Homework 3

Stat 435, Spring 2020

Due Friday, May 8, 11:59pm

The homework comes with a test data set "test-data.R" in "dump" format. For the test data, $n = 100$, and $x_1, \ldots, x_n$ are equi-spaced in $[0, 2\pi]$. The true conditional expectation is $f(x) = \sin(x)$, and the error sd is $\sigma = 0.4$.

1. Experiments with Turbo

Turbo is an expansion based smoother that fits 2nd order (linear) splines. It is described in the article *Flexible Parsimonious Smoothing and Additive Modeling* by J.H. Friedman and B.W. Silverman (Technometrics, Vol. 31, No. 1, 1989, pp 3 – 39.

a) (10 points) Define basis functions $B_i(x) = (x - x_i)_+, i = 1, \ldots, n-1$, and $B_n(x) = 1$. Write a function `truncated.power.design.matrix(x)` that generates the $n \times n$ design matrix for this set of basis functions.

b) (10 points) Install the package "leaps" and take a look at the documentation. Write a function

```
regsubsets.fitted.values <- function(X, regsubsets.out, nterm)
```

that computes the fitted values for a model with nterm terms.

c) (10 points) For the test data produce a plot of residual sum of squares as a function of the number k of basis functions in the model.

d) (10 points) Plot the GCV score as a function of k. Surprised? Why? Explanation?

e) (10 points) F&S (pp 9–10) propose to fix this problem by charging 3 degrees of freedom for each of $B_1, \ldots, B_{n-1}$ entered into the model. Plot this modified GCV score as a function of the number of basis functions in the model. Surprised? Problems with the F&S definition of GCV? (If you have trouble figuring out where the constant term is included in the model, you may charge 3 degrees of freedom for each of $B_1, \ldots B_n$.)

f) (10 points) Restricting yourself to suitable small values of k, find the "forward" and "backward" models with the smallest (modified) GCV scores and plot them.

2. Experiments with order 2 smoothing splines

Training data and basis functions as in (1) above. Define the $n \times n$ matrix X by $X_{ij} = B_j(x_i)$.

An order 2 smoothing spline is a function of the form

$$\begin{aligned} g(x) &= \sum \hat{a}_j B_j(x)\,, \text{where} \\ \hat{\mathbf{a}} &= \text{argmin}_{\mathbf{a}} \left[\|\mathbf{y} - X\mathbf{a}\|^2 + \lambda\, \mathbf{a}^T \Omega \mathbf{a} \right] \text{ with} \\ \Omega &= \text{diag}(0, 1, \ldots, 1, 0)\,. \end{aligned}$$

The vector of predicted values for the training sample is $\hat{\mathbf{y}} = X\hat{\mathbf{a}}$.

(a) (10 points) Show that

$$\begin{aligned} \hat{\mathbf{y}} &= X(X^T X + \lambda\Omega)^{-1} X^T \mathbf{y} \\ &= S_\lambda\, \mathbf{y}\,. \end{aligned}$$

(b) (10 points) Read the data in the file `test-data.R`. Use the glmnet package to plot data and spline for $\lambda = 0, 1, 10, 10^6$. Verify (graphically) that the spline for $\lambda = 10^6$ is very close to the least squares line.

(c) (20 points) Use the glmnet package to find the optimal value of λ by cross-vaildation. Print out λ_{opt} and plot the corresponding spline.

3. ISLR Section 6.8 Problem 1 (25 points)

4. ISLR Section 6.8 Problem 4 (25 points)