Question Paper

Exam Date & Time: 09-Jan-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

FIFTH SEMESTER B.TECH (CCE) MAKEUP EXAMINATIONS, JANUARY 2024

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) A) Calculate dissimilarity between a nominal attribute in the following data set in Table Q1A. Also convert these into Symmetric binary attributes and then compute dissimilarity using suitable technique. (5)

Table Q1A

Object id	Value
1	Class A
2	Class B
3	Class B
4	Class A
5	Class C

B) Find 3 cluster centers after one iteration using K-Means for following one dimensional data. (3)

12,14,16, 18,20,25,27,28,35,39,45,49,55,65.

Consider initial centers as 14,27,49 respectively.

C) Identify and explain an application example for clustering technique of data mining (2)

2) A) Construct a PC-tree for the following transactional data set: T1: {1,2,4,5}, T2:{1,2,3,4,5}, T3:{2,3,1,5}, T4:{2,1,4}, T5:{4}, T6:{2,4}, T7:{1,4,5}, T8:{3,2}. Write all the transactions obtained by the PC tree. Construct FP tree based on the transactions obtained from the PC tree. Assume the minimum support count as 4. (5)

B) Find the total cost by applying K-medoids clustering technique to the following two dimensional data with k=2. Choose suitable medoids by manually looking at the 2D plot of the points. (3)

{(1,1),(1,0),(0,2),(2,4),(3,5)(5,3)(2,3)}

C) It is required to divide a two dimensional data set with few outliers into 2 clusters. Which partition based clustering method studied would be suitable for this case? Justify your answer. (2)

3) A) Suppose a teacher needs to divide her students into groups based on the marks scored by each student in an assignment. The teacher wants to use hierarchical clustering to segment the students into different groups, but since there is no fixed target for the number of groups to have, it cannot be solved as a supervised learning problem. Hierarchical clustering will help the teacher to segment the students into different groups based on their performance. Given a sample student list, calculate the distance using the Euclidean distance formula. (5)

Table Q3A

Student ID	Marks
1	10
2	7
3	28
4	20
5	35

B) (3)

Consider the data given in the following table with target variable "Covid". We need to decide whether a person (Temperature = Normal, Cough = Yes, Headache = Yes, Fever = No) must be quarantined or not according to the Covid results (Positive / Negative) by applying Bayes' theorem.

Table Q3B

Temperature	Cough	Headache	Fever	Covid
High	No	Yes	Yes	Negative
High	Yes	No	No	Positive
High	No	Yes	Yes	Positive
Normal	Yes	Yes	Yes	Positive
Normal	No	No	No	Negative
Normal	Yes	Yes	Yes	Positive
Normal	Yes	Yes	No	Negative
High	Yes	Yes	Yes	Positive

C) Why do we use regression analysis? Consider a simple linear regression equation for 2 sets of data:y=1.5+0.95x, and predict a missing value given x=0.56 (2)

4) Using Gain ratio find out root node of the decision tree for the following data set. (5)

Table Q4A

A) Table Q4A

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

B) Consider a model that predicted 106 cases as benign that were actually benign (true positive), and correctly identified 61 cases as malignant (true negative). It incorrectly predicted 20 cases as benign that were actually malignant and identified 25 cases as malignant that were actually benign. Evaluate the performance of this model based on the sensitivity, specificity, accuracy, misclassification rate. (3)

C) Consider the sequences of transactions for 5 customers Find whether the following subsequences are frequent or not by considering minimum support count as 2. Write the sequence number that supports these sub sequences. (2)

CUSTOMER SEQUENCE:

1 - C, J, J

2 - (AB), (DFG)

3 - C, (CEG), (CD)

4 - C, (DG), (EJ)

5 - J, (AB)

Subsequence to check : C -> J, AD, DG, E->D

5) Consider a positive set of data points {(4,4), (4, -4), (-4,-4), (-4, 4), (5,4), (5, -4)} and negative set of data point {(2,2), (2,-2), (-2 ,-2), (-2,2), (-3, 2), (-3, -2)}. Divide the data sets into 2 classes by using non-linear support vector machine. Assume (2,0), (4,1), (4,-1) as support vectors. (5)

A)

B) Consider a population with a total of 30 individuals, which is being used to create a dataset that predicts whether a person will go to the gym or not. Out of these 30 individuals, 16 people go to the gym while 14 do not. The task is to calculate the entropy and information gain for this dataset. (3)

Table Q5B

	will go to Gym	Will not go to gym	Total
No Motivation	7	1	8

Neutral	4	6	10
Highly motivated	5	7	12
Total	16	14	30

C) Calculate the split info and gain ratio of Income using the provided data. In Table Q4A (Consider Gain (Income) as 0.048 bits) (2)

-----End-----