B.Sc. Engg. SWE 8th Semester 20 February 2023 (Morning)

37

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **SUMMER SEMESTER, 2021-2022**
DURATION: 1 HOUR 30 MINUTES **FULL MARKS: 75**

SWE 4839: Big Data Analysis

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) 'Machine Learning (ML)' techniques are often used to find the hidden patterns from data. Despite a number of Big Data Analysis algorithms that use ML-based ideas, it is not guaranteed to provide better results in all possible scenarios. 6 (CO1) (PO2)

Design one scenario, where ML-based idea improves the situation, and one scenario where it fails to contribute at all.

b) Model a MapReduce system for Matrix-Vector Multiplication. 7 (CO3) (PO2)

c) Table 1 represents a matrix of k-shingles. Compute the minhash signature for each column using the following three hash functions: $h_1 = (2x + 1) \bmod 6$; $h_2 = (3x + 2) \bmod 6$; and $h_3 = (5x + 2) \bmod 6$. 9+3 (CO1) (PO1)

Table 1: Data for Question 1.c)

Element	**D$_1$**	**D$_2$**	**D$_3$**	**D$_4$**
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

Also, determine how close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities.

2. a) Considering the transactions mentioned in Table 2, find all the association rules with support threshold 40% and confidence threshold 75%. 15 (CO1) (PO1)

Table 2: Transactions for Question 2.a)

Transaction ID	**Items**
1	A, B, E
2	A, B, D, E
3	B, C, D, E
4	B, D, E
5	A, B, D
6	B, E
7	A, E

b) Discuss the benefits and limitations of the association rule mining algorithms with ≤ 2 passes. How can the concept of 'Negative Border' improve the scenario? Does it guarantee to grasp all the association rules having confidence higher than a threshold? 4+4 +2 (CO1) (PO1)

3. a) You need to use the *k*-means algorithm and the Euclidean distance to cluster the points (A_1 ... A_8) into three clusters. The coordinates of the points along with distance matrix based on the Euclidean distance is given in Table 3. 5+10 +5 (CO1) (PO1)

Table 3: Sample data points for Question 3.a)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
$A_1(2,10)$	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
$A_2(2,5)$		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
$A_3(8.4)$			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
$A_4(5,8)$				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
$A_5(7,5)$					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
$A_6(6,4)$						0	$\sqrt{29}$	$\sqrt{29}$
$A_7(1,2)$							0	$\sqrt{58}$
$A_8(4,9)$								0

Suppose that the initial seeds (centroids) are $A_1, A_4,$ and A_7.

i. Run the *k*-means algorithm for one iteration to assign each datapoint as a member of a cluster, and find the new centroids.

ii. Show the cluster assignments for each iteration until convergence.

iii. For each epoch, draw the original points and centroids in a graph paper and show the clusters with an approximate decision boundary,

b) Discuss the approaches of determining nearness of clusters in non-Euclidean cases. 5 (CO1) (PO1)