COMP 8 (E-III) – 4 (RC)

B.E. (Computer) (Semester – VIII) (RC) Examination, May/June 2016

DATA MINING (Elective – III)

Duration : 3 Hours Max. 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) What Data Mining Tasks exist ? Describe each of them. 4

   b) Assume that there is a dataset with missing values; explain any two appropriate techniques to deal with them. 4

   c) What is the relationship between Data Mining and knowledge discovery in database ? Explain with the help of a diagram. 4

   d) For the given pair of vectors, calculate the Euclidian distance, cosine similarity, Jaccard coefficient and correlation between them. 8

      i) $x = (0, 1, 1, 1, 0, 1, 1, 0)$, $y = (1, 0, 1, 0, 1, 1, 1, 1)$

      ii) $x = (1, 0, 0, 1, 0, 0, 1, 1)$, $y = (1, 1, 0, 0, 0, 1, 0, 1)$

2. a) What is "the curse of dimensionality" ? Discuss the techniques to reduce the dimensionality. 4

   b) What is the difference between supervised and unsupervised discretization ? Give examples. 3

   c) Give the definition of correlation and covariance, and explain how to use them in data pre-processing. 3

   d) A group of 15 data records representing the age of customers is as given below. 10

      32, 64, 44, 42, 37, 32, 29, 55, 52, 31, 32, 42, 68, 24, 44

      i) Partition into three bins using equal depth approach

      ii) Partition into three bins using equal width approach

      iii) Normalize using min-max normalization by min = – 2 and max = 8

      iv) Normalize using z-score normalization.

P.T.O.

MODULE – II

3. a) What is the difference between data visualization and analytical data mining ? 3

   b) Explain the following visualizations methods. 5

   Histograms, pie charts, box plots, scatter plots, chernoff faces.

   c) What is OLAP ? How OLAP helps in data analysis ? Discuss. 6

   d) Discuss the advantages and disadvantages of sampling. Would simple random sampling without replacement be a good approach to sampling ? Discuss other alternative approaches of sampling. 6

4. a) Define a classifier and explain how to construct it when the target attribute is nominal and when the target attribute is continuous. 2

   b) Describe the following impurity metrics used to select attributes ; entropy/ gain, Gini index, and classification error. Explain how these metrics are applied to nominal and to continuous attributes. 6

   c) How does pre-pruning of decision trees work ? Explain with an example. 2

   d) Construct a decision tree for the given set of data. 10

TID	Age	Income	Student	Rating	Buys Computer
1	Young	High	No	Fair	No
2	Young	High	No	Excellent	No
3	Middle	High	No	Fair	Yes
4	Old	Medium	No	Fair	Yes
5	Old	Low	Yes	Fair	Yes
6	Old	Low	Yes	Excellent	No
7	Middle	Low	Yes	Excellent	Yes
8	Young	Medium	No	Fair	No
9	Young	Low	Yes	Fair	Yes
10	Old	Medium	Yes	Fair	Yes
11	Young	Medium	Yes	Excellent	Yes
12	Middle	Medium	No	Excellent	Yes
13	Middle	High	Yes	Fair	Yes
14	Old	Medium	No	Excellent	No

MODULE – III

5. a) Describe the steps of the sequential covering algorithm in detail. 6

   b) How are the k-nearest neighbors of a data instance found ? Discuss different voting methods and when each of them is appropriate. 5

   c) Given a dataset of n data points in m dimensions, a distance metric d, and a new data instance x, what is the time complexity of finding the k-nearest neighbors of x. Use worst case analysis. 5

   d) What do you understand by the following terms ? 4
      i) Rule based rule ordering
      ii) Class based rule ordering.

6. a) If association analysis is about finding relationships among data attributes, is calculating the correlation matrix for the set of attributes association analysis ? Why or why not ? 3

   b) Discuss the techniques for compact representation of the frequent item sets. 5

   c) Consider the following data set (Assume min_sup = 2, min_conf = 80%)

TID	Items Bought
1	{a, b, d, e}
2	{b, c, d}
3	{a, b, d, e}
4	{a, c, d, e}
5	{b, c, d, e}
6	{b, d, e}
7	{c, d}
8	{a, b, c}
9	{a, d, e}
10	{b, d}

   i) Find the frequent 3-itemsets using apriori algorithm. 4

   ii) Find the frequent 3-itemsets using FP-Growth algorithm. 4

   iii) Generate a set of strong association rules. 4

MODULE – IV

7. a) Discuss effective ways to choose appropriate initial centroids. Illustrate situations in which one way would be more appropriate than the others. 5

   b) Describe the steps of the basic agglomerative hierarchical clustering algorithm with the help of an example. 5

   c) Consider the following set of points in Euclidian space. 10

Point	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
X-Coordinate	2	2	6	5	7	6	1	4	3	5
Y-Coordinate	10	5	4	8	15	4	2	9	6	3

   Generate three clusters using K-Means algorithms. Use P1, P4 and P7 as the initial centroids.

8. a) Define each of the following approaches to anomaly detection and describe the differences between each pair : 9

      i) Model-based

      ii) Proximity-based and

      iii) Density-based techniques.

   b) Consider the case of a dataset that doesn't have labels identifying the anomalies and the task is to find how to assign a sound anomaly score, f(x), to each instance x in the dataset. Is that supervised or unsupervised anomaly detection ? Why ? 3

   c) Discuss the strength and weakness of K-means algorithm. 4

   d) What are the issues that need to be considered during anomaly detection ? 4
