Statistical Data Mining I
Homework 3

Due: Friday November 2 (11:59 pm)
30 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) 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:

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