Mid Term

***K*-means Clustering.**

K-means clustering is a method of cluster analysis, which aims to partition *n* observations into k mutually exclusive and exhaustive clusters in which each observation belongs to the cluster with the nearest mean. The k-means clustering algorithm is as follows:

1) Select a set of initial centres of k clusters

2) Assign each object to the cluster with the closest centre

3) Compute the new centres of the clusters:

$$\vec{C}(S) = \sum_{i=1}^{n} \vec{X}_i / n, \vec{X}_1, \ldots, \vec{X}_n \in S$$

4) Repeat step 2 and 3 until no object changes cluster

Gene expression dataset: Normalized human gene expression profiles over the cell cycle in HeLa cell line were obtained from Whitfield et al. (2002). HeLa cells were synchronized at G1/S stage in arresting medium. After releasing from arrest, cells were collected at 1-2 hour intervals and mRNA extracted was analyzed using cDNA microarrays. The gene_expression.txt file contains the gene expression of 1000 genes at 114 different time points/conditions (file format: column 1- Entrez gene ID; columns 2 to114 – normalized gene expression value).

Problem 1: Implementation of *k*-means clustering [50 points]

Implement the k-means clustering algorithm in Perl to cluster the given gene expression dataset using Pearson correlation coefficient as the distance metric. Partition the genes into 10 clusters, using the first 10 genes in the file as initial centers. In your written report, include a table describing the number of genes in each cluster (in ascending order) and the within cluster sum of squares of each cluster. How many iterations did it take to converge to the final cluster?

To better organize your code, write subroutines to calculate the Pearson correlation coefficient and sum of squares. Comment your code carefully to explain the purpose of your code.

Pearson Correlation Coefficient: $r = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

Sum of squares: $\sum_{i=1}^{c}(X_i - \overline{X})^2$

Where n is the number of observations for each gene, c is the number of genes in the cluster, and X is an n-dimensional vector.

Problem 2: Find the optimal number of clusters by the elbow method [50 points]

The k-means clustering algorithm aims to minimize the within cluster sum of squares (WSS). The percentage of variance explained by clustering is ratio of between cluster sum of squares (BSS) and the total sum of squares (TSS). With increasing number of clusters, the within cluster sum of squares decreases, and the percentage of variance explained increases. If you plot the number of clusters against the percent of variance explained, the marginal gain of variance explained will drop at some point, giving an angle in the graph. The number clusters is chosen based on the 'elbow' point (see Wikipedia article http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set).

For a set of observations (X_1, X_2, …, X_m) where each observation is a n dimensional vector, K-means clustering partitions the m observations to k sets, S = {S_1, S_2, …, S_k}.

$$WSS = \sum_{i=1}^{k} \sum_{x_j \in S_i} \| X_j - \mu_i \|^2$$

$$\text{TSS} = \sum_{i=1}^{m} (X_i - \overline{X})^2$$

BSS=TSS-WSS

Percentage of variance explained = BSS/TSS

Perform k-means clustering and partition data into 20, 40, 60, 80, 100, 150 and 200 clusters using the first k genes in the input file as initial centers. Plot the number of clusters against the percentage of variance explained. What is the optimal number of clusters?

Submission:

Please submit the follow files:

1. Written report with your plots and answers to the problems 1 and 2.
2. Perl script for problem 1 and problem 2.

Please put everything in a zipped folder. The name of the folder should be "YourNetID_BTRY6381_MidTerm.zip". This folder should be sent to me (yg246@cornell.edu) via email. Submissions not adhering to the specified format will be penalized. The deadline for submission is 11:59 pm on April 1, 2013. Penalty will be given to late submissions.