No mobile phones

IS701

USN	1	M	S							

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE, AFFILIATED TO VTU)
BANGALORE – 560 054

SEMESTER END EXAMINATIONS – JANUARY 2015

Course & Branch	: B.E. – INFORMATION SCIENCE & ENGG.	Semester	: VII
Subject	: Data Mining	Max. Marks	: 100
Subject Code	: IS701	Duration	: 3 Hrs

Instructions to the Candidates:

- Answer one full question from each unit.

UNIT – I

1. a) Design an algorithm to generate Frequent Itemset using Apriori principle. Apply the algorithm to the following instance of transactions to generate all possible frequent item sets given ***Minsup=30%***. (10)

TID	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}

Table1

b) Discuss the different attribute types with examples and operations. (10)

2 a) What is data mining? Discuss the two categories of data mining tasks. Explain the knowledge discovery in databases (KDD) with neat diagram. (10)

b) Discuss any five computational complexity of Apriori algorithm. (10)

UNIT - II

3. a) What is multilevel association mining? Illustrate with examples how multilevel association rules can be mined. (10)

b) Consider the market basket transaction given in **Table1**. Apply FP growth algorithm to same to construct FP tree and generate all possible frequent item set using the same. (10)

4. a) Define binary variables. Discuss the terms Similarity and Dissimilarity with respect to binary variables. (05)

b) Discuss the Alternate Methods for Generating Frequent Item sets. (10)

c) Define IS Measure. Consider the following contingency tables for the word pairs {{p,q} and {r,s}}: (05)

	p	$\bar{p}$	
q	880	50	930
$\bar{q}$	50	20	70
	930	70	1000

	r	$\bar{r}$	
s	20	50	70
$\bar{s}$	50	880	930
	70	930	1000

Compute IS measure for the word pairs {p,q} and {r,s}.

UNIT - III

5. Design an decision tree using Information gain for the following training data set, D: (20)

RID	age	Income	Student	Credit_Rating	Class Buys_Computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	mid_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	mid_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	mid_aged	medium	no	excellent	yes
13	mid_aged	high	yes	fair	yes
14	senior	Medium	no	excellent	no

Table2

State the disadvantage of information gain.

6. a) Define Classification. Write a short note on IF-THEN Rules for classification. Compute coverage and accuracy for the rule R: IF (age=youth) Λ (student=yes) THEN buys_computer=yes. (**Note:** Consider the training data set D, given in **Table2**) (08)

   b) Why is naïve Bayesian classification called "naive"? Briefly outline the major ideas of naïve Bayesian classification. Predict a class label for the tuple X=(age=youth, income=medium, student=yes, credit_rating=fair) using naïve Bayesian classification. (**Note:** Consider the training data set D, given in **Table2**) (12)

UNIT - IV

7. a) What is cluster analysis? Explain different types of clustering in detail. (10)

   b) Explain the K-means clustering method. (10)

8. a) Write a short note on (10)
      i. DBSCAN
      ii. Strengths and Weaknesses of K-means clustering method

   b) List out the important issues for cluster validation. (05)

   c) Outline the Basic Agglomerative hierarchical Clustering Algorithm. (05)

UNIT - V

9. a) Discuss various approaches to text mining. (10)

   b) Discuss the usage of Web mining? (07)

   c) Define the terms text mining and web mining. (03)

10. a) Discuss data mining application for Financial Data Analysis. (10)

    b) Write a short note on "Hadoop Schedulers". (05)

    c) What are the two basic measures for assessing the quality of text retrieval? (05)

****