Statistical Data Mining I
Homework 3

Due: Sunday November 19th (11:59 pm)
40 points

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

*New Homework Guidelines: Your write up must be submitted as a *.pdf. document. All materials need to be "bundled" into a compressed file (e.g., zip) and named with your UB IT name. This compressed file can be uploaded into the dropbox.*

1) (10 points) 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) 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) 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) (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:

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

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?

5) (10 points) When the number of features (p) is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when p is large.

a) Suppose that we have a set of observations, each with measurements on p = 1 feature, X. We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0. 6, we will use observations in the range [0. 55, 0. 65]. On average, what fraction of the available observations will we use to make the prediction?

b) Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2 . We assume that (X1, X2) are uniformly distributed on [0, 1] x [0, 1]. We wish to predict a test observation's response using on observations that are within 10% of the range of X1 and within 10% of the range of X2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0. 6 and X2 = 0. 35, we will use observations in the range [0. 55, 0. 65] for X1 and in the range [0. 3, 0. 4] for X2. On average, what fraction of the available observations will we use to make the prediction?

c) Now suppose that we have a set of observations on p = 100 features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

d) Use your answers from (a-c) to argue the drawback of KNN when p is large.