COMP 8E – III (3) (RC)

B.E. (Computer) (Semester – VIII) (RC) Examination, May/June 2017
DATA MINING (Elective – III)

Duration : 3 Hours | Max. Marks : 100

Instructions : *i) Attempt **any five** questions by selecting at least **one** question from **each** Module.*
*ii) Make suitable assumptions **if required**.*

MODULE – I

1. a) With the help of a neat block diagram, explain data mining process. **7**

   b) For the following vectors X and Y, calculate the indicated similarity or distance measures. **3**

   X = (1, 1, 1, 0, 0, 1, 1, 1) Y = (1, 0, 0, 1, 0, 1, 1, 0)

   i) Euclidean

   ii) Cosine

   iii) Jaccard.

   c) List and explain the different types of datasets with appropriate examples. **6**

   d) Classify the following attributes as binary, discrete or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Briefly indicate your reasoning. **4**

   i) Brightness as measure by a light meter.

   ii) Gold, silver, broze medals as awarded at Olympics.

2. a) Explain the process of feature subset selection using a flowchart. **4**

   b) Discuss advantages and disadvantages of using sampling to reduce the number of data objects that need to be displayed. Would Simple Random Sampling (without replacement) be a good approach to sampling ? Why or Why not ? **5**

   c) Explain the following : **5**

   i) Curse of dimensionality

   ii) Aggregation.

d) Calculate the correlation and covariance between the experience and income for the given data records. 6

Record No.	Experience (in years)	Income (per month) Rs.
1	19	5,000
2	20	7,000
3	21	8,000
4	22	9,000
5	23	10,000
6	24	11,000
7	25	10,000
8	26	9,000
9	27	8,000
10	28	7,000

MODULE – II

3. a) What is OLAP ? How does OLAP help in data analysis ? 5

b) Define the following terms with respect to summary statistics : 4

i) Frequency and mode ii) Mean and Median.

c) Describe the Hunt's algorithm to construct decision trees. 4

d) Why do we need a separate data warehouse ? Explain the components of a data warehouse ? 7

4. a) Construct the decision tree for the following data for the target attribute 'Transportation mode' ? 10

Gender	Car Ownership	Travel Cost ($)/km	Income Level	Transportation mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

b) Explain pre-pruning with an example. 4

c) Explain with a suitable examples the following OLAP operations : 4

i) slice ii) dice.

d) Explain with an example stem and leaf plots. 2

MODULE – III

5. a) Consider the given data set. 10

Using Apriori Algorithm generate strong association rules.

Min_Support = 03 Confidence=50%

Transaction Id	Items Bought
1	{Milk, Bread, Eggs}
2	{Bread, Sugar}
3	{Bread, Cereal}
4	{Milk, Bread, Sugar}
5	{Milk, Cereal}
6	{Bread, Cereal}
7	{Milk, Cereal}
8	{Milk, Bread, Cereal, Eggs}
9	{Milk, Bread, Cereal}

b) Explain the Rule based classifier ? How is this different from the Nearest Neighbour Classifier ? 7

c) Define Closed Frequent Itemset with an example. 3

6. a) Explain Candidate Generation and Candidate Pruning in detail. 6

b) Construct FP – Tree for the given data set. Min_Support = 02. 10

Transaction_Id	Items Bought
100	{a, c, d, f, g, i, m, p}
110	{a, b, c, f, l, m, o}
120	{b, f, h, j, o, w}
130	{b, c, k, s, p}
140	{a, f, c, e, l, p, m, n}

c) What do you understand by coverage and accuracy of a rule ? Show the steps to calculate the average and accuracy of a rule with respect to a dataset. 4

MODULE – IV

7. a) Describe in detail the following distance metrics to calculate the distance between two clusters with a suitable example. **8**

   i) single link ii) complete link

   iii) group average iv) ward distance

   b) Consider the following data set : **8**

   Perform clustering using DBSCAN and display the clusters and lable all data points as border point, core points and noise points.

   Min_points = 3 and Epsilon = 03

	x	y
P1	2	10
P2	2	5
P3	8	4
P4	5	8
P5	7	5
P6	6	4
P7	1	2
P8	4	9

   c) Define an outlier. Explain the significance of outlier detection. **4**

8. a) Consider the following data set : **10**

   Construct the dendogram and draw the nested clusters using single linkage clustering.

Data – point	a1	a2
P1	1	1
P2	1.5	1.5
P3	5	5
P4	3	4
P5	4	4
P6	3	3.5

   b) Discuss the important issues that need to be addressed when dealing with anomalies. **5**

   c) Explain proximity based outlier detection and discuss its strength and weaknesses. **5**