**Homework-9**

|  |
| --- |
| Team#: \_\_\_  Team Member-1:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_  Team Member-2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_  Team Member-3:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_ |
| **Submission**   1. Answer the following questions 2. **Problem-1 and Problem-2:** Rename your doc file ““HW8\_YourTeam#” (e.g., HW8\_Team1.doc). 3. **Problem-3** Rename your R file ““HW9\_YourTeam#” HW9\_Team1.rmd/HW9\_Team1.ipynb). 4. Upload the files to Canvas. |

**Problem-1: K-Means Clustering [30 points]**

Suppose that the data mining task is to cluster the following points (with (x, y) representing location): P1(20, 100), P2(20, 50), P3(80, 40), P4(50, 80), P5(70, 50), P6(60, 40), P7(10, 20), P8(40, 90). The distance function is Euclidean distance.

Suppose initially centroids are P1, P4, and P7. Use the k-means algorithm to compute:

1. **The three clusters (K = 3) and their centers after the first iteration. Compute the total Sum of Squared Error for this iteration. [10 points]**

Answer:

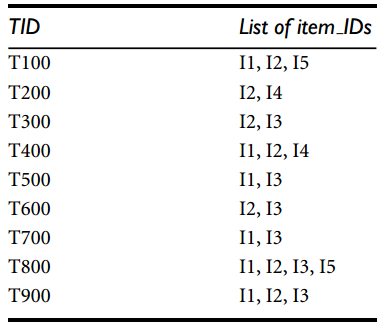
1. **The three clusters (K = 3) and their centers after the second iteration. Compute the total Sum of Squared Error for this iteration. [10 points]**

Answer:

1. **Between the first iteration and the second iteration, which one produces optimal clusters? Justify your answer. [10 points]**

Answer:

**Problem-2: Market Basket Analysis [30 points]**

The following table shows transactional data for a company.

1. **What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)? [5 points]**

Answer:

1. **What is the maximum number of frequent itemsets that can be extracted (assuming min support count >= 0)? [5 points]**

Answer:

1. **Apply Apriori algorithm to generate the candidate itemsets and frequent itemsets for a set of k-itemsets where k = 1 to 3. Min support count ≥ 2. [10 points]**

Answer:

1. **What are the association rules that can be generated from the set of the 3-itemsets calculated in the above question? Compute confidence of each rule. Given that *minconf* ≥ 70%, which rules will be considered for association mining. [10 points]**

Answer:

For Proablem-1 and Problem-2, write your solution here.

**Note:** All the computation should be presented in the step-by-step manner.

**Problem-3: R-Programming [40 points]**

A wholesale customer dataset is available in the homework folder. The dataset refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories. Your task is to apply k-means clustering to find some interesting groupings from the data. The attribute information is as per the following:

1. CHANNEL: Wholesale purchase or Retail purchase
2. REGION: Region1, Region2, or Region3
3. FRESH: annual spending (m.u.) on fresh products
4. MILK: annual spending (m.u.) on milk products
5. GROCERY: annual spending (m.u.)on grocery products
6. FROZEN: annual spending (m.u.)on frozen products
7. DETERGENTS\_PAPER: annual spending (m.u.) on detergents & paper products
8. DELICATESSEN: annual spending (m.u.)on and delicatessen products

Specifically, develop an R script to

1. Read-in the dataset. **[5 points]**
2. Apply k-means algorithm. **[15 points]**
3. Play with the k value to find some interesting groupings **[10 points]**.
4. Write a brief report to explain each of the findings **[10 points]**.