Statistics 315a
Homework 1, due Monday Jan 25, 2010.

1. Compare the classification performance of linear regression and k–nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's, and $k = 1, 3, 5, 7$ and 15. The zipcode data are available at

   `www-stat.stanford.edu/ElemStatLearn`

   (a) Apply both methods and plot both the training and test error for each approach. You will need to come up with some method for choosing k from the training set.

   (b) Plot examples of test digits that are misclassified by one method but not the other, and by both.

   (c) Devise a way of combining the two approaches (regression and K-NN) to come up with a potentially better classifier, and try it out on these data.

2. Exercise 3.2 in ESL

3. 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 $R_{tr}(\beta) = \frac{1}{N}\sum_1^N (y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_1^M (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, 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.

4. The edge effect problem discussed in Chapter 2 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution $X \sim N(0, \mathbf{I}_p)$. The squared distance from any sample point to the origin has a χ_p^2 distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/||x_0||$ be an associated unit vector. Let $z_i = a^T x_i$ be the projection of each of the training points on this direction.

Show that the z_i are distributed $N(0,1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.

Hence for $p = 10$, a randomly drawn test point is about 3.1 standard deviations from the origin, while all the training points are on average one standard deviation along direction a. So most prediction points see themselves as lying on the edge of the training set.

5. Suppose $p \gg N$, you have a data matrix $\mathbf{X}$ and a quantitative response vector $\mathbf{y}$, and you plan to fit a linear regression model.

   (a) Explain why the ordinary least squares solution is not unique. What can you say about the residuals of any of the solutions.

   (b) Is the ridge regression solution unique? why?

   (c) Suppose you compute a series of ridge solutions, letting λ get successively smaller. What can you say about the limiting ridge solution in this case, as $\lambda \downarrow 0$.

   (d) Using the SVD of $\mathbf{X}$, write a closed form expression for this limiting solution.

6. **All-subsets regression.** Consider a linear regression problem with p variables. All-subsets regression creates a sequence of p models, by establishing the best model of size k, $k = 1, \ldots, p$ (using squared error loss on the training data). The parameter k is a tuning parameter that has to be selected. Conduct a small simulation study as follows. Let

$$Y = X^T\beta + \varepsilon,$$

   where $p = 20$, and β is generated from a standard Gaussian distribution (once and for all). Index the variables so that β is in descending absolute value. Assume the inputs and errors are Gaussian, and choose ε so that the the signal to noise variance ratio is approximately 1. Using a fixed test set of size 10000 from the model, your simulation should generate 30 training datasets each of size 50, and produce squared prediction error, bias^2 and variance curves as a function of k. (Hint: In R there is a `leaps` contributed package in the CRAN archive that performs all subsets regression) Produce the corresponding bias, variance and prediction error curves for the simpler sequence of models that uses the first k variables, and include them in your plot. Summarize what you have learned from this exercise.