STA 414S/2104S: Homework #1 Due Feb.11, 2010 at 1 pm

Late homework is penalized at 20% deduction per day. You are welcome to discuss your work on this homework with your classmates. You are required to write up the work on your own, using your own words, and to provide your own computer code.

Answers to the computational questions must be submitted in two parts. The first part presents your conclusions and supporting evidence in a report, written in paragraphs and sentences (not point form) **that does not include computer code**. This part may include tables and figures. The second part is a complete, and annotated, file showing the computer code that you used to obtain the results discussed in the first part. It is important to include readable code, since everyone's answers will be based on different training and test samples.

1. *Likelihood and Bayesian inference in the linear model:*
   Suppose that the $n \times 1$ vector Y follows a normal distribution with mean $X\beta$ and variance $\sigma^2 I$:
   $$Y \sim N(X\beta, \sigma^2 I)$$
   i.e. that
   $$f(y \mid \beta, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\}.$$

   (a) The maximum likelihood estimates $(\hat\beta, \hat\sigma^2)$ are defined to be the values of β and σ^2 that simultaneously maximize the likelihood function, or more conveniently the log-likelihood function
   $$\ell(\beta, \sigma^2) = \log f(y \mid \beta, \sigma^2).$$
   Give expressions for the maximum likelihood estimator of β and σ^2. You may assume that X has full column rank.

   (b) The likelihood ratio statistic for testing the hypothesis $\beta = \beta_{(0)} = (\beta_0, 0, \ldots, 0)$ is defined as
   $$W = 2\{\ell(\hat\beta, \hat\sigma^2) - \ell(\hat\beta_{(0)}, \hat\sigma_0^2)\},$$
   where $(\hat\beta_{(0)}, \hat\sigma_0^2)$ is the maximum likelihood estimate of (β, σ^2) when $\beta = \beta_{(0)}$. Show that $\hat\sigma_0^2 = n^{-1}\Sigma(y_i - \bar y)^2$, and that W is a function of the F-test for regression.

   (c) Assume σ^2 is known. Suppose that we assume a prior distribution for β that is $N(0, \tau^2 I)$, where τ^2 is also known:
   $$f(\beta) = \frac{1}{(\sqrt{2\pi}\tau)^p} \exp(-\frac{1}{2\tau^2}\beta^T\beta).$$
   By Bayes theorem the posterior distribution of β, given y, is
   $$f(\beta \mid y) = f(y \mid \beta)f(\beta) / \int f(y \mid \beta)f(\beta)d\beta.$$

Show that this posterior distribution for β is normal, with

$$\begin{aligned} E(\beta \mid y) &= (X^T X + \lambda I)^{-1} X^T y, \\ \mathrm{cov}(\beta \mid y) &= (X^T X + \lambda I)^{-1} \sigma^2 \end{aligned}$$

where $\lambda = \sigma^2/\tau^2$. What is the limiting posterior distribution as $\tau^2 \to \infty$?

(d) **2104 only**: Continuing with the assumption of known σ^2, show that the mean of the posterior distribution for β using the double exponential prior

$$f(\beta) = \frac{1}{2\tau} \exp(-|\beta|/\tau)$$

gives the lasso estimator.

2. *Exercise 3.8 of HTF*:
Consider the QR decomposition of the uncentered $N \times (p+1)$ matrix X, with a first column of 1's, and the singular value decomposition of the $N \times p$ centered matrix $\tilde{X}$. Show that Q_2 and U span the same subspace, where Q_2 is the sub-matrix of Q with the first column removed. Under what circumstances will they be the same, up to sign flips?

3. *The wine quality data*:
A recently posted regression data set at the UCI machine learning repository is the *wine quality data*. For this exercise we will work with the red wine data set. It can be accessed from within R by read.csv("http://www.utstat.utoronto.ca/reid/sta414/winequality-red.csv", sep=";"). There are 1599 cases, and 11 inputs. The output variable is the quality score, a number between 0 and 10. The goal is to use the features to predict the quality score.

   (a) Choose 1000 cases at random to be your personal training data set. The remaining 599 cases are the test data set.

   (b) Estimate the coefficients in a linear model using least squares with all 11 features, all possible subsets regression, ridge regression, lasso regression, PCR and PLS.

   (c) Evaluate each method on the test data by computing the mean of the squared prediction error.

   (d) Present the results in a Table similar to Table 3.3.

   (e) Although the quality score can range from 0 to 10, most of the values are 5 and 6. What is the range of quality scores in your test data? How might this affect the estimated mean square error?

4. **2104 only** *Exercise 3.10 of HTF* :
Show that the ridge regression estimate can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix X with p additional rows $\sqrt{\lambda} I$, and augment y with p zeros. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data points that satisfy them.

5. **2104 only** *Exercise 3.2 of HTF*:
   Given data on two variables X and Y, consider fitting a cubic polynomial regression model $f(X) = \sum_{j=0}^{3} \beta_j X^j$. In addition to plotting the fitted curve, you would like a 95% confidence band about the curve. Consider the following two approaches:

   (a) At each point x_0, form a 95% confidence interval for the linear function $a^T\beta = \sum_{j=0}^{3} \beta_j x_0^j$.

   (b) For a 95% confidence set for β as in (3.15), which in turn generates confidence intervals for $f(x_0)$.

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

6. *Project description*:
   Submit a brief description of the data set you will analyse for your project, along with the source (usually a web site) and any relevant papers associated with the data set. Be sure to identify the response variable or variables, to clearly state the number of cases and the number of features, as well as whether or not there is any missing data. State the problem that you expect to address in analysing this data. The answer to this question should be about one paragraph.