## 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.

- Bioinformatics is one of the most impactful area of Data Mining. It is the science of storing, 1. a) 15 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.
  - b) Outliers are often discarded as noise. However, one person's garbage could be another's 10 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.
- a) Use these methods to normalize the following group of data: 235, 319, 400, 600,1000
  - min-max normalization by setting min = -1 and max = 7. -1, -0.121, 0'725, 2.817, 7
  - z-score normalization  $\tilde{A} = 50\%$ , 6 = 272.91, -1.01, -0.702, -0.406, 0.326, 1.792ii.

  - z-score normalization using the mean absolute deviation instead of standard deviation (231-34) normalization by decimal scaling
  - b) What are the Data reduction strategies? Explain with appropriate examples. c) Differentiate Interval-scaled attributes from Ratio-scaled attributes. 4
- Data quality can be assessed in terms of several issues, including accuracy, completeness, and 3. 10 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.
  - Suppose that the data for analysis includes the attribute age. The age values for the data tuples 3x3 are (in increasing order):

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

- Give the five-number summary of the data. i.
- Is there any outlier here? What are those? ii.
- iii. Show a boxplot of the data.
- "Manhattan distance and Euclidean distance are variations of Minkowski distance." Justify this statement.
- 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
110		Age	Gender		3400	CSE	Moderate
1	X1X1	20	F	Yes		ODE	Door
2 .	X2X2	23	M·	No	5800		Poor
	ΛΖΛΖ	45	171			CSE/	Good
3	X3X3	21	M	Yes	6000		
1		25		No	2500	CSE	Very Good
1 4	X1X1	1 25	M	No	2300		



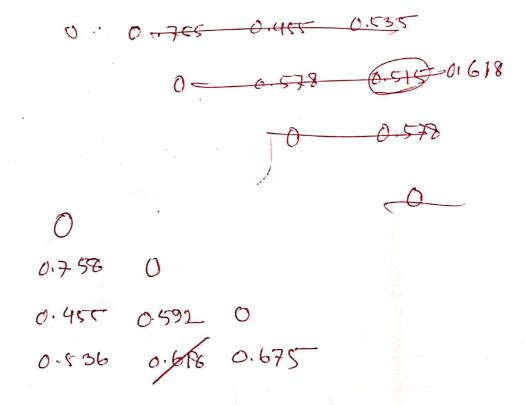
3x4

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)</li>
- a) Differentiate Data Matrix from Dissimilarity Matrix. Show the data matrix for the data of Table 1.
- b) Find the dissimilarity matrix of the data in Table 1.

20

5



0.5