STA 450S/4000S: Homework #3
Due April 6, 2005

1. *A Bayesian model for smoothing* Recall from §5.4 that the smoothing spline evaluated at the training data points minimizes

$$RSS(\theta, \lambda) = (\mathbf{y} - \mathbf{N}\theta)^T(\mathbf{y} - \mathbf{N}\theta) + \lambda\theta^T\mathbf{\Omega}\theta$$

where $\{\mathbf{N}\}_{ij} = N_j(x_i)$ are the spline basis functions, and $\{\mathbf{\Omega}\}_{ij} = \int N_i''(t)N_j''(t)dt$, with solution

$$\hat{\theta} = (\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega})^{-1}\mathbf{N}^T\mathbf{y},$$

leading to

$$\hat{\mathbf{f}} = \mathbf{N}\hat{\theta} = (I + \lambda K)^{-1}\mathbf{y}.$$

Now assume that $\mathbf{y}$, given θ, follows a Gaussian distribution with mean $\mathbf{N}\theta$ and covariance matrix $\sigma^2 I$. Assume a prior distribution for θ that is Gaussian with mean 0 and covariance matrix

$$\frac{\sigma^2}{\lambda}\mathbf{\Omega}^{-1}.$$

Show that the posterior distribution of θ given $\mathbf{y}$ is again Gaussian, and give expressions for the mean and variance. Use this to get expressions for

$$E(\mathbf{N}\theta \mid \mathbf{y}), \quad \text{cov}(\mathbf{N}\theta \mid \mathbf{y}).$$

Does the prior above make sense in this context?

2. *Classification on the wine data* The wine data has 178 cases, and there are three classes. It is available on Cquest in `/u/reid` as `wine.data`; the columns names are given in `wine.names`. These files are also on the course home page.

   (a) Select 120 observations to be a training data set, being sure to maintain approximately the same proportion of each of the 3 classes in your training data as is in the full data. (This will require stratified sampling.)

   (b) Use the Naive Bayes classifier described in §6.6.3 to classify the data.

   (c) Try classification based on the neural network program `nnet` available in the `MASS` library. Provide careful discussion of the choices you made to train the neural network.

   (d) Use `rpart` to construct a classification tree.

   (e) Finally compare the results of (a), (b) and (c) with the results of linear discriminant analysis, by tabulating the classification errors made on the test data by the various methods.

3. *Cross-validation for linear smoothers* Assume $y_i = f(x_i) + \epsilon_i, i = 1, \ldots, N$, for $x_i \in R$ and $\epsilon_i \sim (0, \sigma^2)$. Assume we fit $f(\cdot)$ using a linear smoother, which gives the N-vector

$$\widehat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda\mathbf{y}$$

where $\mathbf{S}_\lambda$ is an $N \times N$ smoother matrix. Assume that $\mathbf{S}_\lambda \mathbf{1} = \mathbf{1}$, which implies that the row sums of $\mathbf{S}_\lambda$ are 1. Define the leave-one-out prediction by

$$\hat{f}_\lambda^{-i}(x_i) = \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{S_{ij}}{1 - S_{ii}} y_j. \tag{1}$$

(a) Note that (1) implies that $\hat{f}_\lambda^{-i}(x_i) = \sum S_{ij} y_j + S_{ii} \hat{f}_\lambda^{-i}(x_i)$. Use this to show that $|y_i - \hat{f}_\lambda^{-i}(x_i)| \geq |y_i - \hat{f}_\lambda(x_i)|$.

(b) Show that the leave-one-out cross-validation sum of squares can be computed from the original fit $\hat{f}_\lambda$: i.e. it is not necessary to re-fit the model.

(c) Give an explicit formula for the cross-validation sum of squares for the linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

4. **4000 only: bonus for 450** (Based on Exercise 11.5, p.368)

(a) Simulate 100 observations from the model

$$Y = \sigma(a_1^T X) + (a_2^T X)^2 + 0.30\epsilon$$

where $\sigma(x)$ is the sigmoid function $1/\{1 + \exp(-x)\}$, $a_1 = (3, 3)$, $a_2 = (3, -3)$, $X = (X_1, X_2)$ and X_1 and X_2 are independent standard normal, and ϵ is standard normal, independent of X_1 and X_2. Fit a single hidden layer neural network with 10 hidden units.

(b) Generate a test sample from the same model of size 1000, and plot the training and test error curves as a function of the weight decay parameter.

(c) Vary the number of hidden units in the network, from 1 to 10, and determine the minimum number needed to perform well for this task.