B.Sc. Engg. CSE 7th Semester — 8 March 2019 (Afternoon)

# ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
## ORGANISATION OF ISLAMIC COOPERATION (OIC)
## Department of Computer Science and Engineering (CSE)

**MID SEMESTER EXAMINATION** — **WINTER SEMESTER, 2018-2019**

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

# CSE 4739: 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) Bioinformatics is one of the most impactful area of Data Mining. It is the science of storing, analyzing, and utilizing information from biological data such as sequences, molecules, gene expressions, and pathways. Though it is one of the promising areas, it comes with a lot of challenges. Outline the major research challenges of data mining in Bioinformatics. 15

   b) Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Give two more examples where outlier information can be useful. 10

2. a) Use these methods to normalize the following group of data: 3x4

   235, 319, 400, 600,1000

   i. min-max normalization by setting min = -1 and max = 7.
   ii. z-score normalization
   iii. z-score normalization using the mean absolute deviation instead of standard deviation
   iv. normalization by decimal scaling

   b) What are the Data reduction strategies? Explain with appropriate examples. 9

   c) Differentiate *Interval-scaled attributes* from *Ratio-scaled attributes.* 4

3. a) Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality. 10

   b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order): 3x3

   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. Give the five-number summary of the data.
   ii. Is there any outlier here? What are those?
   iii. Show a boxplot of the data.

   c) "Manhattan distance and Euclidean distance are variations of Minkowski distance." – Justify this statement. 6

4. The following data in Table 1 are from a survey on some IUTians:

Table 1: Survey Data

| ID | Name | Age | Gender | Residential status | Income | Dept. | Behavior |
|---|---|---|---|---|---|---|---|
| 1 | X1X1 | 20 | F | Yes | 3400 | CSE | Moderate |
| 2 | X2X2 | 23 | M | No | 5800 | | Poor |
| 3 | X3X3 | 21 | M | Yes | 6000 | CSE | Good |
| 4 | X1X1 | 25 | M | No | 2500 | CSE | Very Good |

Here, the attributes types are-

- Age and Income are Numeric,
- Name is Nominal,
- Dept. is Nominal with 6 possible options,
- Gender is asymmetric binary with "M" having higher weight,
- Residential status is symmetric binary,
- Behavior is ordinal with 5 options. (Ve 1ry poor < Poor< Moderate< Good < Very Good)

a) Differentiate *Data Matrix* from *Dissimilarity Matrix*. Show the data matrix for the data of Table 1. 5

b) Find the dissimilarity matrix of the data in Table 1. 20