CSE514 Fall 16 – Datamining Homework 2

Assigned date: 10/25/2016 **Due date: 11/10/2016**, before class starts

Tips: Follow the tips given in HW1.

**P1**: (10pt)

Probabilistic model-based cluster is the following. We are given a set of data points D={xi; i=1… n}, which are assumed to be drawn from a probabilistic model consisting of k probabilistic distributions

F(x) = ∑i=1…k πi fi(x;Θ).

Assume fi(x;Θ) ~ N(µ, σ) (normal distribution with mean µ and standard deviation σ).

1. Describe an E-M algorithm to find F(x) that can best fit the data.
2. Apply your algorithm to the following data D={35, 41, 21, 20, 17, 55, 12, 33, 15, 18, 4, 51, 17, 46}, which are assumed to be drawn from F(x) = π1 N(µ1, σ1) + π2 N(µ2, σ2). Your output will be the parameters of the F(x), i.e., {π1, π2, µ1, µ2, σ1, σ2}.

**P2:** (20 pt)

From Problem 4 of HW1. In your Random Forest, you should use at least 100 trees. Take out the 200 genes that are most frequently used in your Random Forest. Now ignore the labels and mix the data for patients (cases) and normal controls (ctrl) together – i.e., all the data have no label at all.

1. Use the 200 genes as features (dimensions) and the k-means algorithm to cluster all the data (patients and controls) into k=2 clusters. Use Euclidean distance for the similarity measure.
2. Repeat step 1 for k=3 and k=4.
3. Repeat steps 1 and 2 for k=2, 3, 4 using dot product as the similarity measure.
4. Evaluate the qualities of all of the clustering results above using the “Distance btw clusters” (single link, complete link, average and centroid) that we discussed in class (also in my lecture notes). (tip: you may create a table to show your final results)
5. Discuss what you see and learn from this clustering problem.

(note: how do you deal with missing value?)

**P3:** (15 pt)

Feature selection. From Problem 4 HW1. Consider the labeled AD data again (patients vs controls). We can use the linear kernel (i.e., k(x, x’) = (x x’)) and SVM to select the top k features – see the lecture notes for detail.

1. Using this method to select k=200 genes from the AD data.
2. Now use the quadratic kernel k(x, x’) = ((x x’)+1)2 to select the top 200 features in the feature space. What are these features in the input space (or space of genes)? Try to get as much information on genes from your results as possible.