure 5.1

Figure 5.2

Splines – $p = 1$

Define $h_1(x) = 1$, $h_2(x) = x$ and

$$h_{m+2}(x) = (x - \xi_i)_+ \qquad t_+ = \max\{0, t\}$$

for $\xi_1, \ldots, \xi_K$ the *knots*.

$$f(x) = \sum_{m=1}^{M+2} \beta_m h_m(x)$$

is a *piecewise linear, continuous* function. One *order-R spline basis* with knots $\xi_1, \ldots, \xi_K$ is

$$h_1(x) = 1, \ldots, h_R(x) = x^{R-1}, \;\; h_{R+l}(x) = (x - \xi_l)_+^{R-1}, \quad l = 1, \ldots, K.$$

Figure 5.3

Natural Cubic Splines

Splines of order R are polynomials of degree $R-1$ beyond the boundary knots ξ_1 and ξ_K. The *natural cubic splines* are the splines of order 4 that are linear beyond the two boundary knots. With

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3$$

the restriction is that $\beta_2 = \beta_3 = 0$ and

$$\sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \theta_k \xi_k = 0.$$

$$N_1(x) = 1, \quad N_2(x) = x$$

and

$$N_{2+l}(x) = \frac{(x-\xi_l)_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_l} - \frac{(x-\xi_{K-1})_+^3 - (x-\xi_K)_+^3}{\xi_K - \xi_{K-1}}$$

for $l = 1, \ldots, K-2$ form a basis.

Obviously $\beta_2 = \beta_3 = 0$ and then beyond the last knot the second derivative of f is

$$f''(x) = \sum_{k=1}^{K} 6\theta_k (x - \xi_k) = 6x \sum_{k=1}^{K} \theta_k - 6 \sum_{k=1}^{K} \theta_k \xi_k,$$

which is zero for all x if and only if the conditions above are fulfilled. For N_{2+l} we see that

$$\theta_l = \frac{1}{\xi_K - \xi_l}, \ \theta_{K-1} = -\frac{1}{\xi_K - \xi_{K-1}}, \ \theta_K = \frac{1}{\xi_K - \xi_{K-1}} - \frac{1}{\xi_K - \xi_l}$$

and the condition is easily verified. By evaluating the functions in the knots, say, it is on the other hand easy to see that the K different functions are linearly independent. Therefore they must span the space of natural cubic splines of co-dimension 4 in the set of cubic splines.

B Splines

Yet another basis for the splines ...

Defined by a recursion in R;

$$B_{k,1}(x) = \begin{cases} 1 & \text{if } \tau_k \leq x \leq \tau_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

with

$$\tau_1 \leq \ldots \tau_R = \xi_0 < \tau_{R+1} = \xi_1 < \ldots < \tau_{R+K} = \xi_K < \tau_{R+K+1} = \xi_{K+1} \leq \ldots \leq \tau_{2R+K}$$

and

$$B_{k,r} = \frac{x - \tau_i}{\tau_{i+r+1} - \tau_i} B_{k,r-1}(x) + \frac{\tau_{i+r} - x}{\tau_{i+r} - \tau_i} B_{k+1,r-1}(x)$$

for $k = 1, \ldots, K + 2R - r$.

Figure 5.20 – B-splines

Knot Placing Strategies

How do you determine the knots?

- Fix the number (the complexity parameter), spread them uniformly over the whole range of data.
- Fix the number, spread them according to the emprical distribution.
- Adaptive selection of the number and/or the location – ranging from ad hoc adaptation to a full fledged, complete estimation from data.
- Smoothing algorithms determine automatically their location

Smoothing Splines

Allowing $E(Y|X = x) = f(x)$ to be an arbitrary, but twice differentiable functions, define the *penalized residual sum of squares*

$$\mathrm{RSS}(f, \lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_a^b f''(t)^2 \mathrm{d}t$$

If f^λ is a minimizer of $\mathrm{RSS}(f, \lambda)$, the *natural cubic splines* with knots in $x_1, \ldots, x_N$ have the properties that

- they can interpolate; there is a natural cubic spline f with $f(x_i) = f_\lambda(x_i)$
- and among all interpolants f attains the least value of

$$\int_a^b f''(t)^2 \mathrm{d}t.$$

The solution $f^\lambda = \sum_{i=1}^{N} \theta_i N_i(x)$ is a natural cubic spline.

Only requirement above on $a < b$ is that $[a, b]$ contains all the data points. For the interpolation argument we also need that the x_i's are different. See Exercise 5.7 for the second bullet point above.

