Question Paper

Exam Date & Time: 03-Jul-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

INFORMATION & COMMUNICATION TECHNOLOGY DEPARTMENT
VI Semester Makeup Examination June-July 2023

DATA WAREHOUSING AND DATA MINING [ICT 3253]

Marks: 50 **Duration: 180 mins.**

Descriptive Questions

Answer all the questions. Section Duration: 180 mins

Missing data, if any, may be suitably assumed

1) A) Consider a training set that contains 100 positives and 200 negative examples. Compare Rule-1 Covering 5 positive and 2 negative examples, and Rule-2 covering 12 positive and 28 negative examples by using the following metrics: (5)

i. Coverage

ii. Likelihood Ratio

iii. Foil Gain

B) Consider a transactional dataset in which item "pencil" occurs in 3000 transactions, item "eraser" occurs in 2000 transactions. Both the items together- are present in 1700 transactions, and both are not purchased in 700 transactions. Determine the correlation measures (i) lift (ii) all-confidence (iii) cosine. Which of these measures are not null-invariant? Justify. (3)

C) Identify whether the following two scenarios are supervised or unsupervised learning. Justify. (2)

(i) A student gets admission into a new school and meets other students of his class who all are unknown to him. A task assigned to him by the teacher is to classify his classmates based on native and hobbies.

(ii) A student learns about four different objects by his teacher who shows the picture of each object and explains the features of each object. Next day, the teacher shows picture of an object and asks the student to the identify the object and put it into the right cluster.

2) A) Apply K-medoids clustering algorithm for the data given below by using Euclidean distance measure: (5)

i. Initial seed points : X1 and X4

ii. Check whether changing the initial seeds to X1 and X5 would result in better clustering. Justify.

iii. K-medoid does not scale well. Identify one algorithm which solves this problem. Justify

	x	y
X1	2	10
X2	2	5
X3	8	4

X4	9	4
X5	5	8
X6	1	2
X7	4	9

B) Consider the sample data representing the attribute age: 12, 14, 15, 15, 18, 19, 19, 20, 21, 21, 24, 24, 24, 24, 29, 32, 32, 34, 34, 34, 34, 35, 39, 44, 45, 51, 69 (3)

i. Plot an equal width histogram by considering 3 bins

ii. Plot an equal frequency histogram by considering 3 bins

iii. Sketch examples of each of the following sampling techniques: SRSWOR, SRSWR, and stratified sampling with samples of size 5. Given:- the strata:("youth" : age< 25), (middle-age: age>24 AND age< 50), (senior: age>49)

C) A database contains 80 records on a particular topic. A search was conducted on that topic, and 60 records were retrieved. Out of the 60 records that were retrieved, 40 were found relevant. Calculate the precision and recall scores for the search. (2)

3) Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. (5)

A)

(i) Draw a star schema diagram for the data warehouse.

(ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2020?

B) Obtain the dissimilarity matrix for symmetric, asymmetric and categorical values given in the below table. Consider T=0,F=1, Non_transparent =1 , Transparent = 0, Full=1, Semi=0 (3)

Obj ID	A1(categorical)	A2(Symmetric)	A3(Asymmetric)
1	Full	T	Transparent
2	Full	F	Non_Transparent
3	Semi	F	Non_Transparent
4	Full	T	Non_Transparent
5	Semi	T	Transparent

C) Differentiate between directly density-reachable and density reachable by giving an example with respect to DB Scan clustering method (2)

4) Determine the frequent itemsets by using Pincer-Search algorithm for the data given below by assuming minimum support threshold as 2: (5)

A)

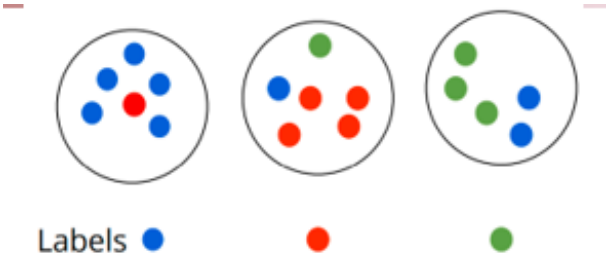
ID	Items
10	1, 3, 4
20	2, 3, 5
30	1, 2, 3, 5
40	2, 5
50	1, 3, 5

.

B) Find the root node which acts as a best split to construct a decision tree by using information gain as a measure for the following data: (3)

Patient ID	Sore throat	Fever	Swollen glands	Headache	**Target Class:** Diagnosis
1	Yes	Yes	Yes	Yes	Allergy
2	No	No	No	Yes	Allergy
3	Yes	Yes	No	No	Cold
4	Yes	No	No	No	Allergy
5	No	Yes	No	No	Cold
6	No	No	No	No	Cold
7	No	No	Yes	No	Cold
8	Yes	No	No	Yes	Allergy
9	No	Yes	No	Yes	Cold
10	Yes	Yes	No	Yes	Cold

C) Find the purity of the 3 clusters and the purity of overall clustering based on the clustered data shown below. Is " purity" a good measure to evaluate clusters? Justify. (2)



5) Determine the frequent itemsets by using Apriori algorithm for the data given below by assuming minimum support threshold as 2: (5)

A)

ID	Rice	Pulse	Oil	Milk	Apple
1	1	1	1	0	0
2	0	1	1	1	0
3	0	0	0	1	1
4	1	1	0	1	0
5	1	1	1	0	1
6	1	1	1	1	1

B) Identify any 3 challenges of outlier detection. (3)

C) During the web server log preprocessing, path completion fills in page references that are missing due to browser and proxy server caching. The problem is to identify important accesses that are not recorded in the access log. Propose a solution to solve this problem (2)

-----End-----