CSE514 Fall 17 – Datamining Homework 3

Assigned date: 11/9/2017 **Due date: 11/21/2017**, before class starts

Note: We use the whole AD dataset exclusively here.

Note: You may (or may not) use WEKA

Note: The problems were designed in such a way to show you how a DM problem can be approached. Think about them and try to see if you can come up with your own ideas.

**P0**: Preprocessing (Write your own program for this, although you don’t have to submit your code)

As you may have noticed, the data varied from sample to sample which may not reflect the reality. We may assume that all genes in a cell (or *one sample*) are in the range of [-1, +1], where -1 and +1 correspond to the min and max values of the sample. Preprocess both AD and normal datasets data first following the above scheme. You must use the normalized data for the problems below.

**P1**: Clustering using k-means (20pt)

**The problem**:

Grouping features (i.e., genes in our case) based on their patterns across a set of samples (AD patients or normal people) may provide new insights to the complex gene regulation (i.e., the underlying system that we are interested) in the brain of AD. For this, we apply the k-means clustering method to this problem. However, at least two issues we need to consider for this problem (as well as for similar DM problems): 1) what is the number of clusters? 2) what is the best similarity measure to define clusters? We try to answer these questions here.

**What to do**: To reduce the amount of your work, simple use the AD patient dataset here.

1. Using the *dot product* as the similarity measure, run k-means, with k varies from 2, 3 to 100, to find k clusters of the ~8600 genes. Plot figures to show
   1. the average in-cluster similarity of the k clusters – call it S; (5 pts)
   2. the average between-cluster similarity across any pair of the k clusters – called it D; and (5 pts)
   3. the ratio of S/D for all values of k considered. (2 pts)

For in-cluster similarity, you need to consider all pairs of genes in a cluster, and for between-cluster similarity, you may use the center of mass for a cluster. For distance measure, we will stick to dot product.

1. Repeat 1) but using the Euclidean distance. (4 pts)
2. Compare the results from 1) and 2) and briefly explain what you find. What this exercise tell us about the best number of k? (4 pts)

**P2:** PCA (20 pts) Use both AD and control datasets together.

Apply PCA to the combined dataset, find the eighenvectors and eighenvalues, and sort them from the largest eighenvalue to the smallest.

1. Plot the sorted eighenvalues; plot the accumulative information up to k eighenvalues as a function of number of k. If we want to retain 80% of information, what is the k to use? (10 pts)
2. Plot the AD cases (labeled in red) and normal controls (labeled in green) in a 3D figure where the 3 coordinates correspond to the first 3 largest eighenvectors. (tip: using MatLab if you like.) Can you see a reasonable separation of the AD cases from the normal controls in your plot? Include in your solution the best plot for the best separation you can get. Repeat with a 2D plot using the first 2 largest eighenvectors. (10 pts)