Question Paper

Exam Date & Time: 25-May-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

INFORMATION & COMMUNICATION TECHNOLOGY DEPARTMENT

DATA WAREHOUSING AND DATA MINING [ICT 3253]

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

Descriptive Questions

Answer all the questions.

Show all the steps while tracing algorithms / solving problems

1) Identify the root node for the data given in the following table to construct a binary tree by using Gini index measure. (5)

A)

Own home	Loan	Startup	Employed	Credit rating	Class
Yes	Yes	Yes	No	Good	B
No	No	No	Yes	Good	A
Yes	Yes	No	Yes	Poor	B
Yes	No	Yes	No	Poor	B
No	Yes	No	Yes	Poor	B
No	No	No	Yes	Poor	A
No	Yes	Yes	No	Poor	B
Yes	No	No	Yes	Good	A
No	Yes	No	Yes	Good	A
Yes	Yes	No	Yes	Good	B

B) Consider the stock prices of 2 company observed at 5 different time points given in the following table: (3)

Time	company-1	company-2
1	6	20
2	5	10
3	4	14
4	3	6
5	2	5

Apply covariance analysis to see whether their prices rise or fall together, if the stocks are affected by the same industry trends.

C) Use the quantile-quantile plot to compare the performance of 2 teams by considering their IPL scores during first 10 overs for 6 matches as given below: (2)

team-1: 40, 32, 38, 26, 35, 44

team-2: 33, 41, 25, 46, 23, 38

2) Find the frequent items for the transactional data given in the following table by using dynamic (5)

A) itemset counting (DIC) algorithm with minimum support of 2. Consider a stop after reading 2 transactions.

ID	Items
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 5

B) Construct a bit-map index for the data given in the following table: (3)

Gender	Marital Status	Children (Y/N)	Income (INR)	Home Owner
M	Married	N	120000	N
M	Single	N	80000	Y
F	Divorced	Y	75000	N
F	Married	Y	70000	Y
F	Married	Y	120000	Y
M	Single	Y	80000	N
F	Married	N	70000	N

C) Searching for only interesting patterns is an optimization problem which is desirable but challenging. Can a data mining system find only the interesting patterns? What could be the two possible approaches? (2)

3) Consider the tuples sorted by decreasing probability score obtained by a probabilistic classifier model. Determine the true positive, false positive, true negative, false negative, true positive rate, false positive rate by considering each tuple based on receiver operating characteristics (ROC) curves. Draw a neat ROC. (5)

A)

Tuple	Class	probability
1	Y	0.89
2	N	0.84
3	Y	0.73
4	N	0.60
5	N	0.57
6	Y	0.54
7	Y	0.52
8	Y	0.51
9	N	0.46
10	Y	0.40

B) Determine the frequent itemsets using vertical data format method for the following database. Assume minimum support as 2. (3)

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

50	1,5
60	2,4,5
70	1,3,4,5
80	1,4,5

C) The metadata is the data defining warehouse objects when used in a data warehouse. Write any four contents that a metadata repository should contain as a part of a data warehouse architecture. (2)

4) Apply agglomerative single linkage clustering on the dataset given below and find the clusters. Use Euclidean distance to compute the distance matrix. (5)

A)

	X	Y
P1	0.40	0.53
P2	0.23	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.45	0.30

B) Identify whether INR. 55 and INR. 99 are outliers using nonparametric method for the following data. Justify. (3)

Daily Wages in INR	Number of Workers
30-40	10
40-50	20
50-60	35
60-70	14
70-80	8
80-90	6

C) Consider a cube with 3 dimensions. Assume no hierarchy levels for dimension 1 & dimension 2, 3 levels of hierarchy for dimension 3. Find the total number of cuboids required to represent this cube as a lattice of cuboid. Justify. (2)

5) Find the frequent items (minimum support=2) and the interesting association rule (minimum confidence=50%) by using the frequent pattern (FP) growth algorithm for the transactional data given in the following table: (5)

A)

ID	items
1	A, B
2	B, D
3	B
4	A, B, D
5	A, C
6	A, C
7	B, C
8	A, B, C, E
9	A, B, C

.

B) Determine the Attribute-Value, Classlabel (AVC) for the tuple data given in the following table: (3)

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

C) Session identification takes all of the page references for a given user in a log and breaks them up into user sessions. For logs that span long periods of time, it is very likely that users will visit the website more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. Propose a simplest method which could solve this problem. (2)

-----End-----