Project 2: Clustering/Classification of Microarray Data

I. K Nearest Neighbors

For this part of the assignment, you will be performing 4-fold cross validation using KNN on the entire data set comprised of 7129 genes over 72 patients. (Note: to make sure that the two cancer types are equally represented, divide the two groups separately and then combine).

1. With p=0.5 run 4-fold cross-validation on the following values of K: K=1,5,10,15,30,50.

a) Plot the cross validation accuracy vs. K, and save the file as "question1.jpg".

b) Qualitatively, how is the clustering performance affected by different values of K? (describe general trends)

ALL (+) mis-classified as AML (-) = False Negative

c) Are there any ALL patients that are consistently misclassified as AML across the different values of k? If so which patient(s) (as identified by column number)?

2. Using the value of K= 30, run 4 fold cross validation for the following different values of p: 0%, 5%, 20%, 50%, 75%, 95%, 100%.

a) Plot ROC curve by plotting the sensitivity vs. (1-specificity) for varying values of p, and save the file as "question2.jpg".

b) Describe in words the trend in specificity and sensitivity as p increases.

II a. K-means clustering

3. Run kmeans using K=3 for the testdata.dat file using the testdata_centroids.dat file as the starting centroids. In this simple, made up case we can imagine the data as being points with x and y coordinates on a 2-dimensional graph. Each line corresponds to one point.

After running the program, graph the points and note which point belongs to which of the three clusters. Save the file as "question3.jpg".

II b. K-means clustering with yeast microarray data

For the rest of the assignment, run K-means on the **yeast microarray data** described in the instructions. We will cluster the genes based on their expression levels.

Of the genes represented on the 79 microarrays, 121 were previously characterized as ribosomal genes. The ribosome is a large complex of many proteins that facilitates the translation of mRNA into protein; they are the cellular machinery responsible for linking together the correct sequence of amino acids from a sequence of codons. The proportion of each protein present in the complex is coordinated in the cell to ensure the correct number of subunits is available to construct complete ribosome molecules. The cell often regulates the amount of protein by controlling the transcription level of the protein's gene. Therefore, we might expect many of the ribosomal genes to be coordinately regulated -- they should have similar mRNA expression levels. Also, given a number of known ribosomal proteins and their expression patterns across a wide range of experiments, we might be able to find other ribosomal proteins by comparison.

Scanning yeast_gene_names.txt you will see that the ribosomes are the last 121 genes in the file.

4. Run K-means with K=2. Pick gene #1 (the 1st gene in yeast.dat, a non-ribosomal gene) and gene #2467 (a ribosomal gene) as your starting centers.

a) Are all the ribosomal genes in the same cluster? If not, list all the ribosomal genes that are in the cluster that is different from the majority of the ribosomal genes (list genes by gene index in yeast.dat).

b) What percentage of genes in each cluster are ribosomal genes?
(enter two % values, separated by a comma)

5. Again run K-means with K=2, but this time choose two random data point as your starting centers (your algorithm should randomly pick 2 genes from yeast.dat, so they will be different for each run).

What percentage of genes in each cluster are ribosomal genes?
(enter two % values, separated by a comma)

6. Now consider the clusters obtained in question 4 and question 5.

a) In 3-5 sentences, compare and contrast the clusters you observed in questions 4 and 5.

b) Based on these observations, what can you say in general about the K-means clustering algorithm?

7. Do K-means clustering on the same dataset for 20 times with K=2 and random starting centers.

a) Out of the 20 runs, are there any ribosomal genes that are often clustered into a different cluster from the majority of the ribosomal genes? If there are, which ones are they (list by gene index)?

b) Out of the 20 runs, how many times does the translation elongation factor EFB1 (gene #1511) cluster with the majority of the ribosomal genes?

c) From a biological standpoint why does your answer to (b) make sense?