Assignment 2

(COMP3605 - Introduction to Data Analytics, 2023-2024)

Date Available: Thursday, October 26, 2023
Due Date: 11:50 PM, Thursday, November 09, 2023
Total Mark: 100 marks

Answer ALL Questions

INSTRUCTIONS
1. Type or write your answers neatly.
2. Show all working of your answers.
3. Your solutions must be your own. You must not share your working or solutions with your peers.
4. You are not permitted to copy, summarize, or paraphrase the work of others in your solutions.
5. Submit your answers in a single zipped file named A2_ID.zip to the email comp3605@gmail.com, where ID is replaced with your student ID. The file A2_ID.zip contains
- a single PDF file containing all of your typed, handwritten, and screenshots answers.
- a signed and dated UWI Plagiarism Declaration indicating that the work submitted is your own.

Question 1 [50 marks]
You are given the transactional data set *D* shown in the table below. The data set has six transactions. Let the minimum support (*min_sup*) count be 3.

The transactional data set *D*

TID	Items
1	J, M, S
2	J, R, S
3	G, M, R, S
4	G, J, M, R, S
5	G, M, S
6	G, M, R

Find all frequent itemsets in *D* using
a. [25 marks] the horizontal Apriori algorithm
b. [25 marks] the vertical Apriori algorithm

PLEASE TURN TO THE NEXT PAGE

Question 2 [50 marks]
You are given six two-dimensional points shown in the table below.

Point	*x* coordinate	*y* coordinate
p_1	1	1
p_2	1.5	2
p_3	4	4
p_4	5	5
p_5	6	4.6
p_6	4	3

a. [5 marks] Use the Euclidean distance to calculate the distance matrix *M* for the six points.
b. [20 marks] Show the results of the **complete linkage** version of the basic agglomerative hierarchical clustering algorithm. The distance between two clusters C_i and C_j is computed by

$$dist_{max}(C_i, C_j) = \max_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \{\|\boldsymbol{p} - \boldsymbol{p}'\|_2\}$$

where $\|\cdot\|_2$ is Euclidean distance (a.k.a. L_2-norm).

c. [25 marks] Show the results of the **group-average linkage** version of the basic agglomerative hierarchical clustering algorithm. The average distance between two clusters C_i and C_j is calculated by using the UPGMA (Unweighted Pair Group Method with Arithmetic mean) approach. That is, we have

$$dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\boldsymbol{p} \in C_i, \boldsymbol{p}' \in C_j} \| \boldsymbol{p} - \boldsymbol{p}' \|_2$$

where $n_i = |C_i|$, $n_j = |C_j|$.

Note: For each iteration of the algorithm, you need to show the found closest two clusters and the updated distance matrix *M*.

End of Assignment 2