7.4.3: The Spline Basis Representation

A cubic spline with K knots can be modeled as
$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_{K+3} b_{K+3} + \varepsilon_i$
for function b_j $j \in [K+3]$. Then fit with least squares.

<u>Direct representation:</u> start off with $1, x, x^2, x^3$, and then add one <u>truncated power basis</u> function per knot ξ, where such a function is

$$h(x, \xi) = (x-\xi)_+^3 = (x-\xi)^3 \mathbb{1}_{(\xi, \infty]}(x)$$

Least squares on $X, X^2, X^3, h(X, \xi_1), \dots, h(X, \xi_K)$

<u>Natural splines</u> are required to be linear in the boundary region.

7.4.4: Choosing the Number and Location of the Knots

2/24/23
Day 17

- Could place the knots more densely where the function seems to change rapidly.
- Usually, knots are placed uniformly
  - Specify the degrees of freedom, then have software automatically place the corresponding number of knots at uniform quantiles of the data.

Use cross-validation to choose the number of knots. The number of knots K giving the smallest CV RSS is chosen.

7.5: Smoothing Splines

7.5.1: Overview

We want to find a function $g(x)$ that is smooth and gives small $RSS = \sum_{i=1}^{n} (y_i - g(x_i))^2$, but we don't want to overfit. We can try minimizing the functional

$$\underbrace{\sum_{i=1}^{n} (y_i - g(x_i))^2}_{\text{loss}} + \underbrace{\lambda \int_X g''(t)^2 \, dt}_{\text{penalty}}, \quad \lambda \geq 0 \text{ is the tuning parameter}$$

The minimizer g is called a smoothing spline.

If g is very smooth, then g' is close to constant and g'' is very small.

The larger λ, the smoother g.

g will be a natural cubic spline with knots at each x_i (WHOA, COOL).

7.5.2: Choosing λ

λ controls the very high flexibility of having knots at each data point.

As $\lambda \nearrow \infty$, the <u>effective degrees of freedom</u>, df_λ, decreases from n to 2.

Defining df_λ is technical. Consider

$$\hat{g}_\lambda = S_\lambda y,$$

where $\hat{g}$ is a length n vector of the smoothing spline g for a given λ evaluated at the n data points.

There is some matrix S_λ (which has a formula elsewhere), and y is the response.

Anyhow, the effective degrees of freedom is

$$df_\lambda = \text{trace}(S_\lambda).$$

We need to choose λ. Cross validation works. LOOCV turns out to be very efficient for smoothing splines, with

$$\boxed{RSS_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{S_\lambda\}_{ii}} \right]^2}$$

where $\hat{g}_\lambda^{(-i)}$ is fit on the data excluding datapoint i.

Similarly, LOOCV can be calculated for regression splines using eqn 5.2.

7.6: Local Regression

Fit a flexible non-linear function by computing a fit at some x_0 using only local data.

- Suffers bad from the curse of dimensionality.

7.7: Generalized Additive Models

Extension of multiple linear regression

Allow nonlinear functions of the variables while maintaining additivity.

7.7.1 GAMs for regression problems

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$$

We calculate a separate f_j for each X_j, then add all of the contributions

Sections 7.1-7.6 fit functions to one variable. GAMs extend these methods to multiple variables.

ex) Fit

$$wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon$$

- Fit year and age with natural splines.
- Fit education using 5 constants, one for each variable using the dummy variable approach.

Note that this whole model is a regression onto spline basis variables and dummy variables.

Pros & Cons of GAMs

Δ Automatically model ~~non~~ linear relationships that linreg will miss.

Δ Can be more accurate

Δ Can still view effect of one variable while holding the others fixed.

Δ The smoothness of f_j can be summarized via degrees of freedom.

∇ Restriction to be additive

7.7.2: GAMs for Classification Problems

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

Ch. 8: Tree-Based Methods

Stratify or segment the predictor space into a number of simple regions. For a prediction, report the mean or mode of the region it belongs to. The set of splitting rules can be summarized by a decision tree, hence the name.

- Not the best predictivity, so we also introduce:
- Bagging
- Random forest
- Boosting

8.1: The Basics of Decision Trees