10161 15 E

COMP 8 (E-III) 4(RC)

B.E. (Computer) (Semester – VIII) (RC) Examination, May/June 2015

DATA MINING

(Elective – III)

Duration : 3 Hours | Total Marks : 100

Instructions : *1) Answer **any five** questions by selecting at least **one** from **each** Module.*
*2) Assume suitable data if **necessary**.*

MODULE – I

1. a) Describe the steps involved in data mining when used for the process of knowledge discovery. **5**

   b) 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. **6**

      i) Distance from the center of the campus

      ii) ISBN number for books

      iii) Ability to pass light of the following values: Opaque, Translucent, Transparent.

   c) What is discretization ? Cite a relevant example and write in points the significance of discretization. **4**

   d) Suppose a group of 12 *sales price* records has been sorted as follows : **5**

      5; 10; 11; 13; 15; 35; 50; 55; 72; 92; 204; 215 :

      Partition them into three bins by each of the following methods.

      i) equal-frequency partitioning

      ii) equal-width partitioning.

2. a) Use a flow chart to summarize 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) Use the two methods below to *normalize* the following group of data : **4**

      200; 300; 400; 600; 1000

      i) min-max normalization by setting min = 0 and max = 1

      ii) z-score normalization

P.T.O.

c) 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

d) For the following vectors X and Y, calculate the indicated similarity or distance measure : 3

X = (0, −1, 0, 1), Y = (1, 0, −1, 0)

i) Cosine

ii) Euclidean

iii) Jaccard

e) What do you understand by aggregation ? 2

MODULE – II

3. a) Why do you need a separate data warehouse ? What are the various components of a data warehouse ? 6

b) Consider the following example : 9

State	Season	Barometer	Weather
AK	Winter	Down	Snow
HI	Winter	Down	Sun
HI	Summer	Up	Sun
CA	Summer	Up	Rain
AK	Winter	Up	Snow
CA	Winter	Down	Sun
AK	Summer	Down	Sun
CA	Winter	Up	Rain
HI	Summer	Down	Sun

Using the ID3 algorithm construct a decision tree for predicting the target attribute weather.

c) What are overfitted models ? Explain their effects on performance of a decision tree. 5

4. a) Classification is supervised learning. Justify. 3

b) Explain what is crossvalidation in classification ? 6

c) Explain the methods for computing best split. 6

d) How is multidimensional data represented ? What do you understand by slice and dice ? Give an example. 5

MODULE – III

5. a) Explain rule induction using sequential covering algorithm. 9

b) Given a transactional database X : 9

TID	Items
T01	A, B, C, D
T02	A, C, D, F
T03	C, D, E, G, A
T04	A, D, F, B
T05	B, C, G
T06	D, F, G
T07	A, B, G
T08	C, D, F, G

Using the Apriori algorithm and the threshold values support = 25% and confidence = 60%, find :

i) All frequent itemsets in database X.

ii) Strong association rules for database X.

c) What are closed patterns ? 2

6. a) Write the algorithm and explain the K-nearest neighbor algorithm for classification. 6

b) What is rule pruning ? How is it carried out ? 5

c) Write the FP tree algorithm and construct the FP tree from the following transaction database : 9

TID	Items in the Transaction	Ordered Frequent Items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

MODULE – IV

7. a) What is an outlier ? Why is outlier mining important ? Describe distance based outlier detection. **8**

   b) Write the K-means algorithm. Consider the data mining task is to cluster the following 8 points into 3 clusters. **10**

   A1 (2, 10)
   A2 (2, 5)
   A3 (8, 4)
   B1 (5, 8)
   B2 (7, 15)
   B3 (6, 4)
   C1 (1, 2)
   C2 (4, 9)

   The distance function is Euclidean distance. Suppose initially we assign A1, B1, C1 as center of each cluster respectively. Use K-means algorithm to show

   i) The three cluster centers after first round execution

   ii) The final 3 clusters.

   c) List the key issues in hierarchical clustering. **2**

8. a) Both K-means and K-medoids algorithms can perform effective clustering. **(4+4)**

   i) State the working principle and illustrate the strength and weakness of K-means in comparison with K-medoids.

   ii) Illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (AGNES)

   b) What are the different approaches to anomaly detection ? **6**
   i) Model based techniques

   ii) Proximity based techniques.

   c) Why outlier detection and mining is important for cluster analysis ? Provide an example to support your answer. **6**