B.E. (Comp.) (Semester – VIII) (RC 2007-08) Examination, Nov./Dec. 2018
DATA MINING

Duration : 3 Hours Total Marks : 100

Instructions : *1) Attempt* ***any five full*** *questions by selecting* ***at least 1*** *full question from* ***each*** *module.*

2) Make suitable assumptions if necessary, state clearly assumptions made.

MODULE – I

1. a) Consider x = (0, – 1, 0, 1) y = (1, 0, –1, 0) :
Compute distance between x and y using SMC, Jaccard, Cosine and Euclidean measures. 4

b) Explain using a flowchart the following procedures for attribute subset selection : 6

i) stepwise forward selection

ii) stepwise backward elimination

iii) a combination of forward selection and backward elimination.

c) Explain the following methods. Cite applications for the same : **(5+5)**

i) Regression

ii) Principal Component Analysis.

2. a) With the help of a neat labeled diagram explain how are DM and KDD related. 8

b) The age values for the data tuples are :
20, 20, 21, 22, 22, 25, 25, 25, 25, 13, 15, 16, 16, 19, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

i) Use min-max normalization to transform the values 20, 30, 40 and 70 in the range [0.0, 1.0].

ii) If the standard deviation of age is 12.94 years, what are the z-scores corresponding to 20, 30, 40 and 70 years ? 6

c) Giving suitable example explain how the concept/class descriptions are derived. 6

P.T.O.

MODULE – II

3. a) Discuss the advantages and disadvantages of using sampling to reduce the number of data objects that need to be selected. Would simple random sampling (without replacement) be a good approach to sampling ? Why or why not ? 4

b) Compare OLTP and OLAP. 4

c) Consider the database of a car insurance company shown in Table 1 : 12

Name	AgeGroup	CarType	CrashRisk
Ben	30-40	Family	Low
Paul	20-30	Sports	High
Bill	40-50	Sports	High
James	30-40	Family	Low
John	20-30	Family	High
Steven	30-40	Sports	High

i) Assume that CrashRisk is the class attribute. Explain which of the remaining attributes are appropriate for classification. Show the complete decision tree that is produced on this dataset.

ii) Grow the tree from this root node, until the leaf nodes are "pure", i.e. contain only records from the same class. Explain what will be the class label for nodes with no training samples. Show the split test used at each node. For each leaf node, show the class and the records associated with it. Explain how you derived the split node using the information gain (here I(2, 4) for the CrashRisk class) and entropy (E(age) and E(cartype) considering AgeGroup and CarType for Classification) concepts.

iii) Using the produced classifier, determine the class label of the following records :

{Pete, 20-30, Sports} and {Bob, 40-50, Family}

4. a) Explain supervised and unsupervised learning approaches stating their strengths and weaknesses. List two algorithms of each type. 4

b) The weather data is stored for different locations in a warehouse. The weather data consists of 'temperature', 'pressure', 'humidity' and 'wind velocity'. The location is defined in terms of 'latitude', 'longitude', 'altitude' and 'time'. Assume that nation ( ) is a function that returns the name of the country for a given latitude and longitude. Propose a OLAP model for this case. 8

c) With the help of an error graph explain the state of overfitting in decision trees. Explain the idea of pre-pruning and post-pruning and the related methods to fix the same. 8

MODULE – III

5. a) Imagine that you are given the following set of training examples shown in Table 2. Feature F1 can take on the values a, b, or c; Features F2 is Boolean-valued; and Feature F3 is always a real – valued number in [0, 1]. **12**

Table 2 :

	F1	F2	F3	Category
Example 1	A	T	0.2	+
Example 2	B	F	0.5	+
Example 3	B	F	0.9	+
Example 4	B	T	0.6	–
Example 5	A	F	0.1	–
Example 6	A	T	0.7	–

i) How might a Naïve Bayes system classify the following test example ? (Discretize the numeric feature into three equal – width bins) F1 = c, F2 = T, F3 = 0.8.

ii) Describe how a 2-nearest – neighbour algorithm might classify part (i) test example.

b) The efficiency of the Apriori algorithm for mining association rules may be improved by using any of the following techniques : **8**

i) Pruning

ii) Transaction reduction

iii) Partitioning

iv) Sampling.

Explain two of these approaches.

6. a) Consider the following **Table 3** : **(6+6)**

Tld	Items
100	bread, cheese, eggs, juice
200	bread, cheese, juice
300	bread, milk, yogurt
400	bread, juice, milk
500	cheese, juice, milk

Consider 50% support and 75% confidence. Write apriori algorithm and demonstrate on the dataset given in Table 3.

b) List the advantages of a FP-tree for mining association rules. 4

c) Write k-NN algorithm and suggest an application of the same. 4

MODULE – IV

7. a) Consider the following dataset of six objects, each with two attributes : (3+7)
   A_1 (4, 6), A_2(2, 5) A_3(9, 3) A_4(6, 9) A_5(7, 5) A_6(5, 7)
   Perform the following :
   i) Create a distance matrix for six objects using Euclidean distance.
   ii) Using the agglomerative method, determine the two objects that should form the basis for splitting the given dataset.

   b) "In a cluster defined in DBSCAN, any two objects are density connected." Do you agree/disagree with this statement ? Why ? 2

   c) Differentiate between k-means and k-mediods clustering w.r.t. strengths and limitations. 2

   d) What is an outlier ? Why is outlier mining important ? Briefly describe distance based outlier detection. 6

8. a) Consider the following items to cluster : 10
   {2, 4, 10, 12, 3, 20, 30, 11, 25}
   i) Using the k-means clustering, cluster the given items considering k = 3.
   ii) Write the k-means algorithm.

   b) You have learned about data mining techniques which deal with outlier detection. Decide which outlier detection method is best for the following problems. Explain each choice and show the working : (4+4)
   i) A model designed to accept or reject credit card applications.
   ii) A model designed to determine those individuals likely to develop colon cancer.

   c) k-means has difficulties to cluster datasets that contain categorical attributes. What is this difficulty and how do you resolve it ? 2