CSE514 Fall 16 – Datamining Homework 3

Assigned date: 11/09/2016 **Due date: 11/29/2016**, before class starts

Tips: Follow the tips given in HW1. Start as early as possible and don’t wait until the last few days.

**P1**: (5 pts each for each individual subproblem below) Hierarchical clustering

We consider the Alzheimer’s Disease dataset(s) again. Consider data on an AD patient or normal person as one data point and the ~7000 genes as features. Ignore the labels (patients or normal controls) and treat each data point equally. We use Pearson Correlation coefficient in this problem.

1. Let’s fix the missing value problem one more time. This time, we use k-NN with k=7 and use the median of the 7-NN to impute a missing value. That is, when imputing a missing value (gene expression value), we use the other genes as features for finding k-NN. Apply this to the patient data and control data separately. (note, this method should be in general better than taking the average.) For the rest of the problem (and the rest of the class, use your missing value imputed dataset. (note – none of you suggested this method in previous HWs ☹
2. Bottom-up. Find two clusters using the bottom-up approach. Specify in the solution where your method starts. Compare the top two clusters with actual patient vs. control groups to see how the clustering results and real labels agree. List the data points that are placed to the wrong group.
3. Top-down. This is basically a partition based approach and we may use it to detect outliers in the data. Recursively use k-means (with all the genes as features) to recursively partition the whole set of data (patients and controls) into two sets. Describe where you stop in your solution. Compare the top two clusters with actual patient vs. control groups to see how the clustering results and real labels agree. List the data points that are placed to the wrong group.
4. Compare the data points that are placed to the wrong groups in 1) and 2) above.
5. (bonus 5 pts) Can you come up with a good way to identify real outliers? Describe your method and back up with some reasoning.

**P2:** (20 pt) PCA

We consider the Alzheimer’s Disease dataset(s). Apply PCA to the whole dataset with no labels (patients and controls combined).

1. (5 pts) Order the eigenvalues in a nonincreasing order and graph them in an X-Y plot to see how these values decrease.
2. (15 pts) Now use the top k, for k=2 to 20, eigenvectors as features and cluster the samples (data points) into 2 groups using the k-means and Pearson Correlation coefficient as distance measure. This means we now work in the “PCA eigen space” using its coordinates (or features that are the eignvectors). Compare the clustering results with the actual sample labels to see when you have a good clustering result. You may use some plots to show your results.

**Note**: this problem also tells as we can take advantage of the PCA space for data analysis.

**P3:** (Bonus 15 pt) SVD – this is optional

You need to repeat P2 but this time using SVD. You need to formulate and solve the problem in SVD. Make it as complete as possible if you decided to work on it. My bonus pts do not come out easily ☺