CSE514 Spring 18 – Datamining Homework 3

Assigned date: 3/24/2018 **Due date: 4/10/2018**, before class starts

Notes:

* Similar to HW2, use the first 800 genes (features) for this HW, but use all instances for AD subjects and normal controls.
* You may use WEKA or a system you are comfortable with.

Concepts to learn in this HW:

* Effects of different similarity measures; we will explore this in the k-means clustering.
* Dimension reduction by PCA and SVD. Particularly, we will study the number of effective dimensions in a complex problem.
* Relationship between PCA and SVD

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

* For this set of problems, we will use the data after missing data imputation using the method from HW2 – i.e., impute missing data using 5-NN. Also remove one problem instance where gene values are the same; this is apparently a corrupted sample in the dataset.
* As you may have noticed, the data varied from sample to sample which may affect the downstream analysis. Since we are interested in relative variations in the data, we may assume that expression values of 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. Normalize both AD and normal samples first by forcing their values to be in the range of [-1, +1]. You must use the normalized data for the problems below.

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

**The problem**:

Grouping features (i.e., **genes** (not samples) 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, simply use the AD patient dataset and consider the first 800 genes 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 ~800 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; (5 pts) and
   3. the ratios 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 here.

1. Repeat 1) but using the Euclidean distance. (4 pts)
2. Compare the results from 1) and 2) and briefly explain what you find by using two different similarity measure. Furthermore, what is the best number of k for the two similarity measure? (4 pts)

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

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

1. Plot the sorted eighenvalues. (2 pts) What can you say about the effective number of dimension from this PCA analysis? (4 pts)
2. Plot the accumulative information up to k largest eighenvalues with an increasing order of k. If we want to retain 80% of information, what is the k to use? (4 pts)
3. 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. (8 pts) Repeat with a 2D plot using the first 2 largest eighenvectors. (2 pts)

**P3:** SVD (15 pts) Use both AD and control datasets together (the datasets after preprocessing above).

Apply SVD to the dataset of AD and control combined, find the left and right singular vectors and their corresponding singular values. Sort the two sets of singular values from the largest to the smallest.

1. Plot the sorted left and right singular values in one figure. (2 pts) Discuss what you observe from the plot. (4 pts). What can you say about the effective number of dimension from this PCA analysis? (4 pts)
2. Which set of (left or right) singular vectors and singular values are related to the eigenvectors and eigenvalues of PCA from P2 above? Discuss their relationship. (4 pts)