Stat 462/862 Assignment 2

(Due on Oct 27th, 2015)

1. Generate 100 data based on the following model,

$$y_i = f(x_i) + \epsilon,$$

where $f(x_i) = -2x_i^2$, $x_i \overset{iid}{\sim} Unif(-a, a)$ for $i = 2, \ldots, 99$, $x_1 = -a$ and $x_{100} = a$, and $\epsilon \sim N(0, \sigma^2)$. Now specify the value of a and σ^2 at your own choice. Do the following.

   (a) Fit the data using four methods, second-order polynomial model (with linear and quadratic terms), cubic spline, natural cubic spline, and smoothing spline.

   (b) Compare these four methods by evaluating the bias, variance and mean square error (MSE) at equally spaced 300 points throughout the range of x. Plot your results versus x and make comments.

2. Analyze the "motor cycle data" (use "library(MASS)", then load "data(mcycle)", the data are x=times, y=accel). Use smoothing splines to fit the data. Try different df's in $[5, 20]$. Find the optimal df in $[5, 20]$ according the cross-validation criterion (in the function "smooth.spline", specify "cv=T"). What is the λ and cross-validation error of the best fit? Return the following three plots:

   (a) The observation points and the optimal smoothing spline fit.

   (b) The observation points and the three smoothing splines with df=5, 10, 15 (three different colored curves). Then you should also add a "legend" to denote these lines.

   (c) Plot the cross validation errors against different df's from 5 to 20 (show both points and lines). The step of df's is 0.5. (Hint: from this plot you can find the optimal df.)

3. Use logistic regression to analyze the data "admit.txt".
Data background: a researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (rank), affect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

Show the estimation of the coefficients as in Table 4.2 on page 122 of HTF, and write the log-ratio as in eq. (4.17) in the textbook.

4. (For graduate students only) Recall that the smoothing spline estimator is given by $\hat{f}(x) = \sum_{j=1}^{N} G_j(x)\hat{\theta}_j$, where $\hat{\boldsymbol{\theta}} = (\mathbf{G}^T\mathbf{G} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{G}^T\mathbf{y}$. Now find $\mathbf{S}_\lambda$ such that $\hat{f}(x) = \mathbf{S}_\lambda\mathbf{y}$ and show that $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$ for a certain matrix $\mathbf{K}$. (Hint: use singular value decomposition)