Stat 6242
Assignment 4, Due Nov. 16

November 7, 2016

Problem 1. (10 pts) Express all the smoothers we discussed (kernel, k-NN, local polynomial, smoothing splines) in the form:

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i$$

or in matrix format as

$$\hat{\mathbf{m}} = \mathbf{S}_x \mathbf{Y}$$

2. (20 pts) For the **Wage** data in library ISLR (library(ISLR); attach(Wage)), fit the response *wage* on *age* using all the different smoothing methods we have discussed (kernel, k-NN, local polynomial fitting, smoothing splines)

   a. (12 pts) Use 5-fold cross-validation to select the optimal smoothing parameter for each smoother.

   b. (4 pts) Plot the corresponding fit for each nonparametric smoother.

   c. (4 pts) Plot the residuals vs the fitted responses for each nonparametric smoother.

   5-fold CV algorithm:

   1. Divide the data into 5 equal parts.
   2. For each $k = 1, \ldots, 5$, fit the model to estimate the smooth $\hat{m}_{-k}$ and compute its mean squared error for predicting the kth part:

$$MSE_k(\hat{m}_{-k}) = \frac{\sum_{\text{i in kth data set}} (y_i - \hat{m}_{-k}(x_i))^2}{\#\text{ of points in the kth data set}}$$

   The notation $\hat{m}_{-k}$ means that the smoother was based on 4/5 of the data excluding the kth part.

3. The overall 5-fold cross-validation error is

$$MSE = \frac{1}{5}\sum_{k=1}^{5} MSE_k(\hat{m}_{-k})$$

4. For each smoothing method, select the bandwidth with the smallest 5-fold CV MSE.

3. (15 pts) How fast do different smoothers converge? In the following simulation study we will examine the effect of different smoother choice to rate of convergence. Draw n observations such that $x_i \sim Unif(0,1)$ and let

$$y_i = \sin(2\pi(1-x_i^2)) + x_i\epsilon_i$$

where $\{\epsilon_i\}$ are iid standard normal variates.

Use the following measure of distance for sample sizes ranging from 100 to 1500 in increments of 100 to examine which of the following smoothers converge faster to the true underlying curve:

$$d(n) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_h(x_i) - m(x_i)\right)^2$$

Compare the Epanechnikov kernel smoother, local linear smoother and smoothing spline. For all three, select the optimal smoothing parameter by minimizing the leave-one-out CV error:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{-i}(x_i))^2$$

where $\hat{m}_{-i}(x_i) = \sum_{j\neq i} w_j(x_i)y_j$.