PH245
Introduction to Multivariate Statistics
Homework Set 5

Due date: December 4, Monday

Problems:

1. Breast cancer continues to be a common and deadly form of cancer among women. The diagnosis of breast tumors has traditionally been performed by biopsy, an invasive and often traumatic surgical procedure. Alternatively, diagnosis might be based on fine needle aspiration (FNA) which can often be performed on an outpatient basis. In 1994 researchers at the University of Wisconsin studied FNA as a method of diagnosis, and developed a dataset that can be found at "Data-HW5-breastcancer.dat". It contains information on 569 FNAs. There are two diagnoses (classes), 212 malignant and 357 benign. The 30 predictors correspond to the mean, standard deviation and a tail average of the empirical distributions of 10 characteristics of the cells extracted by FNA.

   (a) Partition the full data set into a training set of 400 patients, and a testing set of 169 patients. Please use the following command `set.seed(1000)` to set the random seed of your partition, so that you can *reproduce* all your analysis results.

   (b) Fit LDA, QDA, MDA (with number of subclasses equal to $(5, 5)$), Nearest Neighbor (with $k = 5$), and CART. Report the misclassification error rate on the *testing* data set.

   (c) Fit MDA, with number of subclasses equal to $(1, 1)$, $(5, 5)$, and $(10, 10)$, respectively. Report the misclassification error rate on both the *training* and the *testing* data. Please describe the pattern, and see if it agrees with your expectation.

   (d) Fit Nearest Neighbor, with the number of neighbors $k = 1, 5$ and 10, respectively. Report the misclassification error rate on both the *training* and the *testing* data. Please describe the pattern, and see if it agrees with your expectation.

2. (**Optional**) The data file "Data-HW5-university.dat" contains the data on some universities for certain variables used to compare or rank major universities. These variables include X_1 = average SAT score of new freshmen, X_2 = percentage of new freshmen in top 10% of high school class, X_3 = percentage of applicants accepted, X_4 = student-faculty ratio, X_5 = estimated annual expenses, and X_6 = graduation rate (%).

(a) Use Mclust() in R to analyze this data. Report the best model using BIC.

(b) Plot X_2 (Top10) versus X_5 (Expenses) with different clusters and university names marked out.

(c) Apply kmeans to this data with the number of clusters equal to the best number found in (a).

(d) Repeat (b) with clusters found by kmeans now, and compare it with the results found in (b)

(e) Apply hierarchical clustering to this data with average linkage. Report the clustering results using the number of clusters equal to the best number found in (a).

Policy: You must do the homework on your own. Please ask the Instructor or the GSI if you have any question.