42

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

MID SEMESTER EXAMINATION **WINTER SEMESTER, 2022-2023**
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.
Answer **all 3 (three)** questions. Figures in the right margin indicate full marks of questions whereas corresponding CO and PO are written within parentheses.

1. a) Home values among the 8,000 homeowners of Town X are normally distributed, with a standard deviation of $11,000 and a mean of $90,000. Find the the number of homeowners in Town X whose home value is greater than $112,000. 8 (CO2) (PO1)

   b) A researcher is interested in examining the effects of a new drug on pain reports. A total of 90 participants were randomly assigned to one of three conditions (control/drug/placebo) and asked at the end of one week of treatment whether they were experiencing arthritis pain (yes/no). The results are presented in Table 1. 12 (CO2) (PO1)

Table 1: Findings (for Question 1.b)

Condition	No pain	Pain
control	6	24
drug	20	10
placebo	8	22

Conduct a chi-square test of independence on the data and report whether the condition is significantly related to the pain report. Take the level of significance at 5%. The partial chi-square distribution is shown in Table 2.

Table 2: Chi-square distribution table (for Question 1.b)

df	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	6.63
2	0.02	0.10	0.21	0.58	1.39	2.77	4.61	5.99	9.21
3	0.11	0.35	0.58	1.21	2.37	4.11	6.25	7.82	11.34
4	0.30	0.71	1.06	1.92	3.36	5.39	7.78	9.49	13.28

   c) We are generally more interested in association rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Explain why? Also specifically explain why association rules with 99% confidence may be interesting (or what might they indicate). 5 (CO1) (PO1)

2. a) Why do we need a separate Data Warehouse? Distinguish between OLTP and OLAP. 4 + 6 (CO1) (PO1)

b) What is meta-data repository? Describe the four properties of a data warehouse. 3 + 5 (CO1) (PO1)

c) It is important to measure the proximity of attributes using some similarity measure in data analysis. Equations of two of the most common similarity measures named as 'Cosine Similarity' and 'Cosine Distance' are given below. 7 (CO2) (PO1)

$$\text{Cosine Similarity}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \cdot \mathbf{y}}{|\mathbf{x}||\mathbf{y}|} \quad (1)$$

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity} \quad (2)$$

Suppose we have the dataset in Table 3. Given a new data point, $\mathbf{x}^t = (1.3, 1.5)$ as a query, rank the data points based on cosine distance in ascending order.

Table 3: Vector representation of two documents for Question 2.c

D_1	D_2
1.7	1.9
2.1	1.8
1.5	1.7
1.1	1.3
1.7	1.0

3. a) What is the curse of dimensionality? Describe a dimensionality reduction technique. 4 + 5 (CO1) (PO1)

b) Describe the property that the Apriori algorithm exploits to work efficiently. 5 (CO1) (PO1)

c) Consider the transactions of a local shop in Table 4 and construct the FP-tree (Frequent Pattern tree) and find all the frequent itemsets. 11 (CO2) (PO1)

Table 4: Shop Transactions (for Question 3.c)

TID	1	2	3	4	5	6	7	8	9	10
Items	a,b	b,c,d	a,c,d,e	a,d,e	a,b,c	a,b,c,d	a	a,b,c	a,b,d	b,c,e