COMP 8 (E-III) 4 (RC)

B.E. (Computer) (Semester – VIII) (Revised Course) Examination, May/June 2012
DATA MINING (Elective – III)

Duration : 3 Hours Total Marks : 100

Instructions : *1) Answer **any five full** questions by selecting at least **one** from **each** Module.*
*2) Make necessary assumptions wherever **necessary**.*
*3) Answer to the question must be **written** in the **same** sequence.*

Module – 1

1. a) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific example of how techniques, such as clustering, classification, association rule mining and anomaly detection can be applied. **8**

b) Consider the following documents
D1 : Illegal mining should be stopped.
D2 : Mining of data is crucial for analysis
D3 : Knowledge can be extracted from huge corpus
D4 : Data mining and pattern mining are useful.
Compute the cosine similarity matrix for the above sets of documents. **8**

c) Explain the various types of data set. **4**

2. a) With the help of neat block diagram, explain the data mining process. **8**

b) Calculate the correlation and covariance between the age and salary. **6**

Sr. No.	Age	Salary
1	20	10000
2	21	12000
3	22	13000
4	23	14000
5	24	15000
6	25	16000
7	26	15000
8	27	14000
9	28	13000
10	29	12000

P.T.O.

c) Compute the Cosine, Correlation and Euclidean distance measure between X and Y.

i) X = (1, 1, 2, 2) and Y = (2, 2, 3, 3)

ii) X = (3, 2, 4, 1) and Y = (2, 1, –1, 2) 6

Module – 2

3. a) Construct the Decision Induction Tree for the given set of data. 8

Custer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	F	Luxury	Small	C1
6	F	Luxury	Large	C1
7	F	Sports	Large	C1
8	M	Luxury	Medium	C1
9	M	Family	Large	C0
10	M	Sports	Medium	C0

b) Briefly explain two cause of model over fitting. 4

c) List and explain various data visualization techniques. 8

4. a) What is OLAP ? How OLAP helps in Data Analysis ? Discuss. 10

b) What is Data Warehouse ? What is the role of data warehousing in data mining ? What are the different implementation issues of data warehouse ? Explain. 8

c) Define Gini and Classification Error. 2

Module – 3

5. a) Consider the data set given below. 8

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

Using apriori algorithm generate strong association rules. Let min_support = 3 and confident = 80%.

b) Construct FP tree using data set from Q. 5. a). Identify the frequent itemsets using FP tree. Compare the apriori algorithm and FP tree algorithm. 8

c) Define and give one example for the following terms. 4

i) Maximal frequent Itemset

ii) Closed frequent Itemset.

6. a) Classify the data points x = 5.0 according to its 1-, 3-, 5- and 9- nearest neighbor (using majority vote). 8

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y	-	-	+	+	+	-	-	+	-	-

b) Explain the Rule based classifier ? How this method is different from the Nearest neighbor classifier. 8

c) Explain the direct method for rule extraction. 4

Module – 4

7. a) Consider the similarity matrix table as shown below. Construct dendogram and nested cluster using single link clustering for given similarity matrix. 8

	P1	P2	P3	P4	P5
P1	0.0	0.10	0.41	0.55	0.35
P2	0.10	0.0	0.64	0.47	0.98
P3	0.41	0.64	0.0	0.44	0.85
P4	0.55	0.47	0.44	0.0	0.76
P5	0.35	0.98	0.85	0.76	0.0

b) List and explain the various application of anomaly detection. 8

c) Explain the key issues in hierarchical clustering. 4

8. a) Consider the following data points. 8
P1 (1,1), P2 (1,2), P3 (2,2), P4 (2, 1), P5 (1.1, 2.1), P6 (11, 12), P7 (7,7) P8 (6,7), P9 (7,6) P10 (5,4).
Trace the DBSCAN clustering algorithm and identify final cluster and noise. Min_pts = 3, epsilon = 3.

b) List and explain the any two methods of outlier detection. 6

c) Explain the K-means algorithm with suitable data set. 6