Smoothing Splines

In vector notation

$$\mathbf{f} = \mathbf{N}\theta$$

with $\mathbf{N}_{ij} = N_j(x_i)$ and

$$\begin{aligned} \mathrm{RSS}(f, \lambda) &= (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda \int_a^b f''(t)^2 \mathrm{d}t \\ &= (\mathbf{y} - \mathbf{N}\theta)^T(\mathbf{y} - \mathbf{N}\theta) + \lambda\theta^T \mathbf{\Omega}_N \theta \end{aligned}$$

with

$$\boldsymbol{\Omega}_{N,ij} = \int_a^b N_i''(t)N_j''(t)\mathrm{d}t.$$

This *generalized ridge regression problem* has solution

$$\hat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\boldsymbol{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y}$$

and the fitted values are

$$\hat{\mathbf{f}} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\boldsymbol{\Omega}_N)^{-1}\mathbf{N}^T\mathbf{y}$$

Degrees Of Freedom

Writing

$$\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\boldsymbol{\Omega}_N)^{-1}\mathbf{N}^T$$

and by analogy with projection matrices the *effective degrees of freedom* is

$$\mathrm{df}_\lambda = \mathrm{trace}(\mathbf{S}_\lambda).$$

The value of df_λ is monotonely decreasing from N to 0 as λ increases from 0 to ∞.

The matrix $\mathbf{S}_\lambda$ is known as a *spline smoother* and it is common to specify the degrees of freedom instead of λ in practice.

Figure 5.8 – Smoother Matrix

Multidimensional Splines

Two multivariate versions.

- *Tensor products.* Consider a basis consisting of

$$B_{i_1,R}(x_1)B_{i_2,R}(x_2)\dots B_{i_p,R}(x_p)$$

– compare with the multinomial basis for polynomials.

- *Thin plate splines.* If $p = 2$ consider minimizing

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int_A (\partial_1^2 f)^2 + 2(\partial_1\partial_2 f)^2 + (\partial_2^2 f)^2.$$

The solution is a function

$$f(x) = \beta_0 + x^T\beta + \sum_{i=1}^{N} \alpha_i \eta(||x - x_i||)$$

with $\eta(z) = z^2 \log(z^2)$ – thus a *radial basis function expansion.*

Figure 5.10 – Tensor Products of B-splines

Kernel Density Estimation

If $Y \in \{1, \dots, K\}$ and g_k denotes the density for the conditional distribution of X given $Y = k$ the Bayes classifier is

$$f(x) = \operatorname*{argmax}_k \pi_k g_k(x)$$

If $\hat{g}_k$ for $k = 1, \dots, K$ are density estimators – non-parametric kernel density estimators, say – then using *the plug-in principle*

$$\hat{f}(x) = \operatorname*{argmax}_k \hat{\pi}_k \hat{g}_k(x)$$

is an estimator of the Bayes classifier.

This is the non-parametric version of LDA.

Naive Bayes

High-dimensional kernel density estimation suffers from the *curse of dimensionality.*

Assume that the X-coordinates are independent given the Y, then

$$g_k(x) = \prod_{i=1}^{p} g_{k,i}(x_i)$$

with $g_{k,i}$ univariate densities.

$$\begin{aligned}\log\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)} &= \log\frac{\pi_k}{\pi_K}+\log\frac{g_k(x)}{g_K(x)}\\ &= \log\frac{\pi_k}{\pi_K}+\sum_{i=1}^{p}\underbrace{\log\frac{g_{k,i}(x_i)}{g_{K,i}(x_i)}}_{h_{k,i}(x)}\\ &= \log\frac{\pi_k}{\pi_K}+\sum_{i=1}^{p}h_{k,i}(x)\end{aligned}$$

Naive Bayes – Continued

The conditional distribution above is an example of a *generalized additive model.* Estimation of $h_{k,i}$ using univariate (non-parametric) density estimators $\hat{g}_{k,i}$;

$$\hat{h}_{k,i}=\log\frac{\hat{g}_{k,i}(x_i)}{\hat{g}_{K,i}(x_i)}$$

is known as *naive* – or even *idiot's* – *Bayes.*

Naive Bayes – Discrete Version

If some or all of the X variables are discrete, univariate kernel density estimation can be replaced by appropriate estimation of point probabilities.

If all X_i take values in $\{a_1,\ldots,a_n\}$ the extreme implementation of naive Bayes is to estimate

$$\hat{g}_{k,i}(r)=\frac{1}{N_k}\sum_{j:y_j=k}1(x_{ji}=a_r),\quad N_k=\sum_{j=1}^{N}1(y_j=k).$$

This is a possible solution procedure for the first asignment.