This question paper contains 8 printed pages]

Roll No. □□□□□□□□□□□

S. No. of Question Paper : **861**

Unique Paper Code : **234611** **E**

Name of the Paper : **Data Mining [CSHT-616 (*iv*)]**

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

Semester : **VI**

Duration : **3 Hours** Maximum Marks : **75**

(Write your Roll No. on the top immediately on receipt of this question paper.)

All parts of Question 1 (Part A) are compulsory.

Parts of a question must be answered together.

Attempt any *four* questions from Part B.

All questions in Part B carry equal marks.

Part A

(*All* questions are compulsory) 35 marks

1. (*a*) Give a diagrammatical representation of the steps involved in the knowledge discovery from data. Explain in brief. 4

(*b*) Differentiate between classification and regression analysis. Give an example of a d-dimensional dataset to support the difference. 4

(*c*) Define : 2

(*i*) Closed frequent itemset

(*ii*) Maximal frequent itemset.

(*d*) State the Apriori property. Also state the *two* major drawbacks of the Apriori method. 1+2

(*e*) Mention the strategy adopted by the FP growth method. 2

(*f*) Draw a confusion matrix for a binary classification problem. Write down the formula for : 2+4

(*i*) Sensitivity

(*ii*) Specificity

(*iii*) False positive rate

(*iv*) False negative rate.

(*g*) Give an example for the test condition of a binary split and a multiway split for : 2+2

(*i*) Nominal attributes

(*ii*) Continuous attributes.

(*h*) Explain the holdout method for evaluating a classifier. How is two-fold cross-validation different from the holdout method ? 2+2 (*i*) Mention the difference between single and complete linkage of hierarchical clustering. Illustrate with an example. 2

(*j*) Describe the following clustering algorithm in terms of : 2+2

(*i*) Shape of clusters

(*ii*) Limitations

(I) k-means

(II) DBSCAN.

Part B

Attempt any *four* questions from this part.

All questions carry equal marks.

2. (*a*) Describe *two* objective measures of interestingness for association analysis. 4

(*b*) Suppose data for analysis include attribute age whose 20 values in increasing order are : 3+3

13, 15, 16, 16, 20, 20, 20, 20, 22, 22, 25,

25, 25, 25, 25, 30, 33, 35, 40, 45.

(*i*) Give *five* number summary of data.

(*ii*) Show a boxplot of the data.

3. (*a*) Consider a transaction database with two transactions : 4

(*a*1, *a*2, ..., *a*100) and (*a*1, *a*2, ..., *a*50)

Let the minimum support threshold be 1.

(*i*) Find two closed frequent itemsets and their support counts.

(*ii*) One maximal frequent itemset.

(*b*) 3+3

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

For the transaction dataset given above :

(*i*) Generate the complete FP-tree.

(*ii*) Mine the conditional FP-tree for item I3. Given minimum support count is 2.

4. Find all frequent item sets in the following transactional database using Apriori (minimum support 40%). Also, mention steps used in each pass. Derive the association rules having 100% confidence : 6+4

TID	A	B	C	D	E
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

5. (*a*) Write the algorithm for the *k* nearest neighbor algorithm. Why is this algorithm known as a lazy learner ? 1+3

(*b*)

Instance	a1	a2	a3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	–
4	F	F	4.0	+
5	F	T	7.0	–
6	F	T	3.0	–
7	F	F	8.0	–
8	T	F	7.0	+
9	F	T	5.0	–

Consider the training examples shown in the above table for a binary classification problem : 2+4

(*i*) Calculate the overall Gini index.

(*ii*) What is the best split (between a1 and a2) according to the Gini index ?

6. (*a*) Consider the data set shown below : 3+5

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	–
3	0	1	1	–
4	0	1	1	–
5	0	0	1	+
6	1	0	1	+
7	1	0	1	–
8	1	0	1	–
9	1	1	1	+
10	1	0	1	+

(*i*) Estimate the conditional probabilities for P(A = 1|+), P(B = 1|+), P(C = 1|+), P(A = 1|–), P(B = 1|–), P(C = 1|–), P(A = 0|+), P(B = 0|+), P(C = 0|+), P(A = 0|–), P(B = 0|–), and P(C = 0|–),

(*ii*) Use the estimate of conditional probabilities given in part (*i*) to predict the class label for a test sample (A = 0, B = 1, C = 0) using the naive Bayes approach.

(*b*) What is a core object in the DBSCAN algorithm for clustering. 2

7. (*a*) Use the similarity matrix in the given table to perform complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. 5

Data	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

(*b*) For the given data, compute two clusters using k-means algorithm for clustering where initial cluster centers are (1.0, 1.0) and (5.0, 7.0). Execute for two interations : 5

Record number	**A**	**B**
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5