[This question paper contains 8 printed pages.]

Your Roll No. 20035570003

Sr. No. of Question Paper : 4711 E

Unique Paper Code : 32347611

Name of the Paper : Data Mining

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

Semester : VI

Duration : 3 Hours Maximum Marks : 75

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. Question No. **1 (Section A)** is compulsory.
3. Attempt any **4** Questions from Nos. **2** to **8 (Section B)**.
4. Parts of a question must be answered together.

Section A

1. (a) Determine the attribute type for the following : (2)

(i) Bronze, Silver, Gold medals awarded at Olympics

(ii) Number of patients in hospital

(iii) Car color

(iv) Dates in a Calendar

(b) List two applications where graph data structure is used to model the data. (2)

(c) Consider an association rule between items from market basket domain which has high support and high confidence. What does it signify? (2)

(d) Explain the following terms with respect to a density-based clustering algorithm: Core point and Border point (2)

(e) Consider a categorical attribute with five values {awful, poor, OK, good, great}. Covert this attribute to asymmetric binary attributes. (3)

(f) List any two ways in which a noise object differs from an outlier? Explain with the help of the example. (4)

(g) Consider the given dataset with two attributes Age and Salary measured on different scales. What problems might arise if the dataset is directly used for k-means clustering? What steps will you suggest to handle the problem? (4)

	Age (in years)	Salary (in rupees)
1	44	72000
2	27	48000
3	30	54000
4	38	61000
5	50	83000
6	37	67000

(h) How is an eager learner classifier different from a lazy learner classifier? Support your answer with an example from both category of classifiers. (4)

(i) Specify whether each of the following activities should fall under the purview of a data mining task or a database query. Justify your answer.

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

(ii) Predicting the future stock price of a company using historical records. (4)

(j) Explain the concept of following types of clustering schemes :

(i) Fuzzy clustering

(ii) Hierarchical based clustering (4)

(k) Consider the following values for two attributes corresponding to four data points: P1(0,2), P2(2,0), P3(3,1), and P4(5,1). Compute the proximity matrix using the metric as Euclidean Distance. (4)

Section B

2. (a) Consider the following dataset for binary classification problem : (6)

Instance	A	B	C	Target Class
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(b) For evaluating the performance of a classifier, how does holdout method ~~and 4~~ differ from k-fold cross validation? For k=5 and datapoints- D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10 in the dataset, mention one possible dataset distribution between training and test partition for k-fold cross-validation. (4)

3. Consider the dataset shown below : (10)

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

(i) Estimate the conditional probabilities for *P(Outlook|Yes)*, *P(Temperature|Yes)*, *P(Humidity|Yes)*, *P(Windy|Yes)*, *P(Outlook|No)*, *P(Temperature|No)*, *P(Humidity|No)*, and *P(Windy|No)*.

(ii) Use these estimate of conditional probabilities to predict the class label (*Play Golf*) for a test sample (*Outlook = Rainy, Temperature = Cool, Humidity = High, Windy = True*) using the naive Bayes approach.

4. The DM Pizza Parlour sells pizzas with optional toppings: pepperoni, pineapple, and pickled-onion. Suppose, you have tried five pizzas (P1 to P5) and kept a record of which you liked :

	Pepperoni	pineapple	pickled-Onion	liked
P1	True	True	True	False
P2	True	False	False	True
P3	False	True	True	False
P4	False	True	False	True
P5	True	False	False	True

Show binarization of the above data and use it to calculate Euclidean distances, to demonstrate how the k-Nearest-Neighbor (k-NN) classifier with majority voting would classify a tuple <*False, True, True*>, for k = 1 and k = 3 respectively. (10)

5. Suppose we have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Use k-Nearest-Neighbor classifier to classify the tuple <161, 64>. Assume

k = 5. Note that the data set should be scaled to range [0-1] prior to classification, using min-max normalization. (10)

Height (in cms)	Weight (in kgs)	T-shirt Size
157	58	S
158	59	S
158	63	M
160	59	M
157	56	S
163	60	M
163	61	M
160	64	M
168	64	L
165	61	L
171	62	L
169	63	L

6. Considerthe following data set with nine transactions. Use Apriori algorithm to compute all frequent itemsets of size one and two, considering 1/3 as the minimum support. Also, generate strong association rules using frequent 2-itemsets, considering 0.65 as the minimum confidence. (10)

TID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

7. Consider the following data set: {4, 8, 12, 20, 32, 36, 48}. Assuming that k = 2, and initial cluster centers for k-means clustering are 32 and 48. Perform the k-means clustering to arrive at final set of cluster solutions. Also, at the end of every iteration, compute the SSE. (10)

8. Use the distance matrix given below to perform hierarchical clustering using single link and show the dendrogram. (10)

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.24	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0