CSE 482: Big Data Analysis (Spring 2020) Homework 4 Part 1

Due date: April 5, 2020 (before midnight)

Submit a PDF file containing the solutions for all the questions below.

1. Consider the training data shown in the table below:

Data point	x_1	x_2	class
p1	0.3	0.2	+
p2	0.2	0.45	+
p3	0.5	0.2	+
p4	0.1	0.1	+
p5	0.4	0.1	+
p6	0.25	0.8	-
p7	0.3	0.5	-
p8	0.4	0.8	-
p9	0.15	0.7	-
p10	0.3	0.7	-

Consider a test instance, $(x_1 = 0.2, x_2 = 0.55)$.

(a) Compute the Euclidean distance between the test instance to all the training instances.

(b) Based on your answer in part (a), classify the test instance using the 1-nearest neighbor approach.

(c) Based on your answer in part (a), classify the test instance using the 5-nearest neighbor approach.

(d) Which classification result (part (b) or (c)) do you think is more reliable for the given test point? Explain your reason.

(e) Consider the following logistic regression model constructed from the training set shown in the table above:

$$\log\left(\frac{P(\hat{y}=1|x_1,x_2)}{P(\hat{y}=-1|x_1,x_2)}\right) = w_2x_2 + w_1x_1 + w_0$$

where $\hat{y}$ is the predicted class, $w_2 = -122.1774$, $w_1 = -73.36$, and $w_0 = 73.3023$. Apply the logistic regression model on the given test instance to predict its class label. Show your computations clearly.

2. Consider the following set of one-dimensional points: {0.1, 0.2, 0.8, 0.9, 1.0,1.3, 1.8, 1.9}.

(a) Suppose we apply kmeans clustering to obtain three clusters, A, B, and C. If the initial centroids of the three clusters are located at {0.1,

0.2, 1.9}, respectively, show the cluster assignments and locations of the centroids after the first three iterations by filling out the following table.

	Cluster assignment of data points (enter A, B, or C)								Centroid Locations		
Iter	0.10	0.20	0.80	0.90	1.00	1.30	1.80	1.90	A	B	C
0	-	-	-	-	-	-	-	-	0.10	0.20	1.90
1											
2											
3											

(b) Compute the sum-of-squared errors (SSE) for the clustering solution in part (a).

(c) Repeat part (a) using {0.8, 1.0, 1.8} as the initial centroids. Show the cluster assignments and locations of the centroids after the first four iterations by filling out the following table.

	Cluster assignment of data points (enter A, B, or C)								Centroid Locations		
Iter	0.10	0.20	0.80	0.90	1.00	1.30	1.80	1.90	A	B	C
0	-	-	-	-	-	-	-	-	0.80	1.00	1.80
1											
2											
3											
4											

(d) Compute the sum-of-squared errors (SSE) for the clustering solution in part (c). Which solution is better in terms of their SSE?

3. Consider the transaction database shown in the table below.

Table 1: Transaction database=.

Transaction ID	Items Purchased
1	Bread, Coffee, Sugar
2	Bread, Eggs, Milk
3	Bread, Butter, Milk
4	Coffee, Milk
5	Bread, Butter, Eggs, Cookies
6	Milk, Sugar
7	Bread, Butter, Eggs, Milk, Sugar
8	Bread, Butter, Milk, Cookies
9	Bread, Butter, Eggs, Milk
10	Butter, Coffee, Milk

(a) Assuming the minimum support threshold for frequent itemsets is 40%, list all the frequent 2-itemsets of the data along with their support values.

(b) Based on your answer in part (a), generate all the candidate 3-itemsets using the candidate generation approach described in the

lecture. You may assume items in an itemset are ordered in increasing alphabetical order.

(c) Assuming the minimum support threshold for frequent itemsets is 40%, which of the candidate 3-itemsets in part (b) are frequent?

(d) Extract all the candidate rules from the frequent itemsets found in part (c).

(e) Based on your answer in part (e), find all the rules whose confidence is more than 70% and support is at least 40%. For this question, you need to focus only on the rules that can be extracted from the frequent 3-itemsets found in part (e). Note: you do not have to use the Apriori implementation to extract the rules.