IS711

USN	1	M	S							

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE, AFFILIATED TO VTU)

BANGALORE – 560 054

SEMESTER END EXAMINATIONS –JANUARY 2016

Course & Branch	**: B.E.- Information Science & Engg.**	**Semester**	**: VII**
Subject	**: Data Mining**	**Max. Marks**	**: 100**
Subject Code	**: IS711**	**Duration**	**: 3 Hrs**

Instructions to the Candidates:

- Answer one full question from each unit.

UNIT – I

1 a) Discuss challenges that motivate development of data mining. C01 (10)

b) Consider the following set of frequent 3 item sets: C01 (04)
{1,2,3},{1,2,4},{1,2,5},{1,3,4},{1,3,5},{2,3,4},{2,3,5},{3,4,5} Assume that there are only five items in the data set.
   i) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$
   ii) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

c) Discuss at least three ways of dealing with missing data listing and their advantages and disadvantages. C01 (06)

2 a) Consider the following transactional table to find out the frequent item sets using Apriori algorithm with the minimum support =20% C01 (08)

TID	List of item-ids
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

b) Discuss the three issues on measurement and data collection aspect of data quality C01 (06)

c) Indicate the type of the data preprocessing task and elaborate. C01 (06)
   i) Replace all the transactions at a given store with a single storewide transaction
   ii) Transform a continuous attribute to a categorical attribute

UNIT – II

3 a) The following contingency table summarizes supermarket transaction data C02 (06)

	Hot dogs	hotdogs	Σrow
hamburgers	2000	500	2500
hamburgers	1000	1500	2500
Σrow	3000	2000	5000

a) Suppose that the transaction rule "hot dogs →hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%

b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

b) Discuss the various methodologies for applying association analysis to continuous data. C02 (06)

c) Define maximal frequent item sets. Illustrate with an example maximal frequent item sets. C02 (08)

4 a) Define the following terms C02 (04)
i)Discretization-based approach ii) Scaling Property

b) Describe the inversion and null addition properties of the objective measures. Indicate a measure for each that satisfies the properties. C02 (06)

c) a)Draw a contingency table for each of the following rules using the transactions shown below C02 (10)

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

Rules {b}→{c},{a}→{d},{b}→{d},{e}→{c},{c}→{a}

a) Use the contingency tables in part (a) to compute the rank the rules in decreasing order according to the following measures

i)Support

ii)Confidence

iii)$\text{interestfactor}(X \rightarrow Y) = \frac{P(X,Y)}{P(X)}P(Y)$

UNIT – III

5 a) Consider the data set shown below C03 (10)

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

i) Estimate the conditional probabilities for P(A=1|+), P(B=1|+), P(C=1|+),P(A=1|-), P(B=1|-) and P(C=1|-).
ii) Use the conditional probabilities in part(a) to predict the class label for a test sample (A=1,B=1,C=1)using the Naïve Bayes Approach.
iii) Compare P(A=1),P(B=1) and P(A=1,B=1).State the relationships between A and B

b) Compare the rule based and class based Ordering scheme. C03 (04)
c) Provide and discuss the K-nearest neighbor classification algorithm. C03 (06)

6 a) Consider the training examples shown below in Table for a binary classification problem C03 (08)

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	-

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?

b) Provide the Sequential Covering algorithm that is used to build a rule based classifier. C03 (06)
c) What is boosting? State why it may improve the accuracy of decision tree induction. C03 (06)

UNIT – IV

7. a) Explain any three different types of clusters. C04 (06)
b) Use the similarity matrix in the following table to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. C04 (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

c) Explain the DBSCAN algorithm. Illustrate with an example. C04 (08)

8. a) Highlight the strength and weakness of K-means algorithm. C04 (06)

b) Given a table where patients are described by binary attributes. Excluding name, all other attributes are asymmetric binary variables. C04 (06)

Name	Attr1	Attr2	Attr3	Attr4	Attr5	Attr6
J	1	0	1	0	0	0
M	1	0	1	0	1	0
k	1	1	0	0	0	0

Compute the distance between each pair of the 3 objects. Provide suitable explanation.

c) Explain the single link or MIN version of Hierarchical clustering with an example. C04 (08)

UNIT – V

9 a) List the challenges posed by the web for effective resource and knowledge discovery. Describe ways of resolving these challenges. C05 (10)

b) Write a note on mining the web page Layout structure. C05 (04)

c) Discuss the role of data mining in financial sector. C05 (06)

10 a) Discuss in detail the application of data mining for the retail industry. C05 (06)

b) Write a short notes on C05 (08)

i) Web usage mining

ii) Spatial Data mining

c) What are the basic measures for text retrieval? Discuss the Text retrieval methods in detail. C05 (06)

****