CP8210 - Assignment 2

Thakar, Viral Bankimbhai
vthakar@torontomu.ca
501213983

Gole, Montgomery
mgole@torontomu.ca
501156495

February 2023

Summary

1 Question 1: K-Means Clustering

1.1 Part A: Manual K-Means Clustering

RID	Dimension 1	Dimension 2
1	8	4
2	5	4
3	2	4
4	2	6
5	2	8
6	8	6

- **Step 1: Randomly Select $k = 3$ Initial Centroids** It is given in the question that we should select samples with RID 1, 3 and 5 as the initial set of centroids. Let's use following assignment table as reference:

Cluster ID	Centroid Dimension 1	Centroid Dimension 2
1	8	4
2	2	4
3	2	8

- **Step 2: Distance Calculation and Cluster Assignment** We will use Sum of Square Error as distance metrics.
  - **RID 2**
    - Distance between RID 2 and Cluster ID 1

      $$d_{21} = (5-8)^2 + (4-4)^2 = 9$$

    - Distance between RID 2 and Cluster ID 2

      $$d_{22} = (5-2)^2 + (4-4)^2 = 9$$

    - Distance between RID 2 and Cluster ID 3

      $$d_{23} = (5-2)^2 + (4-8)^2 = 25$$

    - As the distance $d_{21} = d_{22} = 9$ is the minimum distance we can place sample RID 2 in Cluster 1 or 2. We randomly pick to place it in Cluster 1.
  - **RID 4**
    - Distance between RID 4 and Cluster ID 1

      $$d_{41} = (2-8)^2 + (6-4)^2 = 40$$

- Distance between RID 4 and Cluster ID 2

$$d_{42} = (2-2)^2 + (6-4)^2 = 4$$

- Distance between RID 4 and Cluster ID 3

$$d_{43} = (2-2)^2 + (6-8)^2 = 4$$

- As the distance $d_{42} = d_{43} = 4$ is the minimum distance we can place sample RID 4 in Cluster 2 or 3. We randomly pick to place it in Cluster 3.

- **RID 6**
  - Distance between RID 6 and Cluster ID 1

$$d_{61} = (8-8)^2 + (6-4)^2 = 4$$

  - Distance between RID 6 and Cluster ID 2

$$d_{62} = (8-2)^2 + (6-4)^2 = 40$$

  - Distance between RID 6 and Cluster ID 3

$$d_{63} = (8-2)^2 + (6-8)^2 = 40$$

  - As the distance $d_{61} = 4$ is the minimum distance we place the sample RID 6 to Cluster 1.

At the end of Step 2 the following table represent the cluster assignment details.

Cluster ID	Centroid Dimension 1	Centroid Dimension 2	RIDs of Members
1	8	4	1, 2, 6
2	2	4	3
3	2	8	4, 5

- **Step 3: Calculate New Centroids for Each Cluster**
  - **Cluster 1**

$$Dimension\ 1 = \frac{8+5+8}{3} = 7$$

$$Dimension\ 2 = \frac{4+4+6}{3} = 4.6$$

  - **Cluster 2**

$$Dimension\ 1 = 2$$

$$Dimension\ 2 = 4$$

- **Cluster 3**

$$Dimension\ 1 = \frac{2+2}{2} = 2$$

$$Dimension\ 2 = \frac{6+8}{2} = 7$$

Updated assignment table as reference:

Cluster ID	Centroid Dimension 1	Centroid Dimension 2
1	7	4.6
2	2	4
3	2	7

- **Step 4: Distance Calculation and Cluster Assignment** We will use Sum of Square Error as distance metrics.
  - **RID 1**
    - Distance between RID 1 and Cluster ID 1

      $$d_{11} = (8-7)^2 + (4-4.6)^2 = 1.36$$

    - Distance between RID 1 and Cluster ID 2

      $$d_{12} = (8-2)^2 + (4-4)^2 = 36$$

    - Distance between RID 1 and Cluster ID 3

      $$d_{13} = (8-2)^2 + (4-7)^2 = 45$$

    - As the distance $d_{11} = 1.36$ is the minimum distance we can place sample RID 1 in Cluster 1.
  - **RID 2**
    - Distance between RID 2 and Cluster ID 1

      $$d_{21} = (5-7)^2 + (4-4.6)^2 = 4.36$$

    - Distance between RID 2 and Cluster ID 2

      $$d_{22} = (5-2)^2 + (4-4)^2 = 9$$

    - Distance between RID 2 and Cluster ID 3

      $$d_{23} = (5-2)^2 + (4-7)^2 = 18$$

    - As the distance $d_{21} = 4.36$ is the minimum distance we can place sample RID 2 in Cluster 1.
  - **RID 3**

    - Distance between RID 3 and Cluster ID 1

$$d_{31} = (2-7)^2 + (4-4.6)^2 = 25.36$$

    - Distance between RID 3 and Cluster ID 2

$$d_{32} = (2-2)^2 + (4-4)^2 = 0$$

    - Distance between RID 3 and Cluster ID 3

$$d_{33} = (2-2)^2 + (4-7)^2 = 9$$

    - As the distance $d_{32} = 0$ is the minimum distance we can place sample RID 3 in Cluster 2.

- **RID 4**
    - Distance between RID 4 and Cluster ID 1

$$d_{41} = (2-7)^2 + (6-4.6)^2 = 26.96$$

    - Distance between RID 4 and Cluster ID 2

$$d_{42} = (2-2)^2 + (6-4)^2 = 4$$

    - Distance between RID 4 and Cluster ID 3

$$d_{43} = (2-2)^2 + (6-7)^2 = 1$$

    - As the distance $d_{43} = 1$ is the minimum distance we can place sample RID 4 in Cluster 3.

- **RID 5**
    - Distance between RID 5 and Cluster ID 1

$$d_{51} = (2-7)^2 + (8-4.6)^2 = 36.56$$

    - Distance between RID 5 and Cluster ID 2

$$d_{52} = (2-2)^2 + (8-4)^2 = 16$$

    - Distance between RID 5 and Cluster ID 3

$$d_{53} = (2-2)^2 + (8-7)^2 = 1$$

    - As the distance $d_{53} = 1$ is the minimum distance we can place sample RID 5 in Cluster 3.

- **RID 6**
    - Distance between RID 6 and Cluster ID 1

$$d_{61} = (8-7)^2 + (6-4.6)^2 = 2.96$$

- ∗ Distance between RID 6 and Cluster ID 2

$$d_{62} = (8-2)^2 + (6-4)^2 = 40$$

- ∗ Distance between RID 6 and Cluster ID 3

$$d_{63} = (8-2)^2 + (6-7)^2 = 37$$

- ∗ As the distance $d_{61} = 2.96$ is the minimum distance we place the sample RID 6 to Cluster 1.

At the end of Step 4 the following table represent the cluster assignment details.

Cluster ID	Centroid Dimension 1	Centroid Dimension 2	RIDs of Members
1	7	4.6	1, 2, 6
2	2	4	3
3	2	7	4, 5

- **Step 5: Convergence** As from the previous step, we can see that our clusters are now stable and they are not changing. This state is called convergence and we can stop the iterations. Above table indicates the final cluster assignment.