**STA 414S/2104S**: Homework #2 Due Mar. 4, 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. *Classification of the wine quality data*:
   Another approach to analysing the wine quality data from HW 1 is to consider the quality scores as categorical, and use the features to classify wines into various categories. For this exercise, create 4 categories: Bad (0– 4), Acceptable (5), Good (6), Excellent (8–10). Using training and test set as before, compare the classification errors on the test data using linear discriminant analysis, quadratic discriminant analysis, and the naive Bayes classifier of §6.6.3.
   **STA 2104 only**: Exercise 6.12 describes a version of local discriminant analysis, which is build on a kernel function $K_\lambda(\cdot, x_0)$, and uses this to provide a set of weights to linear discriminant analysis. Write a program to implement this on the wine data, and compare the predictions to those above.

2. *A Bayesian model for smoothing*: Assume $y_i = f(x_i) + \epsilon_i, i = 1, \ldots, N$, for $x_i \in R$ and $\epsilon_i \sim (0, \sigma^2)$ and that we fit $f(\cdot)$ using a cubic smoothing spline, i.e. $f(x) = \sum_{j=1}^N N_j(x)\theta_j$, as at (5.10). Show that if the prior distribution for $\theta$ is Gaussian with mean 0 and covariance matrix
   $$\frac{\sigma^2}{\lambda}\mathbf{\Omega}^{-1}.$$
   that the posterior distribution of $\theta$ given $\mathbf{y}$ is again Gaussian, and give expressions for the mean and variance. Use this to get expressions for
   $$E(\mathbf{N}\theta \mid \mathbf{y}), \quad \mathrm{cov}(\mathbf{N}\theta \mid \mathbf{y}).$$

3. *Project progress report*:
   In two or at most three pages, provide a well-written introduction to the data set and the problem you are addressing. Describe the approaches you intend to take, and give a brief description of what has been achieved so far.