21

[This question paper contains 8 printed pages.]

Your Roll No..............

Sr. No. of Question Paper : 936 **G**

Unique Paper Code : 2342202302

Name of the Paper : Data Mining – I

Name of the Course : **B.A. Programme**

Semester : III

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. **Section A** is compulsory.

3. Attempt any **four** questions from **Section B**.

4. Parts of a question must be answered together.

5. Use of scientific calculator is allowed.

P.T.O.

Section A

1. (a) Why is k-nearest neighbour algorithm called as a lazy learner? (2)

(b) In a dataset it is found that an itemset {ab} is infrequent. Will the itemset {abc} be infrequent or frequent? Give reason. (2)

(c) Write the formula for calculating Euclidean distance. Calculate the Euclidean distance between the two points P1(3, 4, 2) and P2(4, 7, 4). (3)

(d) Convert the given categorical attribute X with possible values {RED, GREEN, BLUE, YELLOW, PINK, WHITE} into asymmetric binary attributes. (3)

(e) Describe any two feature subset selection approaches. (4)

(f) Determine the type of attribute for the following : (5)

(i) Height of students in a class

(ii) Date of joining of employees

(iii) Color of cars

(iv) Grades of the students

(v) Gender of people

(g) Consider the attribute X with values {2, 2, 3, 2, 4, 3, 2, 5, 4}. Compute mean, median, mode, range and variance. (5)

(h) What is meant by standardization? Write its formula. In a dataset, the mean and standard deviation of the variable 'age' is 22 and 3 respectively. Standardize the value of age = 32.5. (6)

Section B

2. (a) Based on the dataset given below, classify a new instance with x1 = 2 and x2 = 5 as positive or negative using k-Nearest Neighbour technique for k = 5. Use the distance function as Euclidean distance. (10)

x1	x2	Y
2	3	-
5	7	+
6	6	+
1	2	-
7	5	+
1	4	-
4	1	-
7	3	+

(b) Briefly describe the process of knowledge discovery in databases. (5)

3. (a) How does underfitting and overfitting affect a model's generalization capability? Describe one measure to address each. (5)

(b) Using K-Means clustering technique, cluster the given dataset into 2 groups. Compute the new cluster centres after two iterations of K-Means algorithm. Which records will belong to these clusters? Calculate the sum of squared error (SSE) at the end of each iteration. Use (1, 1) and (5, 7 as initial centroids and Euclidean distance as distance metric. (10

Record No.	A	B
R1	1	1
R2	2	2
R3	3	3
R4	5	7
R5	3	5

4. (a) Identify the type of data mining task for each of the tasks given below : (5)

(i) Projecting the sales volumes for the next quarter using historical data.

(ii) Identifying relationships between items that are frequently purchased together in a supermarket.

(iii) Finding different customer groups for targeted marketing strategies.

(iv) Identifying fraudulent credit card transactions.

(v) Estimating amount of rainfall over next few days.

(b) Consider the given dataset for determining whether a person will buy a car or not based on his income, marital status and gender. (10)

(i) Calculate the overall Gini Index.

(ii) Determine the attribute that will be selected as root of the decision tree according to the Gini Index.

P.T.O.

Income (x1)	MaritalStatus (x2)	Gender (x3)	BuyCar (y)
Low	Yes	MALE	No
Medium	Yes	FEMALE	Yes
High	Yes	FEMALE	Yes
Low	No	MALE	No
Medium	No	FEMALE	No
High	No	MALE	Yes
Low	Yes	MALE	Yes
Medium	Yes	FEMALE	No
Medium	No	MALE	Yes
High	No	MALE	No

5. (a) State two assumptions of Naive Bayes classifier. Using Naive Bayes 10 classification algorithm, estimate the class label Y for a new instance with values x1 = BLUE, x2 = LARGE and x3 = FULL for the dataset given below. Calculate all the prior and posterior probabilities. (10)

X1	X2	X3	Y
BLUE	LARGE	FULL	YES
RED	SMALL	HALF	YES
BLUE	SMALL	FULL	NO
RED	LARGE	HALF	NO
RED	LARGE	FULL	YES
BLUE	SMALL	HALF	NO
BLUE	SMALL	HALF	YES
RED	LARGE	FULL	YES

(b) What is dimensionality reduction? Why is it used? (5)

6. (a) Explain the process of rule generation in association rule mining. Consider the given dataset and answer the following : (10)

(i) Compute all frequent item sets of size one and two, considering 1/3 as the minimum support.

(ii) Generate association rules using frequent two item sets with a minimum confidence of 60%.

Transaction	Items
T1	A, B, C
T2	A, B
T3	B, C
T4	A, C
T5	B

(b) Differentiate between random sampling and stratified random sampling. Give one example of each. (5)

7. (a) Briefly describe the potential problems in data collection. (5)

P.T.O.

(b) Differentiate between supervised and unsupervised techniques. Give examples for both. (5)

(c) Consider the given confusion matrix given below and compute the following : (5)

		Actual Values	
		True	False
Predicted Values	True	10	11
	False	4	20

(i) Number of false positives

(ii) Number of true negatives

(iii) Accuracy

(iv) Precision

(v) Recall