

Roll No.

--	--	--	--	--	--	--

H

TCS-812

B. Tech. (CSE) (Eighth Semester)
End Semester EXAMINATION, 2017
DATA WAREHOUSING AND MINING
(Elective)

Time : Three Hours] [Maximum Marks : 100

Note : (i) This question paper contains *five* questions.

(ii) All questions are compulsory.

(iii) Instructions on how to attempt a question are mention against it.

(iv) Total marks assigned to each question are twenty.

1. Attempt any *two* questions of choice from (a), (b) and (c). (2×10=20 Marks)

(a) What is Data Cleaning ? In real world data, tuples with missing values for some attributes are common occurrence. Describe the various methods for handling this problem.

[2]

TCS-812

- (b) (i) Suppose a group of 12 sales price records has been sorted as follows :
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215
Partition them into three bins by each of the following methods :
- (1) Equal-frequency partitioning
 - (2) Equal-width partitioning
 - (3) Clustering
- (ii) Write a short note on data characterization and data discrimination.
- (c) List some of the major applications of data mining. Are you able to see some disadvantages of dangers of using data mining in business or government.

2. Attempt any *two* questions of choice from (a), (b) and (c). (2×10=20 Marks)

- (a) Apply the Apriori algorithm for discovering frequent item sets of the following data set :

T-ID	Item purchased
101	Kiwi, Grapes, Star fruit
102	Kiwi, Gooseberry
103	Gooseberry, Pear
104	Kiwi, Grapes, Star fruit
105	Lemon, Star fruit

A-77

[3]

TCS-812

106	Lemon
107	Lemon, Gooseberry
108	Kiwi, Grapes, Mango, Star fruit
109	Mango, Pear
110	Kiwi, Grapes, Star fruit

Use 0.3 for the maximum support value.
Illustrate each step of the apriori algorithm.

- (b) Consider the following transactional data set :

T-ID	Item Id
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

- (i) Construct the FP tree.
- (ii) Find the support and confidence for the given rules :
 $I1 \Rightarrow I2$ and $I2 \Rightarrow I3$.

A-77

P. T. O.

[4]

TCS-812

(c) Discuss the following :

- (i) Multilevel association rules
- (ii) Correlation analysis

3. Attempt any *two* questions of choice from (a), (b) and (c). (2×10=20 Marks)

- (a) What is Bayes classification theorem ?
Classify the Tuple X = (Married = yes, Gender = Male, Employed = No, Credit-Rating = B, Risk Class = A) using naive Bayesian classification for the following data set. The class label attribute, Owns home, has two distinct values {yes, no}.

Married	Gender	Employed	Credit-rating	Risk Class	Owns Home
Yes	Male	Yes	A	B	Yes
No	Female	Yes	A	A	No
Yes	Female	Yes	B	C	Yes
No	Male	No	B	B	Yes
Yes	Female	Yes	B	C	No
No	Female	Yes	B	A	No
No	Male	No	B	B	No
No	Female	Yes	A	A	Yes
Yes	Female	Yes	A	C	No
Yes	Female	Yes	A	C	Yes

A-77

[5]

TCS-812

(b) Discuss the following with example :

- (i) Entropy
- (ii) Information Gain

(c) (i) What are the critical points along an ROC ? Plot the ROC for the following data :

Tuple #	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

(ii) Confusion matrix for classification model.

4. Attempt any *two* questions of choice from (a), (b) and (c). (2×10=20 Marks)

- (a) Suppose that the data mining task is to cluster the ahead eight points [with (x, y) representing location] into three clusters :

A-77

P. T. O.

[6]

TCS-812

$A_1(4, 6)$, $A_2(2, 5)$, $A_3(9, 3)$, $A_4(6, 9)$,
 $A_5(7, 5)$, $A_6(5, 7)$, $A_7(2, 2)$, $A_8(6, 6)$.

Suppose initially we assign A_1, A_2, A_3 as the seeds of three clusters that we wish to find. Use K-means method to show :

- (i) The three cluster centroids after the first iteration.
 - (ii) The final three clusters.
- (b) What is the Clustering ? How is it different that classification ? In what situations clustering can be useful ? List and explain the desired features of cluster analysis.
- (c) Write short notes on the following :
- (i) Outlier analysis
 - (ii) Data mining trends
5. Attempt any *two* questions of choice from (a), (b) and (c). (2×10=20 Marks)
- (a) What is data warehouse ? Diagrammatically illustrate and discuss the data warehousing architecture.
- (b) (i) Distinguish between Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP).
- (ii) Star schema for multidimensional data model.

[7]

TCS-812

- (c) Discuss the following OLAP operations :
- (i) Roll-up
 - (ii) Drill-Down
 - (iii) Slice and Dice
 - (iv) Pivot

TCS-812

90