COMP8 (E – III) 4(RC)

B.E. (Comp.) (Semester – VIII) (RC 2007-08) Examination, May/June 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) Describe three main reasons that would motivate an enterprise to take an interest in data mining. Describe the steps that are required in a typical data mining process. 7

b) Describe the working model for Principal Component Analysis. (6+3)

Explain its application to the following case :

The people in marketing would like a better understanding of their diferent customers. They want to know what distinguishes customers - what are the key attributes that make a customer unique ? The idea isn't to group similar customers, but to identify the attributes that set customers apart. Justify your answer.

c) Given two objects represented by the attribute values (1, 6, 2, 5, 3) and (3, 5, 2, 6, 6). 4

i) Compute the Euclidean distance between the two objects.

ii) Compute the Cosine distance between the two objects.

2. a) Each customer visiting a supermarket is characterized by the following attributes : 6

- Ssn
- Items_Bought (The set of items the bought last month)
- Amount_spend (Average amount spent per purchase; it has a mean of 50.00, a standard deviation of 40, the minimum is 0.05 and the maximum is 600)

P.T.O.

- Age

i) Identify the data type of each attribute.

ii) Investigate and suggest the retention of attributes based on their importance.

iii) Suggest an appropriate distance function is to compute the similarity of customers of a supermarket.

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. 5

Describe various data smoothing methods. Using one of the appropriate methods of data smoothing, demonstrate the working on the given age dataset.

c) Give a brief description using illustration of the following : **(3+3+3)**

i) Regression

ii) Generalization

iii) Aggregation.

MODULE – II

3. a) Explain the multidimensional cube model used for OLAP. How does a snowflake scheme differs from a STAR schema ? Explain with an example. Name any two disadvantages of snowflake schema. **6**

b) 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**

c) What is Attribute Oriented Induction ? Consider Table 1 and Table 2. **10**

Table 1 : Graduate Student Table

Gender	Major	Count
M	Science	20
F	Science	30
M	Engineering	10
F	Engineering	40

Table 2 : Under-Graduate Student Table

Gender	Major	Count
M	Science	15
F	Science	40
M	Engineering	10
F	Engineering	10

i) Compute Gain (Gender)

ii) Gain (Major)

iii) Of the two attributes Gender, Major which one is more relevant.

4. a) What is overfitting of a decision tree ? What problems can overfitting lead to ? What are the suggested approaches for dealing with overfitting ? **8**

b) List strengths and weaknesses of supervised and unsupervised learning approaches. Also list two algorithms of each type. **2**

c) Explain the main steps and construct a Decision Tree using the training data in the Table 3. Divide the Height attribute into ranges as follows : (0, 1.6], (1.6, 1.7], (1.7, 1.8], (1.8, 1.9], (1.9, 2.0], (2.0, 5.0]. **10**

Table 3 :

Gender	Height	Class
F	1.6 m	Short
M	2 m	Tall
F	1.9 m	Medium
F	1.88 m	Medium
F	1.7 m	Short
M	1.85 m	Medium
F	1.6 m	Short
M	1.7 m	Short
M	2.2 m	Tall
M	2.1 m	Tall
F	1.8 m	Medium
M	1.95 m	Medium
F	1.9 m	Medium
F	1.8 m	Medium
F	1.75 m	Medium

MODULE – III

5. a) Why is the k-nearest neighbor called as a lazy classifier ? Explain with the help of an example by comparing it with the decision tree classifier. **6**

b) Consider the dataset given in Table 4. **10**

Table 4 :

No.	Color	Type	Origin	Stolen ?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

i) Using Table 4, construct the Naïve Bayes Classifier.

ii) Classify the test case ‘Red, Domestic, SUV’.

c) Explain why the Apriori algorithm for mining association rules may be improved by the following techniques ? Explain.

i) Pruning

ii) Transaction reduction. **4**

6. a) Compare and contrast lazy and eager learning. **4**

b) Explain maximal frequent itemsets and closed frequents itemset in association mining. **6**

c) Consider the following Table 5 :

Table 5 :

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

Construct FP tree for the dataset given in Table 5. Consider minimum support = 50%. **6**

d) Explain rule ordering scheme in a rule based classifier. State its advantages. **4**

MODULE – IV

7. a) You have learned about data mining techniques which deal with data while learning. Decide which technique is best for the following problems ? Explain the method chosen in detail : **(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.

b) Why is outlier mining important ? Briefly describe any three of the following different approaches used for outlier analysis : **12**

i) Statistical-based outlier detection

ii) Distance-based outlier detection

iii) Density-based outlier detection

iv) Deviation based outlier detection.

8. a) Suppose that a data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. **8**

A_1 (4, 6) A_2 (2, 5) A_3 (9, 3) A_4 (6, 9) A_5 (7, 5) A_6 (5, 7) A_7 (2, 2) A_8 (6, 6)

Suppose initially we assign A_1, A_2 and A_3 as the seeds of three clusters that we wish to find. Use the k-means to show the final three clusters.

b) Provide example and enlist the characteristics of the following types of outliers :

i) Global outlier

ii) Contextual outlier

iii) Collective outlier. **(3×2=6)**

c) Explain how the clusters found by the agglomerative clustering algorithm differ to those found by the k-means clustering algorithm ? **6**

———————