B.Sc. Engg. CSE 7th Semester

25 February 2020(Afternoon)

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION

DURATION: 1 Hour 30 minutes

WINTER SEMESTER, 2019-2020

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. **"Ha Ha!"- Nelson Muntz**

Bart Simpson is student of Springfield Elementary School who is known for his mischievous, rebellious and "potentially dangerous" attitude. As per academic records Bart is an underachiever while her sister Lisa Simpson excels in study. Class test marks of Bart and Lisa are given in Table 1:

Table 1: Class test marks of Bart and Lisa

Bart	13, 15, 16, 17, 19, 20, 20, 21, 22, 25, 27, 68
Lisa	33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Draw two boxplots to compare Bart and Lisa's performance. 2×6

b) Springfield Elementary School keeps record of all the students performance. Sometimes students miss class tests and the entry is kept blank. 2+5

i. During mining what problems may arise for this practice?

ii. What are the possible solutions for this case?

c) Suppose that we plot the Inter-Quartile ranges for attribute D for the + and − classes, and we have the plot in Figure 1. Based on this plot, do you think attribute D would be an effective feature for distinguishing between the + and − class? Explain why or why not. 6

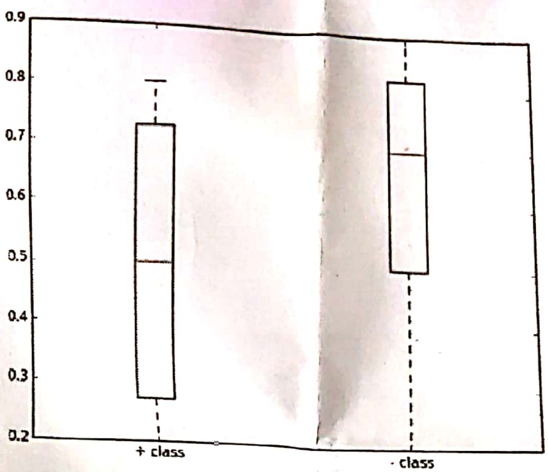

Figure 1: Plot for attribute D

2. **"Simpson, eh?"** - Charles Burns

Mr. Burns is an arrogant, billionaire who owns most of the important properties in Springfield which ranges from Nuclear power plant to University. Even the "Isotope Stadium" is one of his properties. Isotope Stadium management maintains a data warehouse consisting of the four dimensions *date*, *spectator*, *location*, and *game*, and the two measures *count* and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

a) Draw a star schema diagram for the data warehouse. 8

b) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should you perform in order to list the total charge paid by student spectators at GM Place in 2010? 7

c) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful for this warehouse and state the reasons behind your answer. 10

3. **"Thank you! Come again."** - Dr. Apu Nahasapeemapetilon
Dr. Apu is an immigrant from India and holds a Ph.D. in Computer Science. He is the owner of the famous convenience store of Springfield named "Kwik-E-Mart". He has a large customer base including most of the habitats of Springfield. Dr. Apu wants use these customer data to make further business decisions.

a) What kind of patterns can be mined from these data? Explain with example. 4+4

b) What are the characteristics make a pattern interesting? 4

c) What are the major issues this data mining process may face? 8

d) Why dimensionality reduction works? Name a few dimensionality reduction methods that can be used by Dr. Apu. 5

4. **"Do'h"** - Homer Simpson
Homer Simpson is a lazy, selfish, dim witted employee of Springfield Nuclear power plant. Whenever he works in a team, his team's performance leads to a disaster. Some of his colleagues complained to the owner Mr. Burns that they miss the performance bonus for Homer. Mr. burns made a list of all the teams that had Homer in recent past. The list is given in Table 2:

Table 2: Teams having Homer as a member

Team ID	Teams Members
T1	Lenny, Joe, Homer, Carl, Kathy
T2	Lenny, Bernie, Joe, Homer, Carl
T3	Joe, Homer, Eugene
T4	Lenny, Homer, Carl
T5	Lenny, Joe, Homer, Eugene, Carl
T6	Bernie, Joe, Homer, Eugene, Carl, Kathy

a) Find all the representative sets from the team list using apriori algorithm considering minimum support = 50% and confidence = 75% . 8

b) There is a complain- *"Homer always teams up with **Lenny and Carl**"*. Verify this complain from given data. 3

c) Though apriori algorithm increases efficiency, still it has a costly candidate generation process. How can you solve this problem? 6

d) Find all the frequent employee sets using your proposed method. 8