Question Paper

Exam Date & Time: 04-Jan-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

FIFTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, DEC 2022 / JAN 2023

DATA MINING AND PREDICTIVE ANAYSIS [ICT 3171]

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

A

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) Categorize and brief about various techniques for assessing accuracy of classification models. (5)

A)

B) Point out pros and cons of support vector machines. (3)

C) Summarize the following with respect to Linear Regression. (2)

i) Types of Linear Regression ii) Types of relationship shown by regression line

2) Perform 2-medoids clustering using Manhattan distance for the dataset given below for one iteration and compute the total cost. x1(2,6), x2(3,4), x3(3,8), x4(4,7), x5(6,2), x6(6,4), x7(7,3), x8(7,4), x9(8,5), x10(7,6). Consider x2 and x8 as initial medoids of cluster1 and cluster2 respectively. Do you suggest swapping of x8 with x7? Justify. (5)

A)

B) Explain the concept of tree pruning. Summarize and illustrate the two common approaches to tree pruning. (3)

C) Draw basic level PC tree for the dataset given in Table Q2C (2)

Table Q2C

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}

3) Consider the dataset given in table Q.3A. Find the root node and first level split using information gain. (5)

A) **Table Q.3A**

Day	Outlook	Temperature	Humidity	Wind	Play cricket
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes

8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

B) Given Fig Q.4B, find the page rank for each node considering damping factor as 0.7. (3)

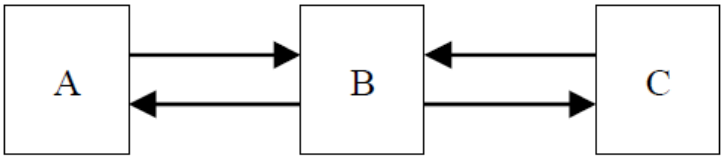

Fig Q.4B

C) Compare and contrast CLARA and CLARANS algorithm. (2)

4) Consider the dataset given in Table Q.4A and find all frequent patterns satisfying a support threshold of 50% and confidence of 60% using Apriori algorithm. Also, find the strong association rules. (5)

A)

Table Q.4A

Transaction	List of items
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

B) Table Q.4B shows ticket prices (in $) for Padman and Black Panther movies respectively in Big Cinemas, Manipal. Find the covariance between the two movies and also state the type of covariance between the two movies (3)

Table Q.4B

Days of the week	Padman	Black Panther
1	7	21
2	6	11
3	5	15
4	4	6
5	3	6

C) Give the five-number summary of the below data and draw the box plot. (2)

5 7 1 9 11 22 15

5) List and summarize the various methods to improve efficiency of Apriori algorithm. (5)

A)

B) Discuss Web Usage Mining. Briefly describe two main approaches in Web Usage Mining (3)

C) Identify whether the following task requires data mining or not (2)

i) By looking at a CT scan, a doctor wants to classify if a patient id covid +ve or not. He uses many labeled CT scans for making the decision.

ii) Monitoring heart rate of a patient for abnormalities

iii) Predicting the outcome of tossing a fair pair of dice

iv) Extracting the frequencies of a sound wave

-----End-----