Statistical Data Mining I
Homework 3

Due: Monday November 4 (11:59 pm)
40 points

Directions: If you complete all exercises, only the first four will be graded. Please adhere to the homework guidelines posted in UB learns.

1) (10 points) **MS + MPS**
Using the Boston data set (ISLR package), fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA and kNN models using various subsets of the predictors. Describe your findings.

2) (10 points) **MS + MPS**
Download the diabetes data set (http://astro.temple.edu/~alan/DiabetesAndrews36_1.txt). Disregard the first three columns. The fourth column is the observation number, and the next five columns are the variables (glucose.area, insulin.area, SSPG, relative.weight, and fasting.plasma.glucose). The final column is the class number. Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data.

(Note: this data can also be found in the MMST library)

(a) Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?

(b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?

(c) Suppose an individual has (glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?

3) **MS only**
a) Under the assumptions in the logistic regression model, the sum of posterior probabilities of classes is equal to one. Show that this holds for k=K.
b) Using a little bit of algebra, show that the logistic function representation and the logit representation for the logistic regression model are equivalent.
In other words, show that the logistic function:

$$p(X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

is equivalent to:

$$\frac{p(X)}{1-p(X)} = \exp(\beta_0 + \beta_1 X).$$

4) **MS + MPS**

(10 points) We will now perform cross-validation on a simulated data set. Generate simulated data as follows:

```
> set.seed(1)
>x=rnorm(100)
>y=x-2*x^2+rnorm(100)
```

a) Compute the LOOCV errors that result from fitting the following four models using least squares:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$$

a) Which of the models had the smallest LOOCV error? Is this what you expected? Explain your answer.

b) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in part c using least squares. Do these results agree with the conclusions drawn from the cross-validation?

4) (10 points) **MPS only**

This question uses the "Weekly" dataset in the ISLR package. The data contains information for weekly returns for 21 years, beginning in 1990 and ending in 2010.

a) Produce some numerical and graphical summaries of the "Weekly" data. Do there appear to be any patterns?
b) Use the full data to perform logistic regression with "Direction" as the response and the five lag variables, plus volume, as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? Comment on these.
c) Compute the "confusion matrix" and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
d) Fit the logistic model using a training data period from 1990-2008, with "Lag2" as the only predictor. Compute the confusion matrix, and the overall correct fraction of predictions for the held out data (that is, the data from 2009 and 2010).
e) Repeat (d) using LDA.
f) Repeat (d) using KNN with k=1.
g) Which method appears to provide the best results?
h) Experiment with different combinations of predictors, including possible transformations and interactions, for each method. Report the variables,

method, and associated confusion matrix that appears to provide the best results on the held-out data. Note that you should also experiment with values for K in the kNN classifier.