CS 6350 − Big Data Management and Analytics
Spring 2016
Assignment #4 Part A
Due: 4/17/2016

Note: You must show your work to receive credit. A correct answer without showing your reasoning/work will not receive credit. Please write legibly, ensure electronic submissions are readable and check that pages/problems are in correct ordering.

Problem 1 Given the following utility matrix which contains ratings (1 to 5 star(s) scale) on six items (i_1 to i_6) by four users (U_1, U_2, U_3 and U_4):

	i_1	i_2	i_3	i_4	i_5	i_6
U_1		$\star\star\star$		$\star$	$\star\star$	$\star\star\star$
U_2		$\star\star\star\star\star$	$\star\star\star\star$		$\star\star\star\star$	
U_3			$\star\star$	$\star\star\star\star\star$		$\star$
U_4	$\star\star\star\star$	$\star$		$\star\star\star\star$		$\star\star$

Compute the following from the data of this matrix. Please show all your work.

(a) Compute the Jaccard distance between each pair of users.

(b) Compute the cosine distance between each pair of users.

(c) Cluster the six items hierarchically into four clusters. Use Jaccard distance to measure the distance between the resulting column vectors. For clusters of more than one element, take the distance between clusters to be the minimum distance between pairs of elements, one from each cluster.

Problem 2 Given the following eight points ($P_1, \ldots, P_8$) in 1-dimensional space with corresponding x-coordinates:

Point	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
x	1.25	2.5	3.4	5.1	6.25	7.2	8.3	10.5

(a) For each of the following hierarchical agglomerative clustering approaches, draw the resulting dendrogram using Euclidean distance. Please show all distance calculations and the dendrogram should *clearly show the order* in which the points are merged.

(i) Single link

(ii) Complete link

(b) Assuming you have the following two clusters: $\{P_1, P_2, P_3, P_4\}$ and $\{P_5, P_6, P_7, P_8\}$. Calculate the Euclidean distance between the two clusters according to:

(i) Single link

(ii) Complete link

(iii) Group average

Problem 3 Given the following 8 points in 2-dimensional space with corresponding coordinates:

Point	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
x_1	2	3	4	5	6	8	8	9
x_2	2	4	7	3	7	7	1	3

Using k-means and Euclidean distance, with $k = 2$ and centroid locations at $(x_1 = 2, x_2 = 2)$ and $(x_1 = 3, x_2 = 5)$, calculate the coordinates of the centroids and the cluster assignments for each iteration until convergence.

Problem 4 Running the BFR (Bradley, Fayyad, Reina) algorithm results in a discard set containing the following points (in high dimensional Euclidean space)

Point	P_1	P_2	P_3	P_4
x_1	5	3	1	4
x_2	2	3	9	1
x_3	7	3	6	2
x_4	4	3	1	7

with the centroid at $(x_1 = 2,\ x_2 = 7,\ x_3 = 4,\ x_4 = 5)$. Calculate the following for the discard set.

(i) N (ii) SUM (iii) SUMSQ

(iv) normalized Euclidean distance between the centroid and each point

Hint: the normalized Euclidean distance for point $(x_1, \ldots, x_N)$ and centroid $(c_1, \ldots, c_N)$ is

$$d(x, c) = \sqrt{\sum_{i=1}^{N} \left(\frac{x_i - c_i}{\sigma_i} \right)^2}$$

where σ_i is the standard deviation of points in the cluster in the ith dimension.

Problem 5 List the differences between BFR and CURE (Clustering Using REpresentatives) clustering algorithm.

Problem 6 Given the following training data. X_1, X_2, X_3, X_4 are the features and Y is the class.

X_1	X_2	X_3	X_4	Y
a	e	j	u	$+1$
b	f	k	v	$+1$
b	f	j	v	-1
a	f	k	u	-1
a	f	j	u	$+1$
a	f	k	u	-1
b	e	k	v	$+1$
a	f	j	v	-1
a	f	k	v	-1
b	e	k	v	$+1$

Using the ID3 decision tree algorithm with information gain heuristic, learn the first two levels of the decision tree. Show all information gain and entropy calculations ($\log_2$). Draw the decision tree. What is the classification accuracy of the training data set using the learned two level decision tree?