U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

September / October 2023 Supplementary Examinations

Programme: B.E. **Semester: V**
Branch: Information Science and Engineering **Duration: 3 hrs.**
Course Code: 20IS5PEDMG **Max Marks: 100**
Course: Data Mining **Date: 19.09.2023**

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

SUPPLEMENTARY EXAMS 2023

UNIT - I

1 a) Enumerate on aggregation process and feature subset selection in data preprocessing for following scenario: **06**
i. Detection of fraud in credit card transactions.
ii. Grouping of related customers.

b) How to perform feature subset selection process in data preprocessing? **08**

c) Compute the hamming distance and the jaccard similarity between the following two binary vectors **06**
X = 0101010001
Y = 0100011000
Which approach is more similar to the simple matching coefficient and which approach is more similar to the cosine measure?

UNIT - II

2 a) Give the general approach to solve a classification problem **06**

b) Consider the training examples shown in Table below for a binary classification problem. **08**
i. What is the entropy of this collection of training examples with respect to the positive class?
ii. What are the information gains of a1 and a2 relative to these training examples?
iii. What is the best split (between a1 and a2) according to the Gini index?

Instance	a_1	a_2	a_3	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	−

c) Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules, **06**

R1 A .-----*+ (covers 4 positive and I negative examples),
R2: B .-----*+ (covers 30 positive and 10 negative examples),
R3: C ------*+ (covers 100 positive and 90 negative examples),
Determine which is the best and worst candidate rule according to:

i. Rule accuracy.
ii. FOIL's information gain.

OR

3 a) What are the characteristics of rule based classifier? **06**

b) Consider the data set below to compute the following **08**

i. Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+),P(A| −), P(B| −), and P(C| −).

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

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	+

c) Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules, **06**

R1 A .-----*+ (covers 4 positive and I negative examples),
R2: B .-----*+ (covers 30 positive and 10 negative examples),
R3: C ------*+ (covers 100 positive and 90 negative examples),
Determine which is the best and worst candidate rule according to:

i. The Laplace measure
ii. The m-estimate measure (with k=2 and p_+=0.2)

UNIT - III

4 a) What are the factors that affect the computational complexity of the apriori algorithm? **06**

b) Generate association rules for the below transactions with support threshold =50%. Also find confidence of the rule generated. **08**

Transactions	Itemsets
T1	I1,I2,I3
T2	I2,I3,I4
T3	I4,I5
T4	I1,I2,I4
T5	I1,I2,I3,I5
T6	I1,I2,I3,I4

SUPPLEMENTARY EXAMS 2023

c) Consider an association rule between items from market basket domain which has high support and high confidence. What does it signify? What is the use of support and confidence? **06**

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

OR

5 a) What are maximal itemsets? Justify their application with respect to frequent itemset generation. **06**

b) The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k + I are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if-any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table – 1 with minsup: 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent. **08**

Table -1

Transaction ID	Items Bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

(i) Draw an itemset lattice representing the data set given in Table - . Label each node in the Iattice with the following letter(s):

- N: If the itemset is not considered to be a candidate itemset by the Apriori. algorithm. The-re are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

- F: If the candidate itemset is found to be frequent by the Apriori algorithm.

- I: If the candidate itemset is found to be infrequent after support counting.

(ii) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

SUPPLEMENTARY EXAMS 2023

c) Construct a FP tree for the given data. Calculate minimum support. **06**

Transaction id	Items
T1	{1, 3, 4}
T2	{2, 3, 5}
T3	{1, 2, 3, 5}
T4	{2, 5}
T5	{1, 2, 3, 5}

UNIT - IV

6 a) Explain the steps for generating agglomerative hierarchical clustering. **06**

b) Compute the entropy and purity for the confusion matrix in Table below. **08**

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total
#1	1	1	0	11	4	676	693
#2	27	89	333	827	253	33	1562
#3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

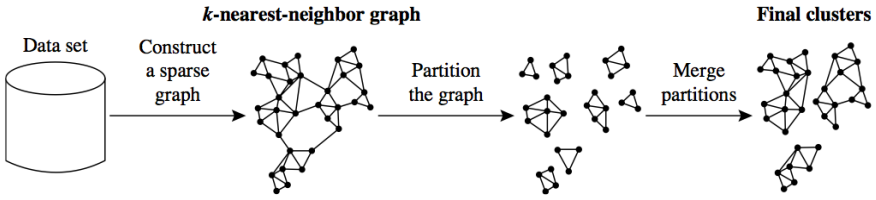
c) Use the similarity matrix in the Table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendogram. The dendogram should clearly show the order in which the points are merged. **06**

	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

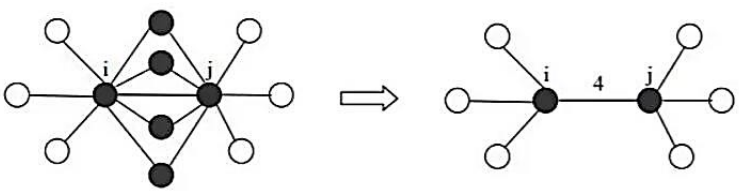
UNIT - V

7 a) Compare K-means and DBSCAN. **06**

b) Explain the overall process involved in the figure below. Which algorithm best suits for this process? **08**



c) How to compute SNN similarity between two points i and j? Explain the steps in detail. **06**



******

SUPPLEMENTARY EXAMS 2023