BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

First Semester 2023-24

COMPREHENSIVE EXAMINATION

PART A (OPEN BOOK)

CS F415 Data Mining

**Date : 06-12-2023(FN)** **Max.Duration : 120 minutes** **Max.Marks : 50 (25%)**

Instructions: Write precise and to the point answers. *Only text book, reference books, class notes, & lecture slides are allowed.*

1. In a high-dimensional space, it is often desirable to find contextual outliers. Suppose Waltmart may issue its customers (loyalty) cards and store its customers' shopping records, together with product information. Outline a method that may find extremely valuable customers in different contexts. [5]

2. Explain how AdaBoost, while iteratively training weak learners, adapts by assigning increasing importance (higher weights) to incorrectly classified samples. Clarify how this process minimizes the exponential loss function, emphasizing the relationship between weight adjustments and the algorithm's focus on correcting earlier mistakes. [5]

3. A database has five transactions. Let *min sup* = 60% and *min conf* = 80%.

| *TID* | *Items-bought* |
|---|---|
| T100 | M, O, N, K, E, Y |
| T200 | D, O, N, K, E, Y |
| T300 | M, A, K, E |
| T400 | M, U, C, K, Y |
| T500 | C, O, O, K, I ,E |

(a) Find all frequent itemsets using Apriori method. and FP-growth, respectively. Compare the efficiency of the two mining processes.

(b) List all of the *strong* association rules (with support *s*=60% and confidence *c*=80%) matching the following metarule, where *X* is a variable representing customers, and *item$_i$* denotes variables representing items (e.g., "*A", "B"*, etc.):

*buys*(*X; item*1) *and buys*(*X; item*2) *) => buys*(*X; item*3) [*s; c*]

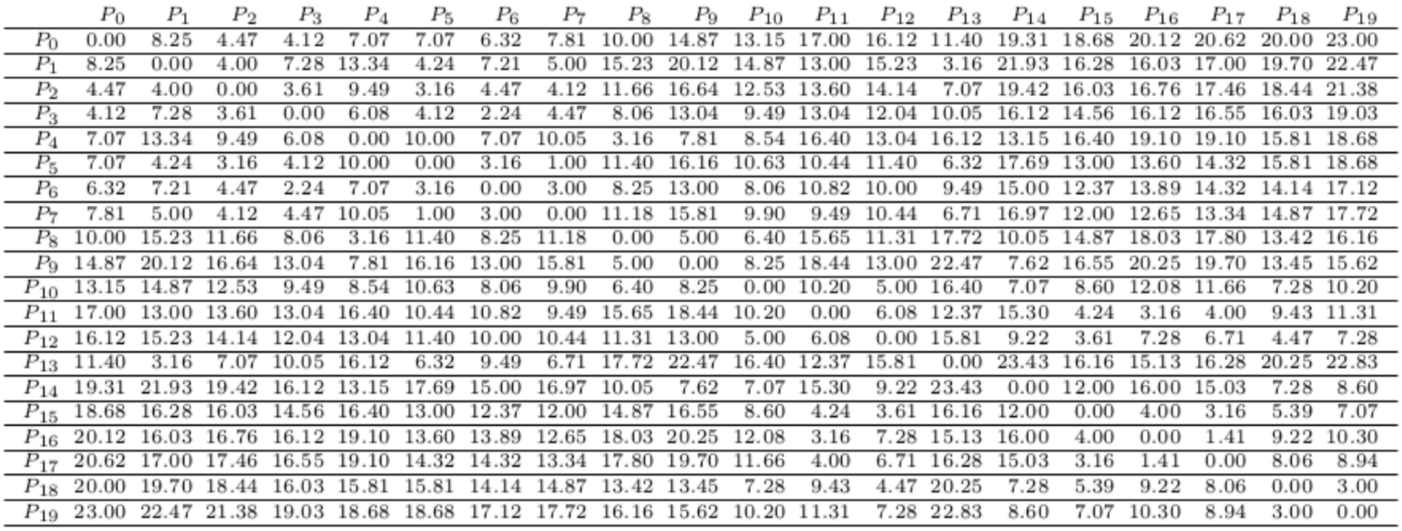
[10+4=14]

