[This question paper contains 8 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 9220 **J**

Unique Paper Code : 2343012005

Name of the Paper : Data Mining – I

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : IV

Duration : 3 Hours Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. The paper has **two** sections. **Section A** is compulsory.
3. Attempt any **four** questions from **Section B.**
4. Parts of the question must be answered together.
5. The use of simple calculator is allowed.

Section A

1 (a) Give an application where spatio-temporal data is used. (2)

(b) Why is the decision tree classifier an early learner? (2)

(c) Identify whether or not the given activities are data mining tasks. Justify your choice.

(i) Extracting the frequencies of a sound wave

(ii) Monitoring the heart rate of a patient for abnormalities

(iii) Dividing the customers of a company according to their gender. (3)

P.T.O.

(d) Given five objects in the dataset (01, 02, 03, 04 & 05), mention all train and test set distributions for performing leave one out cross-validation. (3)

(e) Is clustering considered a supervised learning task or an unsupervised learning task? Justify your answer. (3)

(f) In Naïve Bayes classifier, if certain combination of attributes and class labels are never observed, it results in a zero conditional probability. What problems may arise in this scenario? Which alternate estimates can be used in such a scenario? (4)

(g) Consider the attribute Website-Accessed with the following values :

{Amazon, Myntra, Flipkart, Ajio, Nykaa}

Convert this attribute to asymmetric binary attribute. (4)

(h) Define the splitting method Gain Ratio used by the decision tree classifier. How is it used to compensate for attributes that produce a large number of child nodes? (4)

(i) Given a two-class problem with class labels as "A" or "B", draw the confusion matrix and compute error rate, precision and specificity. (5)

Instance Id	Predicted Class	Actual Class
1	A	A
2	A	A
3	A	B
4	B	B
5	A	A
6	B	A
7	B	B
8	B	B
9	B	A
10	B	A

Section B

2. (a) Given the following training dataset, compute all class conditional and prior probabilities. Use the Naïve Bayes approach to predict the class label for the test instance : (12)

Gives Birth = Yes, Can Fly = No, Lives in Water = Yes, Have Legs = No

Name	Gives Birth	Can Fly	Live in Water	Have Legs	Class
Human	Yes	No	No	Yes	Mammal
Salmon	No	No	Yes	No	Non-Mammal
Whale	Yes	No	Yes	No	Mammal
Frog	No	No	Sometimes	Yes	Non-Mammal
Komodo	No	No	No	Yes	Non-Mammal
Bat	Yes	Yes	No	Yes	Mammal
Pigeon	No	Yes	No	Yes	Non-Mammal
Cat	Yes	No	No	Yes	Mammal
Leopard Shark	Yes	No	Yes	No	Non-Mammal
Turtle	No	No	Sometimes	Yes	Non-Mammal
Porcupine	Yes	No	No	Yes	Mammal
Gila Monster	No	No	No	Yes	Non-Mammal
Platypus	No	No	No	Yes	Mammal
Owl	No	Yes	No	Yes	Non-Mammal
Dolphin	Yes	No	Yes	No	Mammal

(b) Differentiate between noise and outliers with one suitable example for each. (3)

3. Consider the following market basket transactions :

TID	List of Items
T1	{a, b, e}
T2	{b, d}
T3	{b, c}
T4	{a, b, d}
T5	{a, c}
T6	{b, c}
T7	{a, c}
T8	{a, b, c, e}
T9	{a, b, c}

(a) Apply the Apriori algorithm on the given transactional dataset to compute the candidate and frequent itemsets for each dataset scan. Assume a support threshold of 25%. (8)

(b) Enumerate all association rules generated from the largest frequent itemsets. Compute the confidence of each generated rule. Assuming that the minimum confidence threshold is 50%, find all the strong association rules. (7)

4. Consider the following dataset :

Weekend	Weather	Parents	Income	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Shopping
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis
W11	Rainy	Yes	Poor	Tennis
W12	Sunny	No	Rich	Shopping

(a) Compute the Gini Index of Weather, Parents, arid Income attributes. Given that you construct a decision tree using the Gini Index as the splitting criteria, which of these three attributes would you choose at the root? Justify your choice. (12)

(b) Compute the Gini Index of 'Weekend' attribute. Is it a good splitting attribute for constructing a decision tree? Justify your answer. (3)

5. (a) Why is it important to standardize data before performing distance- based computations?

Given the dataset, show computations for standardizing the values of Height and Income attributes such that the transformed values have a mean of 0 and a standard deviation of 1.

Name	Height (in cms)	Income
Manas	170	40,000
Shreya	160	30,000
Surbhi	180	50,000
Hemant	175	45,000
Paarth	165	55,000

(7)

(b) Why is K-Nearest Neighbor algorithm considered as an instance-based learner? (8)

P.T.O.

Consider the two-dimensional labelled data set with attributes Brightness and Saturation and class label Color given below :

Brightness	Saturation	Color
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

Classify the data point where Brightness = 20 and Saturation = 35 according to the 5-nearest neighbor algorithm. Use the majority voting scheme and Euclidean distance.

6. (a) Given the following sorted market price records, partition them into three bins using : (6)

(i) equal depth partitioning

(ii) equal width partitioning.

5, 10, 11, 14, 16, 34, 50, 55, 77, 93, 201, 215

(b) Consider a sample dataset of employees working in an organization :

Emp. Code	Name	Designation	Date of Joining	Salary (per annum)	Age
E1A	Malini	Director	14-10-1995	50 Lac	50
E6A	Manan	Sr. Manager	16-1-2005	30 Lac	40
E3A	Gautam	Sr. Manager	18-10-2004	28 Lac	54
E7A	Meeta	Manager	4-6-2004	29 Lac	39

(i) Identify the type of attributes as nominal, ordinal, interval, or ratio. Give justification for each. (6)

(ii) What are the advantages of aggregation while data pre-processing for Data Mining? Suggest a method to aggregate Salary attribute. (3)

7. (a) Given the data of nine points (P1 – P9) for K-means clustering. Perform two iterations of the K-means clustering algorithm and show the clusters, their new centroids, and compute SSE after each iteration. Assume that k=3 and initial cluster centroids are P7 (3, 3), P8 (9, 4) and P9 (3, 7) as C1, C2 and C3 respectively. Use Manhattan distance as the distance metric. (12)

Points	X	Y
P1	1	3
P2	2	2
P3	5	8
P4	8	5
P5	3	9
P6	10	7
P7	3	3
P8	9	4
P9	3	7

(b) Explain the following types of clusters by giving suitable examples :

(i) Centre-based clusters

(ii) Density-based clusters (3)

(100)