B.Sc. TE (1-Yr) 1st Semester
B.Sc. TE (2-Yr) 3rd Semester

~~180~~ 190

13 March 2018 (Afternoon)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION — **WINTER SEMESTER, 2017-2018**

DURATION: 1 Hour 30 Minutes — **FULL MARKS: 75**

CSE 4775: Introduction to Data Mining

Programmable calculators are not allowed. Do not write anything on the question paper.

There are **4 (four)** questions. Answer any **3 (three)** of them.

Figures in the right margin indicate marks.

1. a) Describe three challenges to data mining regarding data mining methodology and user interaction issues. 10

   b) Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need 9

   c) What are the characteristics for a pattern to be interesting? 6

2. a) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order): 4x4

   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70, 82, 86.

   i. Compute the mode and the median of the data.
   ii. Give the five-number summary of the data.
   iii. Is there any outlier here? What are those?
   iv. Show a boxplot of the data.

   b) What is the difference between *Quantile* and *Quartile*? 4

   c) What is *Interquartile Range*? How IQR is used for outlier analysis? 5

3. a) "Manhattan distance and Euclidean distance are variations of Minkowski distance." – Justify this statement. 8

   b) Table 1 shows the data about the results of different tests for disease detection. All the attributes are symmetric binary. Find the dissimilarity matrix for the data of table 1. 12

Table 1: Patient report for different tests

Patients	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
P1	Positive	Negative	Positive	Negative	Positive	Positive	Negative	Positive
P2	Negative	Positive	Positive	Negative	Positive	Negative	Positive	Negative
P3	Negative	Positive	Negative	Positive	Negative	Negative	Negative	Negative

   c) What is the difference between Interval-Scaled Attributes and Ratio-Scaled Attributes? 5

4. a) There are multiple factors comprising data quality. Describe those factors in brief. 10

   b) *iFashion* is a renowned fashion brand in Bangladesh which has several outlets in different cities. They have a central database to store all the information of their sales and customers. This year they are planning to start loyalty program by giving special offers to their loyal customers. For classifying the customer they hired you to analyze *iFashion* sales and customer data. What steps will you follow to perform the task? 10

   c) What is noise in data? What are the techniques used for removing noise? 5