

M.Sc. Examination, 2019
Semester-IV
Computer Science
Course : MCSO-48
(Data Mining and Warehousing)

Time : 3 Hours

Full Marks : 40

Questions are of value as indicated in the margin

Answer Question No.1 and **any four** from the rest.

1. a) What is the purpose of data preprocessing?
b) Describe the basic notions of classification, training dataset, test dataset, and accuracy to a non-technical friend or relative.
c) Consider an association rule between items from market basket domain which has high support and high confidence. What does it signify?
d) Consider the transactions occurring in the given order:

$\{\{a,b\}, \{b,c,d\}, \{a,b,c\}, \{a,b,c,d\}, \{a,b,c\}\}$

Draw FP tree after each transaction.

2×4=8

2. a) Define entropy and Gini index. How are these related to the impurity of nodes?
b) Consider the training dataset for a binary classification problem:

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	-

- i) What is the best split between A_1 and A_2 according to the Gini index?
ii) What is the entropy of the collection of training examples with respect to the Target Class?
iii) What are the information gains of A_1 and A_2 relative to these training instances?
iv) For A_3 , which is a continuous attribute, compute the information gain for the split ≤ 5 and > 5 .
v) Draw the decision tree based on ID3 algorithm.
3. a) Describe the Bayesian classification technique.
b) By using this classification technique and the dataset given in 2(b), find the outputs for the inputs $(T, F, 6.0)$ and $(F, T, 6.0)$.

2+6=8

4+4=8

P.T.O.

(2)

4. a) Let c_1, c_2 , and c_3 be the confidence values of the rules $\{p\} \rightarrow \{q\}, \{p\} \rightarrow \{q, r\}$, and $\{p, r\} \rightarrow \{q\}$, respectively. If it is assumed that c_1, c_2 , and c_3 have different values, what are the possible relationships that may exist among c_1, c_2 , and c_3 ? Which rule has the lowest confidence?
- b) Apply FP-Growth algorithm to find all infrequent items for the following market-basket transactions, where $\text{min support} = 3$:

Transaction ID	Items brought
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	$\{b, d\}$

2+6=8

5. a) Define negative border. How is it used in the sampling algorithm?
- b) Apply sampling algorithm to find all large itemsets in the following grocery data, where $\text{support} = 20\%$ and $D_S = \{T_1, T_2\}$:

Transaction ID	Items Bought
T_1	A,B,C
T_2	A,C
T_3	A,C,D
T_4	A,E
T_5	D,E

- c) Show the dendrogram created by the single, complete, and average link clustering algorithms using the following using the following adjacency matrix:

Items	A	B	C	D
A	0	1	3	5
B	1	0	2	6
C	3	2	0	3
D	5	6	3	0

2+3+3=8

(3)

6. a) Write some applications of data mining.
b) Compare between OLTP and OLAP systems.
c) Write K – means algorithm. Analyse its time complexity. Execute this algorithm over the dataset $\{8.9, 3.4, 22, 16, 30, 8.5, 9.9, 21, 7, 8.9, 10\}$. $2+2+4=8$
7. Write short notes on each of the following :
a) Roll-up and Drill-down operations,
b) Web structured mining,
c) Outliers,
d) Hunt's Algorithm. $2 \times 4 = 8$
-