B555: Homework 4

1. The code below can be used to create an $n \times k$ data matrix, X, and an $n \times 1$ class vector, c, for a classification problem. Here c gives the true class of each vector (row of X). The $k \times 2$ matrix p is such that $p[j, z]$ gives the probability that the jth feature is 1 when the class is z. The code fragment below generates X and c in two different ways that are each consistent with the model given by p

```
k = 13;
n = 1000;
p = matrix(0,nrow=k,ncol=2);
p[,1] = seq(.2,.8,length=k);
p[,2] = seq(.8,.2,length=k);
X = matrix(0,n,k);
c = 1+rbinom(n,1,.75);
for (i in 1:n) {
    u = runif(1);
    for (j in 1:k) {
#         X[i,j] = rbinom(1,1,p[j,c[i]]);       # use this line for dataset 1
        X[i,j] = if (u < p[j,c[i]])  1 else 0; # use this line for dataset 2
    }
}
```

   (a) For both data sets construct the Naive Bayes clasifier using the first half of the data set and test it on the 2nd half, computing your error rate.

   (b) One classifier performs better than the other. Explain in detail exactly why this is the case.

   (c) For both data sets either

      i. train and test, as above, the approximate Bayes' classifier (Bayes' classifier using the parameters estimated from the training examples), or

      ii. argue that Naive Bayes classifier *is* the approximate Bayes classifier.

2. This problem extends the logistic classification example worked in class. For the cases below build a two-class logistic classifier that distinguishes each class: *setosa*, *virginica*, *versicolor* from the other two combined. For this example construct the classifier using all of the data and test using all of the data. Normally it is inadvisiable to test on the training data as we do here, though it is generally not too misleading when working with a low-parameter model.

   (a) Implement Newton-Raphson gradient descent to improve on the optimization method presented in class, and test on training data, computing your error rate. NR should converge very quickly to a local optimum.

   (b) As we saw in class the classifier isn't always able to distinguish the classes due to the linear decision boundary implicit in the parametric form of logistic regression. Create new features by augmenting the original features with the squares and pairwise products of these features. Construct the classifiers optimizing with Newton-Raphson and report your error rates.

3. The standard regression model is

$$y = Xw + e$$

where X is an $n \times d$ matrix of predictor variables, y is an n-vector of response variables, and $e \sim N(0, \sigma^2 I)$ is an n-vector of model errors. Suppose we introduce a prior distribution on w as $w \sim N(0, \rho^2 I)$

   (a) What is the density of y in terms of X, w, σ^2?

   (b) Write down the joint density function $p(w, y)$

   (c) $\hat{w}_{\text{MAP}}$ is defined to be the value of w that maximizes the posterior probability on w. Show that

$$\hat{w}_{\text{MAP}} = \arg\min_{w} \frac{||w||^2}{\rho^2} + \frac{||y - Xw||^2}{\sigma^2}$$

(d) How does $\hat{w}_{\text{MAP}}$ compare with ridge regression?

4. Consider the regression model with $e_1, e_2, \ldots, e_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Note that the errors in the case are not 0-mean, but rather have an unknown mean, μ. Derive an unbiased estimate for μ in this model.

5. The class website has two data sets for this homework assignment. The first has an $n \times d = 1000 \times 50$ data matrix (X) "pred1.dat" with a 1000×1 response vector (y) "resp1.dat." The second has a 1000×500 data matrix "pred2.dat" with a response vector "resp2.dat." These data sets were generated according to the standard linear regression model, as described above.

   (a) For each data set use the first half of the data (observations $i = 1, \ldots, n/2$, all d predictors) to get the minimum varianced unbiased estimate of w, $\hat{w}$.

   (b) For the first data set using only the first half of the data compute and report an unbiased estimate of σ^2. Describe any difficulties you would have in performing the same task for the 2nd data set.

   (c) For each data set, use your estimate of w on the 2nd half of the data set $(n/2+1, \ldots, n)$, to get your estimated response variables, $\hat{y}$ and compute and report your total squared error:

$$SSE = \sum_{i=n/2+1}^{n} (\hat{y}_i - y_i)^2$$

6. In *variable selection* we iteratively grow a collection of variables that are used predictor variables. We begin by finding the single predictor variable that gives the smallest value of $SSE = \sum(y_i - \hat{y}_i)^2$, making this our first predictor variable. We then add another predictor variable (column of our data matrix) by again choosing the variable that, when combined with our first variable, gives the smallest value of SSE. In our simple implementation we implement this greedy search until we have have a fixed number of variables.

   (a) Implement variable selection on the first half of the first data set to identify the 3 best predictor variables. Report the three variables you get and the three decreasing values of SSE they produce.

   (b) Compute the value for SSE on the 2nd half of the data set using the model you have learned. Compare this SSE with that obtained using all predictor variables. Which approach gives a better SSE and why?

7. Use the first half of the 2nd data set to perform ridge regression on w using a parameter of $\lambda = 20$ to get a new $\hat{w}$.

   (a) Using your new $\hat{w}$, compute the estimated response, $\hat{y}$, for the 2nd half of the dataset.

   (b) Compare the resulting sum of squared errors on the 2nd half of the dataset, using both ridge regression and plain regression. If one of the two methods performs better, explain why.

   (c) In general, we do not know what the best choice of the smoothing parameter, λ will be. One way to choose the parameter would be to try a variety of values estimated using the first half of the dataset, choosing the value that gives the best performance on the 2nd half of the dataset. This is known as cross validation. Use this idea to estimate your best choice of smoothing parameter, λ.

8. A time series is a sequence of observations taken over time, usually with constant time between measurements. The data on the website "time_series.dat" was taken from a time series model

$$x_i = \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + e_i$$

   where the $\{e_i\}$ are modeled as independent $N(0, \sigma^2)$ random variables. Estimate the parameters $\alpha_1, \alpha_2, \sigma^2$ from these data.