4. Consider data points with two attributes x1 and x2, and a corresponding target value, y as shown in the table below.

| x1 | x2 | y |
|---|---|---|
| 1 | 7 | 1 |
| 3 | 7 | 1 |
| 3 | 9 | 1 |
| 4 | 5 | 1 |
| 5 | 7 | 1 |
| 6 | 2 | 0 |
| 7 | 3 | 0 |
| 8 | 2 | 0 |
| 8 | 5 | 0 |
| 9 | 4 | 0 |
| 9 | 7 | 0 |
| 9 | 8 | 0 |
| 10 | 6 | 0 |

A Logistic Regression based classifier is being trained for the data. Consider an iteration when weights are w0=-0.4, w1=-0.1, and w2=0.3 at this stage determine the accuracy of the current model. What would be the value of weights in the immediate next iteration? [8]

5. Single-Link is an agglomerative hierarchical clustering approach that repeatedly merges the two closest datum points until the desired number of clusters is obtained. Merging does not perform any consolidation. Consider the following data point in 2D space P0=(1,11), P1=(3,19), P2=(3,15), P3=(5,12), P4=(6,6), P5=(6,16), P6=(7,13), P7=(7,16), P8=(9,5), P9=(12,1), P10=(14,9), P11=(16,19), P12=(17,13), P13=(4,22), P14=(19,4), P15=(19,16), P16=(19,20), P17=(20,19), P18=(21,11), P19=(24,11). Euclidean distances between every pair of points is given in the below. Apply Single-Link to discover four clusters in this data. [6]



