IS701

USN	1	M	S							

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(AUTONOMOUS INSTITUTE, AFFILIATED TO VTU)

BANGALORE – 560 054

SEMESTER END EXAMINATIONS – DEC 2013 / JAN 2014

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. Indicate the attribute type (Nominal, Ordinal, Interval or Ratio) for the following attributes and elaborate. (05)
   i. employee id number ii. Calendar Dates iii. Length iv. Counts v. grades

   b. Discuss at least 3 ways of dealing with missing data listing their advantages and disadvantages. (09)

   c. Provide the definitions of the following metrics i. support count ii. Confidence. (06)
   Given the dataset below.
   i. Evaluate the support for the itemsets ,{e},{b,d},{a,b}.
   ii. Compute the confidence for the association rules
   {b,d}->{e} and {e} ->{a,b}.

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

c. (06)

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

Given the above transaction dataset obtain the frequent itemsets using Apriori principle and $F_{k-1}XF_k$ Method. Minimum Support – 2.

UNIT- II

3. a. Construct an FP-tree for the following dataset. Given Minimum Support= 3. Also generate rules from the frequent itemsets. (08)

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

b. Discuss any two alternate methods of generating frequent itemsets with suitable illustrations. (06)

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

4. a. Given the contingency matrix compute the lift(T->C) and correlation analysis(T->C). (10)

	C	~C	
T	150	50	200
~T	650	150	800
	800	200	1000

i. Indicate the correlation between T and C. Explain.
ii. Indicate whether lift and correlation analysis satisfy the inversion property. Explain.
~C indicates C not present and ~T indicates T not present.

b. Describe the issues to be considered when applying association analysis to the binarized data. (06)

c. Define the following terms (04)
i. maximal frequent itemset ii. closed frequent itemset
iii. Discretization-based approach iv. Scaling Property

UNIT-III

5. a. Discuss the general approach for building a classification model with a diagram. (06)

b. Given nodes (07)
N1(number of instances in class0 = 0, number of instances in class1 = 6),
N2(number of instances in class0 = 1, number of instances in class1 = 5),
N3(number of instances in class0 = 3, number of instances in class1 = 3)
Compute the entropy for each of the nodes and indicate which node has the lowest impurity value. Elaborate.

c. Provide the Sequential covering algorithm that is used to build a rule based classifier. (07)

6. a. List the characteristics of a decision tree. (06)

b. Provide and discuss the k-nearest neighbor classification algorithm. (06)

c. Consider a binary classification problem with the following set of attributes and attribute values: (08)

Air Conditioner = {Working, Broken}
Engine = {Good, Bad}
Mileage = {High, Medium, Low}
Rust = {Yes, No}

Suppose a rule-based classifier produces the following rule set:

Mileage = High ⟶ Value = Low
Mileage = Low ⟶ Value = High
Air Conditioner = Working, Engine = Good ⟶ Value = High
Air Conditioner = Working, Engine = Bad ⟶ Value = Low
Air Conditioner = Broken ⟶ Value = Low

i. Are the rules mutually exclusive? Give Reasons
ii. Is the rule set exhaustive? Give Reasons
iii. Is ordering needed for this set of rules? Give Reasons
iv. Do you need a default class for the rule set? Why?

UNIT- IV

7. a. Explain 3 different types of clusters. (06)

b. Given a table where patients are described by binary attributes. Excluding name, all other attributes are asymmetric binary variables. (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. Elaborate the DBSCAN algorithm listing its strengths and weaknesses. (08)

8. a. Choosing the proper initial centroids is the key step of the basic k-means procedure. Explain why? (06)

b. Given below are the x and y coordinates of the points and Euclidean distances between them. (08)

Point	**X coordinate**	**Y coordinate**
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

	P1	P2	P3	P4	P5	P6
P1	0.00	0.24	0.22	0.37	0.34	0.23
P2	0.24	0.00	0.15	0.20	0.14	0.25
P3	0.22	0.15	0.00	0.15	0.28	0.11
P4	0.37	0.20	0.15	0.00	0.29	0.22
P5	0.34	0.14	0.28	0.29	0.00	0.39
P6	0.23	0.25	0.11	0.22	0.30	0.00

1. Show the single link clustering and single link dendrogram for the above dataset. Also compute the distance between the clusters {3,6} and {2,5} .
2. Show the complete link clustering and complete link dendrogram for the above dataset. Also compute the distance between the clusters {3,6} and {2,5} .

c. Define with examples (06)
1. Cluster Analysis 2. Partitional Clustering 3. Clustering Tendency

UNIT- V

9. a. Discuss any 5 methodologies for stream data processing and stream data systems. (10)

b. What are basic measures of text retrieval? Elaborate them. (06)

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

10. a. Discuss the application of data mining for the Retail Industry. (10)

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

****