No of Pages : 3 Course Code : 18XT83

Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE - 641 004

SEMESTER EXAMINATIONS, APRIL 2023

MSc - THEORETICAL COMPUTER SCIENCE Semester : 8

18XT83 DATA MINING

Time : 3 Hours **Maximum Marks :100**

INSTRUCTIONS:

1. Answer **ALL** questions. Each question carries 20 Marks.
2. Subdivision (a) carries 3 marks each, subdivision (b) carries 7 marks each and subdivision (c) carries 10 marks each.
3. Course Outcome : Table

Qn.1	CO1	Qn.2	CO2	Qn.3	CO3	Qn.4	CO4	Qn.5	CO5

1. a) For each of the following data sets, identify whether or not data privacy is an important issue. Justify your answer.
   - Census data collected from 1900 to 1950.
   - IP addresses and visit times of Web users who visit your Website.
   - Images from Earth-orbiting satellites. [3]

   b) i) Consider the following *person* table.

Name	MaritalStatus
John	Married
Peter	Single
Margret	Divorced
Robert	Married
Kingsly	Single
Reena	Widow

   Convert the attribute "MaritalStatus" to binary type by replacing it with suitable new attributes and rewrite the above table. [3]

   ii) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8)
   - Compute the Euclidean distance between the two objects.
   - Compute the Manhattan distance between the two objects
   - Compute the Minkowski distance between the two objects, using h = 3. [4]

   c) Explain the computation procedure of finding dissimilarity between objects described by nominal attributes and asymmetric binary attributes.

2. a) An educational psychologist wants to use association analysis to analyze test results. The test contains 10 multiple choice questions with four possible answers each. How would you convert this data into a form suitable for association analysis? (Write sample transaction database with 5 tuples). [3]

   b) i) Consider the following 1-item candidate set

Item	Count
Cake	3
Bread	4
Cola	2
Coffee	4
Milk	4
Eggs	1

Find the total number of 2-item candidate sets and 3-item candidate sets without any minimum support (Called as brute force strategy). Again find the total number of 2-item candidate sets and 3-item candidate sets with minimum support 3 (Apriori Pruning). Find the percentage of reduction in the apriori pruning strategy when compared with brute force strategy. [3]

ii) Draw FP tree for the following transaction table

TID	items_bought
T100	{ I6, I1, I3}
T200	{ I1, I2, I4, I5, I3}
T300	{ I3, I2, I5}
T400	{ I6, I7}
T500	{ I1, I3, I2, I4, I5}
T600	{ I1, I3, I6}
T700	{ I1, I2, I5, I7}
T800	{ I2, I3, I5, I1}
T900	{ I1, I6}
T1000	{ I1, I2, I5 }

[4]

c) Consider the following transaction database with min sup = 50%

TransID	Items
T100	A, B, C, D
T200	A, B, C, E
T300	A, B, E, F, H
T400	A, C, H

- Convert this into vertical format.
- Using ECLAT algorithm list all frequent itemsets together with their support in percentage.

3. a) What are the measures used for data quality? Outline each with an example. [3]

b) i) Suppose that a data warehouse for Big-University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. Draw a snowflake schema diagram for the data warehouse. [3]

ii) What are the common strategies for dealing with missing data? [4]

c) Given the following data (in increasing order) for the attribute age:

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40,45, 46, 52, 70.

- Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.
- How might you determine outliers in the data?
- What other methods are there for data smoothing?

4. a) Give any two examples of sequence data. [3]

b) i) What are the challenges in sequential pattern mining? [3]

ii) Outline any two methods for stream data processing. [4]

c) Explain the Generalized Sequential Pattern (GSP) mining algorithm with an example.

5. a) Outline the applications of data mining in Financial and Retail domains. [3]

b) i) What is bootstrap method for evaluating classifier accuracy? In 0.632 bootstrap method, what is the meaning of 0.632? [3]

ii) Suppose that we would like to select between two prediction models, M1 and M2. We have performed 10 rounds of 10-fold cross validation on each model, where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5 and 26.0. The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2 and 35.0. Comment on whether one model is significantly better than the other considering a significance level of 1%. Use the following t statistical table.

df	0.1	0.05	0.025	0.02	0.01	0.005
1	3.078	6.314	12.706	15.895	31.821	63.657
2	1.886	2.920	4.303	4.849	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.999	3.747	4.604
5	1.476	2.015	2.571	2.757	3.365	4.032
6	1.440	1.943	2.447	2.612	3.143	3.707
7	1.415	1.895	2.365	2.517	2.998	3.499
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
10	1.372	1.812	2.228	2.359	2.764	3.169

[4]

c) Consider the following data. Build a decision tree using ID3 algorithm using Profit as classification label.

AGE	COMPETITION	TYPE	PROFIT
old	yes	swr	down
old	no	swr	down
old	no	hwr	down
mid	yes	swr	down
mid	yes	hwr	down
mid	no	hwr	up
mid	no	swr	up
new	yes	swr	up
new	no	hwr	up
new	no	swr	up

/END/

FD/JU