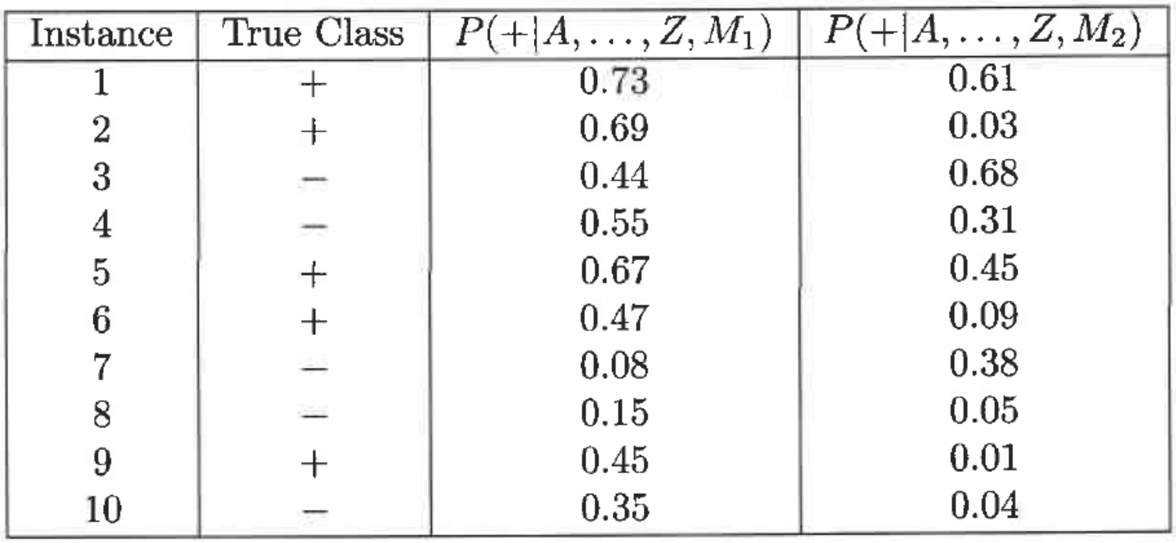
CSE 5334: Data Mining
Homework 2 (100 points)

SUBMISSION DEADLINE: 05/02/2023, 11:59 PM

1. Classifier Evaluation (25 Points)

You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as, A through Z. The table below shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(-) = 1- P(+) and P(- I A, . . . , Z) = 1 - P(+ I A, . . . , Z). Assume that we are mostly interested in detecting instances from the positive class.

a) Construct the ROC curve table to show the TPR(True Positive Rate) and FPR(False Positive Rate) values for each threshold that you should consider. Plot the ROC curve and decide which classifier is better. (15 points)



Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

Table 1

b) For model M1, suppose you choose the cutoff threshold to be t = 0.5. In Other words, any test instances whose posterior probability is greater than 0.5 will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value. (5 points)

c) For model M2, choose the same cutoff threshold t = 0.5 and compute the precision, recall, and F-measure for the model M2. (5 points)

2. K-means Clustering (25 Points)

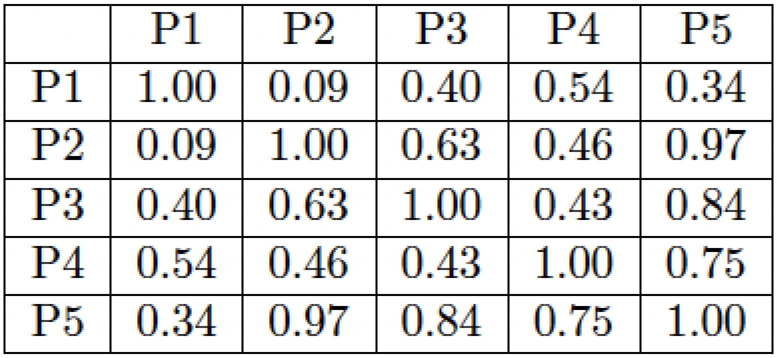
Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as: P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration and show the points in each cluster.

3. Hierarchical Clustering (30 points)

Below is a similarity matrix. (the larger the value is, the more similar two points are)



	P1	P2	P3	P4	P5
P1	1.00	0.09	0.40	0.54	0.34
P2	0.09	1.00	0.63	0.46	0.97
P3	0.40	0.63	1.00	0.43	0.84
P4	0.54	0.46	0.43	1.00	0.75
P5	0.34	0.97	0.84	0.75	1.00

Perform the following two hierarchical clusterings.

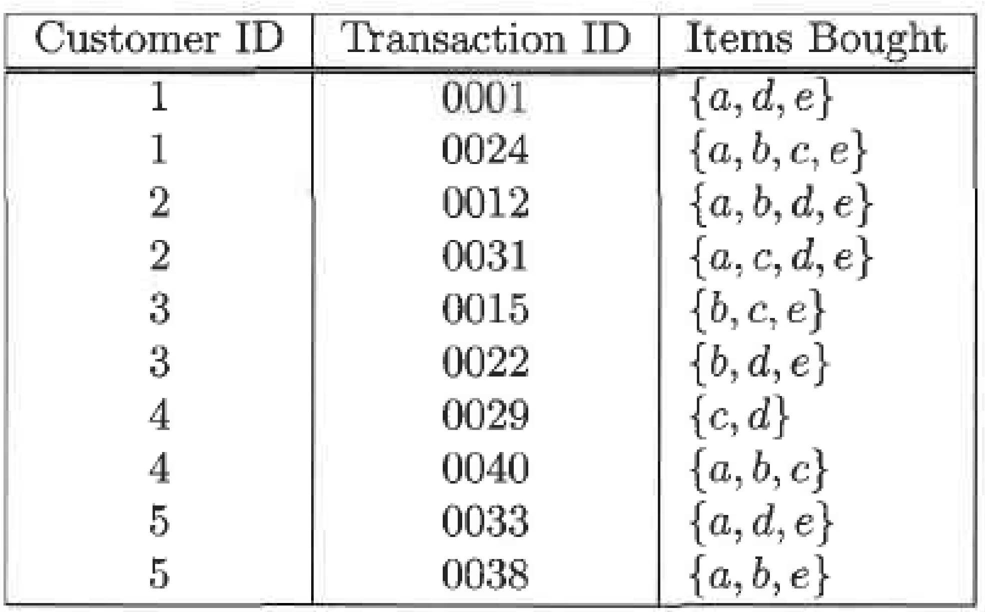
single link (15 points): the proximity of two clusters is defined as the minimum of the distance (maximum of the similarity) between any two points in the two different clusters.

complete link (5 points): the proximity of two clusters is defined as the maximum of the distance (minimum of the similarity) between any two points in the two different clusters.

Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

4. Association Rule Mining (20 points)

Consider the market basket transactions in the data set shown below:



Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

a) Compute the support for itemsets {e}, {b,d}, and {b,d,e} by treating each transaction ID as a market basket. (5 points)

b) Use the results in part (a) to compute the confidence for the association rules {b,d} → {e} and {e} → {b ,d}. (5 points)

c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.) (5 points)

d) Use the results in part (c) to compute the confidence for the association rules {b, d} → {e} and {e}→ {b,d,}. (5 points)