Statistics 315a
Homework 1, due Wednesday January 30, 2013.

"ESL" refers to the course textbook, and ESL 2.4 refers to exercise 2.4 in ESL. Since the homework assignments count 70% of your final grade, you must do them on your own. Problem 1 is computing intensive, and is partly there to get you up to speed in R. You can form teams of up to 3 students to collaborate on problem 1, but must still write up your results on your own. If so, clearly indicate in your writeup who is on the team.

1. *Error curves*

   (a) Write a function to simulate data as described on pages 16-17 in ESL for one of the classes. Your function should take as inputs a 10×2 matrix of centroids, the sample size, and the noise variance. Generate a training sample of size 100 for each class, as well as a test sample of 5,000 per class. (Best to generate the centroids matrices per class once and store them). Try and write elegant code, that makes use of the matrix/vector facilities in `R`.

   (b) Evaluate the misclassification performance of K-nearest neighbor classification on the training and test set (`library(class)` in `R`), for $k = \{1, 3, 5, 9, 15, 25, 45, 83, 151\}$. Evaluate also the performance of the linear regression procedure. Produce a plot as in Figure 2.4.

   (c) Using the training data, use 10-fold cross-validation to estimate the errors in the cases above. Include these errors in your plot (average fold errors and estimated standard error of this average).

   (d) Summarize what you see.

2. ESL 2.4

3. ESL 2.7

4. Given data on two variables X and Y, consider fitting a quartic polynomial regression model $f(X) = \sum_{j=0}^{4} \beta_j X^j$. In addition to plotting the fitted curve, you would like a 90% confidence band about the curve. Consider the following two approaches: (1) At each point x_0, form a 90% confidence interval for the linear function $a^T\beta = \sum_{j=0}^{4} \beta_j x_0^j$, or (2) Form a 90% confidence set for the vector β as in eqn (3.15) of the text, which in turn generates confidence intervals for $f(x_0)$.

(a) How do these approaches differ? Which band is likely to be wider? Conduct a small simulation experiment to compare the two methods.

(b) Do either of the confidence bands constructed in (a) have simultaneous 90% coverage for all x_0? Explain.

5. Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_1^N (y_i - x_i\beta)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_1^M (\tilde{y}_i - \tilde{x}_i\beta)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})],$$

where the expectations are over all that is random in each expression.

6. (a) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau\mathbf{I})$, and Gaussian sampling model $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ and σ^2.

(b) For the lasso, draw a similar relation between the objective function and a Bayesian log-posterior with a Laplacian (double exponential) prior.

(c) Plot the density functions of the Gaussian and Laplacian distributions on the same figure (each having mean 0 variance 1).

(d) For a specific dataset, if one were to sample from the posterior distribution corresponding to the lasso objective, would the realizations look sparse? Explain.

(e) If you answered "no" to (d), suggest a different Bayesian specification that would produce sparse realizations.