| | $P_0$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $P_6$ | $P_7$ | $P_8$ | $P_9$ | $P_{10}$ | $P_{11}$ | $P_{12}$ | $P_{13}$ | $P_{14}$ | $P_{15}$ | $P_{16}$ | $P_{17}$ | $P_{18}$ | $P_{19}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_0$ | 0.00 | 8.25 | 4.47 | 4.12 | 7.07 | 7.07 | 6.32 | 7.81 | 10.00 | 14.87 | 13.15 | 17.00 | 16.12 | 11.40 | 19.31 | 18.68 | 20.12 | 20.62 | 20.00 | 23.00 |
| $P_1$ | 8.25 | 0.00 | 4.00 | 7.28 | 13.34 | 4.24 | 7.21 | 5.00 | 15.23 | 20.12 | 14.87 | 13.00 | 15.23 | 3.16 | 21.93 | 16.28 | 16.03 | 17.00 | 19.70 | 22.47 |
| $P_2$ | 4.47 | 4.00 | 0.00 | 3.61 | 9.49 | 3.16 | 4.47 | 4.12 | 11.66 | 16.64 | 12.53 | 13.60 | 14.14 | 7.07 | 19.42 | 16.03 | 16.76 | 17.46 | 18.44 | 21.38 |
| $P_3$ | 4.12 | 7.28 | 3.61 | 0.00 | 6.08 | 4.12 | 2.24 | 4.47 | 8.06 | 13.04 | 9.49 | 13.04 | 12.04 | 10.05 | 16.12 | 14.56 | 16.12 | 16.55 | 16.03 | 19.03 |
| $P_4$ | 7.07 | 13.34 | 9.49 | 6.08 | 0.00 | 10.00 | 7.07 | 10.05 | 3.16 | 7.81 | 8.54 | 16.40 | 13.04 | 16.12 | 13.15 | 16.40 | 19.10 | 19.10 | 15.81 | 18.68 |
| $P_5$ | 7.07 | 4.24 | 3.16 | 4.12 | 10.00 | 0.00 | 3.16 | 1.00 | 11.40 | 16.16 | 10.63 | 10.44 | 11.40 | 6.32 | 17.69 | 13.00 | 13.60 | 14.32 | 15.81 | 18.68 |
| $P_6$ | 6.32 | 7.21 | 4.47 | 2.24 | 7.07 | 3.16 | 0.00 | 3.00 | 8.25 | 13.00 | 8.06 | 10.82 | 10.00 | 9.49 | 15.00 | 12.37 | 13.89 | 14.32 | 14.14 | 17.12 |
| $P_7$ | 7.81 | 5.00 | 4.12 | 4.47 | 10.05 | 1.00 | 3.00 | 0.00 | 11.18 | 15.81 | 9.90 | 9.49 | 10.44 | 6.71 | 16.97 | 12.00 | 12.65 | 13.34 | 14.87 | 17.72 |
| $P_8$ | 10.00 | 15.23 | 11.66 | 8.06 | 3.16 | 11.40 | 8.25 | 11.18 | 0.00 | 5.00 | 6.40 | 15.65 | 11.31 | 17.72 | 10.05 | 14.87 | 18.03 | 17.80 | 13.42 | 16.16 |
| $P_9$ | 14.87 | 20.12 | 16.64 | 13.04 | 7.81 | 16.16 | 13.00 | 15.81 | 5.00 | 0.00 | 8.25 | 18.44 | 13.00 | 22.47 | 7.62 | 16.55 | 20.25 | 19.70 | 13.45 | 15.62 |
| $P_{10}$ | 13.15 | 14.87 | 12.53 | 9.49 | 8.54 | 10.63 | 8.06 | 9.90 | 6.40 | 8.25 | 0.00 | 10.20 | 5.00 | 16.40 | 7.07 | 8.60 | 12.08 | 11.66 | 7.28 | 10.20 |
| $P_{11}$ | 17.00 | 13.00 | 13.60 | 13.04 | 16.40 | 10.44 | 10.82 | 9.49 | 15.65 | 18.44 | 10.20 | 0.00 | 6.08 | 12.37 | 15.30 | 4.24 | 3.16 | 4.00 | 9.43 | 11.31 |
| $P_{12}$ | 16.12 | 15.23 | 14.14 | 12.04 | 13.04 | 11.40 | 10.00 | 10.44 | 11.31 | 13.00 | 5.00 | 6.08 | 0.00 | 15.81 | 9.22 | 3.61 | 7.28 | 6.71 | 4.47 | 7.28 |
| $P_{13}$ | 11.40 | 3.16 | 7.07 | 10.05 | 16.12 | 6.32 | 9.49 | 6.71 | 17.72 | 22.47 | 16.40 | 12.37 | 15.81 | 0.00 | 23.43 | 16.16 | 15.13 | 16.28 | 20.25 | 22.83 |
| $P_{14}$ | 19.31 | 21.93 | 19.42 | 16.12 | 13.15 | 17.69 | 15.00 | 16.97 | 10.05 | 7.62 | 7.07 | 15.30 | 9.22 | 23.43 | 0.00 | 12.00 | 16.00 | 15.03 | 7.28 | 8.60 |
| $P_{15}$ | 18.68 | 16.28 | 16.03 | 14.56 | 16.40 | 13.00 | 12.37 | 12.00 | 14.87 | 16.55 | 8.60 | 4.24 | 3.61 | 16.16 | 12.00 | 0.00 | 4.00 | 3.16 | 5.39 | 7.07 |
| $P_{16}$ | 20.12 | 16.03 | 16.76 | 16.12 | 19.10 | 13.60 | 13.89 | 12.65 | 18.03 | 20.25 | 12.08 | 3.16 | 7.28 | 15.13 | 16.00 | 4.00 | 0.00 | 1.41 | 9.22 | 10.30 |
| $P_{17}$ | 20.62 | 17.00 | 17.46 | 16.55 | 19.10 | 14.32 | 14.32 | 13.34 | 17.80 | 19.70 | 11.66 | 4.00 | 6.71 | 16.28 | 15.03 | 3.16 | 1.41 | 0.00 | 8.06 | 8.94 |
| $P_{18}$ | 20.00 | 19.70 | 18.44 | 16.03 | 15.81 | 15.81 | 14.14 | 14.87 | 13.42 | 13.45 | 7.28 | 9.43 | 4.47 | 20.25 | 7.28 | 5.39 | 9.22 | 8.06 | 0.00 | 3.00 |
| $P_{19}$ | 23.00 | 22.47 | 21.38 | 19.03 | 18.68 | 18.68 | 17.12 | 17.72 | 16.16 | 15.62 | 10.20 | 11.31 | 7.28 | 22.83 | 8.60 | 7.07 | 10.30 | 8.94 | 3.00 | 0.00 |

6. Consider the following data points and mark outliers using a Gaussian distribution based approach.

   103, 31, 106, 91, 102, 104, 105, 101, 104, 50

   Report a value above 110 that can be marked as an outlier with 95% confidence. [6]

7. Recall the SVM classifier as discussed in the class. Apply the same to get the equation of the linear classifier for the data points specified below.

| | |
|---|---|
| Positive Poins | (1,7), (3,7), (3,9), (5,5) |
| Negative Points | (6,2), (7,3), (8,2), (9,4) |

   Provide weights for the classifier. (Hint: Use geometry) [6]