BST 764: Applied Statistical Modeling
Breheny

Assignment 3: Nonparametric regression
Due: Tuesday, November 8

Mathematical concepts and derivations

1. Write the set of truncated spline basis functions for representing a cubic spline function with three knots.

2. Show that $\boldsymbol{\beta}^T \boldsymbol{\Omega} \boldsymbol{\beta} = \int f''(u)^2 du$, where all terms are defined on slides 9 and 10 of the 10-4 notes.

3. What is the difference between the Nadaraya-Watson kernel-based estimator and k-nearest neighbors (described in our first lecture)? Briefly compare the two approaches: what are their relative strengths and weaknesses?

Simulation

4. Suppose $x_i \overset{\text{iid}}{\sim} Unif(-3, 3)$ for $i = 1, 2, \ldots, 100$ and that $y_i = f(x_i) + \epsilon$, where ϵ follows a standard normal distribution and $f(x) = -x^2$. Conduct a simulation study comparing three methods: polynomial regression (with linear and quadratic terms), smoothing splines, and local linear regression. NOTE: Some methods are unable to predict outside the observed range. To avoid this, set two values of $\{x_i\}$ equal to -3 and 3 and let the other 98 be uniformly distributed.

   (a) On average, how many degrees of freedom do splines and local linear regression need to represent f?

   (b) At equally spaced points throughout the range of x, evaluate the bias, variance, and MSE of the three methods. Plot your results versus x and comment on what you see.

5. Repeat the above exercise, only change f to be the following piecewise function:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } x > 0 \end{cases}.$$

   One interpretation of such an f is that it represents a risk factor for which low levels have no impact, but past a certain threshold, there is an increasingly severe risk. In addition to the three methods in the above exercise, add a fourth method: polynomial regression with terms up to x^5.

   (a) How many degrees of freedom do splines and local linear regression need to represent f?

   (b) At equally spaced points throughout the range of x, evaluate the bias, variance, and MSE of the three methods. Plot your results versus x and comment on what you see.

Application

NOTE: Use spline-based methods for one of the problems below, and use kernel-based methods for the other. The choice of which to use for which problem is up to you.

6. Earlier in the semester, we looked at a 1989 prostate cancer study (`prostate.txt`); the study is described in the 9-1 notes. Re-analyze the data using generalized additive models. Write a ≈ 2 page report (not including figures) on what you find. In particular, comment on which of the variables are the most important, which variables display nonlinear effects, and interpret the associations (*i.e.*, interpret the slopes of the linear effects, the differences between levels of categorical variables, the shapes of the nonlinear effects), as well as any interactions you observe, if any.

7. The course web page contains a data set (`commute.txt`) gathered by me that records the time and date that my evening commute began and ended, over a period lasting from June to November of last year. Each trip constitutes an observation. Many dates are missing (weekends, and days in which I was out of town, didn't go directly home, or simply forgot to record the trip). Analyze the data and write a ≈ 2 page report (not including figures), pointing out any interesting trends you observe. Interesting questions include, but are not limited to: Is there a window of "rush hour" time that I should try to avoid? If so, what is the magnitude of the rush hour effect? If I leave at time x, what time will I get home? Does travel time differ by day of the week? Is there a day by departure time interaction? Are travel times shorter when school is not in session? By how much? Is day of the week a confounder? Why might it be a confounder? Do I tend to leave earlier on some days than others? What are the limitations of the data?