Project 2: Clustering/Classification of Microarray Data (40 points total)

You will be asked to generate several output files and images while completing this quiz. These should all be submitted in a directory named "quiz_output" within your submission directory.

I. Pseudocode

1. Following guidelines under the "Quiz" section on the project page, write pseudocode for:

a. knn with n-fold validation (1.5 points)

b. k-means clustering (1.5 points)

II. 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.

2. With p=0.5 run 4-fold cross-validation on the following values of K: K=1,5,10,15,20,30.

a. Save the output of each run as knn.out.k. For example, knn.out.1 will contain the results reported from running with p=.5 and k=1. Save these 6 files in your "quiz_output" directory. (3 points)

b. Plot the cross validation accuracy vs. K, and save the file as "question2.jpg". (2 points)

3. Qualitatively, how are clustering accuracy, sensitivity, and specificity affected by different values of K? Describe general trends observed in the data generated for question 2 above. (3 points)

4. Are there any ALL patients that are consistently misclassified as AML in question 2 across the different values of K? If so, which patient(s), as identified by column number? (2 points)

5. Using the value of K= 10, run 4-fold cross-validation for the following different values of p: 5%, 10%, 20%, 50%, 75%, 90%, 100%. Plot an ROC curve by graphing the sensitivity vs. (1-specificity) for these values of p, and save the file as "question5.jpg". (4 points)

6. Qualitatively, how are clustering accuracy, sensitivity, and specificity affected with increasing values of p? Describe general trends observed in the data generated for question 5 above. (3 points)

III a. K-means clustering with test data

7. 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.

a. Save the output file generated from this run as kmeans_testdata.out in your "quiz_output" directory. (2 points)

b. After running the program, graph the points and note which point belongs to which of the three clusters. Save the file as "question7.jpg". (2 points)

III b. K-means clustering with yeast microarray data

For the rest of the assignment, run K-means on the **yeast microarray data**. 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.

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

8. Run K-means with K=3. Use experimental_centroids.txt as your starting centroids. Save the output file generated from this run as kmeans_experimental.out in your "quiz_output" directory. (2 points)

9. 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 centroids.

a. Are all the ribosomal genes (the last 121 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. (1 point)

b. Ribosomal genes comprise what percentage of the genes assigned to each of the 2 clusters? Enter two % values, separated by a comma. (1 point)

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

For each run, list the 2 centroids selected and what percentage of genes in each cluster are ribosomal genes. (2.5 points)

Example:

run 1. 1111, 2222: 0.0%, 100.0%
run 2. 101, 202: 5.0%, 95.0%
etc

11. Now consider the clusters obtained in question 8 and question 9.

a. In 3 - 5 sentences, compare and contrast the clusters you observed in questions 8 and 9. (2.5 points)

b. Based on these observations, what can you say in general about the K-means clustering algorithm? (2 points)

12. Again run K-means with K=2. Choose two random data points as your starting centers, just as you did in question 9. Run this 20 times.

a. Out of the 20 runs, are there any ribosomal genes (the last 121 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)? (1 points)

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? (1 points)

c. From a biological standpoint, why does your answer to (b) make sense? Cite at least 1 reference. (3 points)