COMP 8 E – III (4) (RC)

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

Duration : 3 Hours | Total Marks : 100

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

MODULE – I

1. a) Describe the following procedures for attribute subset selection : **6**
   i) Stepwise Forward Selection.
   ii) Stepwise Backward Elimination.
   iii) Combination of Forward Selection and Backward Elimination.

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

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

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

   i) Euclidean ii) Cosine iii) Jaccard

   c) List and explain the different challenges that motivated the development of data mining. **5**

   d) What is data pre-processing ? Explain Dimensionality Reduction as the techniques for performing data pre-processing. **6**

2. a) Describe the Predictive and Descriptive tasks of data mining with an appropriate example. **4**

   b) What is an attribute ? Explain different types of attributes with an example. **5**

   c) Consider the following group of data points. **6**

   125, 45, 175, 95, 775, 625, 675, 215, 335, 285, 625, 175

   Normalize the data points by using :
   i) Min-Max Normalization Method where Min = –3 and Max = 5.
   ii) Z-Score Normalization Method.

   d) Explain any two types of datasets with an appropriate example. **5**

P.T.O.

MODULE – II

3. a) What is OLAP ? How is OLAP different from OLTP ? 5

   b) Explain in detail the causes of Model overfitting. 7

   c) Write short notes on :
      i) Histograms
      ii) Scatter plots. 5

   d) Classification is supervised learning. Justify. 3

4. a) Construct the decision tree for the following data for the target attribute 'Play Outdoors'. 12

Day	Climate	Humid	Breezy	Play_outdoors
D1	S	H	W	N
D2	S	H	S	N
D3	O	H	W	Y
D4	R	H	W	Y
D5	R	N	W	Y
D6	R	N	S	Y
D7	O	N	S	Y
D8	S	H	W	N
D9	S	N	W	Y
D10	R	N	W	Y
D11	S	N	S	Y
D12	O	H	S	Y
D13	O	N	W	Y
D14	R	H	S	N

   b) Explain post-pruning with an example. 4

   c) How is multi-dimensional data represented ? Explain with a suitable example the OLAP operation roll-up. 4

MODULE – III

5. a) Write the FP Tree Algorithm and Construct FP – Tree for the given data set. **12**
Min_Sup = 03

Transaction id (T_{Id})	Items Bought
t_1	strawberry, litchi, oranges
t_2	strawberry, butter _ fruit
t_3	butter _ fruit, vanilla
t_4	strawberry, litchi, oranges
t_5	banana, oranges
t_6	banana
t_7	banana, butter _ fruit
t_8	strawberry, litchi, apple, oranges
t_9	apple, vanilla
t_{10}	strawberry, litchi

b) Explain rule induction using sequential covering algorithm. **8**

6. a) Explain and give example for the following : **5**
   i) Maximal Frequent Itemset.
   ii) Closed Frequent Itemset.

b) What do you understand by the following terms ? Provide suitable example. **6**
   i) Rule based rule ordering.
   ii) Class based rule ordering.

c) Write the algorithm and explain the K-Nearest Neighbour algorithm for classification. **6**

d) How has the association rule mining problem traditionally formulated ? Also state the Apriori Principle. **3**

MODULE – IV

7. a) State the K-means algorithm. Perform K-means clustering for the following data points into 3 clusters. The distance function is Euclidean distance. Initially assign P_2, P_4 and P_5 as center of each cluster respectively. **10**

	x	y
P_1	2	7
P_2	2	5
P_3	8	5
P_4	5	6
P_5	7	10
P_6	6	3
P_7	1	2
P_8	4	9

b) Explain in detail the different types of clusters. **10**

8. a) Explain Density based outlier detection technique and list its strengths and weaknesses. **8**

b) Consider the following data set :

Construct the dendrogram and draw the nested clusters using complete linkage clustering. **12**

data point	a_1	a_2
P_1	0.40	0.53
P_2	0.22	0.38
P_3	0.35	0.32
P_4	0.26	0.19
P_5	0.08	0.19
P_6	0.45	0.30