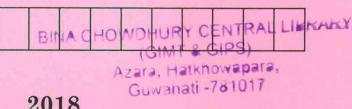Total No. of printed pages = 3

$6/6/18$

CS 131603

Roll No. of candidate

2018

B.Tech. 6th Semester End-Term Examination

DATA MINING

Full Marks – 100   Time – Three hours

The figures in the margin indicate full marks
for the questions.

Answer question No. 1 and any *six* from the rest.

1. Say whether the following statements are TRUE or FALSE :   (10 × 1 = 10)

    (a) All types of information discovery tasks are not data mining.

    (b) Data security issue is one of the key challenges of Distributed Data Mining.

    (c) Discrete attributes cannot be numeric under any circumstances.

    (d) Spatial autocorrelation must be considered with time-series data.

    (e) Timeliness of data is very important in data mining, from application point of view.

                                                                    [Turn over

(f) Variable transformation is a part of data pre-processing.

(g) In Association Rule analysis, discovery of patterns from large transaction dataset may be computationally expensive.

(h) The Apriori principle says that if an itemset is infrequent, then all of its supersets must also be infrequent.

(i) Summarization and compression are among the advantages of Cluster analysis.

(j) A classification model is useful for descriptive modeling but not for predictive modeling.

2. (a) Explain the relationship of Data Mining with Knowledge Discovery in Databases (KDD).   (5)

   (b) Write a note on the motivating challenges towards the development of Data Mining.   (5)

   (c) Distinguish between Predictive and descripting Data Mining tasks giving suitable examples. (5)

3. (a) Explain different types of attributes and their significances.   (8)

   (b) Explain the importance of discretization and binarization in data preprocessing, giving necessary examples.   (7)

4. (a) Write a note on Market-Basket data analysis and its benefits.   (5)

   (b) Describe the Apriori principle of frequent itemset generation and explain it with the help of an example.   (10)

CS 131603   2

5. (a) Explain the benefits of Hierarchical clustering algorithms over partitioning algorithms. (5)

   (b) Explain the K-means clustering algorithm along with its advantages and disadvantages. (10)

6. (a) Compare and contrast classification and Clustering giving examples. (5)

   (b) Give the Hunt's algorithm for building a Decision Tree classifier and explain the same with examples. (10)

7. (a) Explain different classes of Web Mining along with their applications. (7)

   (b) Compare and contrast Text Mining and Web Mining. (8)

8. (a) Given the Frequent Itemsets $F_1 = \{\{A\}, \{B\}, \{C\}, \{D\}\}$ $F_2 = \{\{A,B\}, \{A,C\}, \{A,D\}, \{B,C\}, \{C,D\}\}$, find the candidate Itemsets C_3 using $F_{k-1} X F_1$ method, showing step-by-step computation. (5)

   (b) Consider three data points

   X=(1,1,0,0,1,1,0,0,0,1,1), Y=(1,1,1,1,1,1,0,0,0,1,1) and Z=(1,1,0,0,1,1,1,1,1,1,1).

   Find out which pair of points are nearer among the three pairs XV, YZ and ZX, using the following distance measures

   (i) Jaccard coefficient. (5)

   (ii) Extended Jaccard coefficient. (5)

9. Write short notes on (any *three*) : (3 × 5 = 15)

   (a) Rule-based classifier.

   (b) PAM algorithm.

   (c) FP-growth algorithm

   (d) Applications of Data Mining

   (e) Data cleaning.

———————