2-33

M.Sc. TE 2nd Semester 22 August 2019 (Afternoon)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **SUMMER SEMESTER, 2018-2019**

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

CSE 6293: Data Warehousing and 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) Briefly outline how to compute the dissimilarity between objects described by mixed attribute. 3×4

   b) What are the challenges faced during Data Integration? 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): 3×3

   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. a) Given two objects represented by the tuples (-2, 1, 42, 10) and (21, 0, -6, 10): 5×4

   i. Compute the Euclidean distance between the two objects.

   ii. Compute the Manhattan distance between the two objects.

   iii. Compute the Minkowski distance between the two objects, using h = 4.

   iv. Compute the supremum distance between the two objects.

   v. Which distance among them is the most suitable one. Justify your Answer.

   b) What are the different types of data used in Data Mining applications? 5