K- means

CS550 Homework

Shoumya Singh
ID-19566

2. Please refer K-means example to calculate 2-cluster K-means for the following subjects .

Subject	A	B
1	1.5	1.0
2	1.0	2.0
3	2.0	3.5
4	5.0	6.0
5	3.5	4.0
6	4.5	5.0
7	2.5	4.5

Solution:

Step 1:

Data: the scores of two variables on each of seven individuals

Subject	A	B
1	1.5	1.0
2	1.0	2.0
3	2.0	3.5
4	5.0	6.0
5	3.5	4.0
6	4.5	5.0
7	2.5	4.5

Note:

- Two known information before k-means clustering
  - The data in matrix format
  - Assuming that the data set is to be grouped into 2

Step 2: Initial Partition

1. Calculate the centroid

Subject	A	B	Centroid = (A+B)/2
1	1.5	1.0	1.25
2	1.0	2.0	1.5
3	2.0	3.5	2.75
4	5.0	6.0	5.5
5	3.5	4.0	3.75
6	4.5	5.0	4.75
7	2.5	4.5	3.5

2. Find the minimum and maximum centroids
3. Let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means.

	Individual	Mean Vector (centroid)
Group 1	1	(1.5, 1.0)
Group 2	4	(5.0, 6.0)

Step 3: First clustering

Process:

1. Calculate the distance of each subject and the 2 centroids

Subject	A	B	Centroid = (A+B)/2	Distance from Centroid 1.25	Distance from Centroid 5.5
1	1.5	1.0	1.25	0	4.25
2	1.0	2.0	1.5	0.25	4.0
3	2.0	3.5	2.75	1.5	2.75
4	5.0	6.0	5.5	4.25	0
5	3.5	4.0	3.75	2.5	1.75
6	4.5	5.0	4.75	3.5	0.75
7	2.5	4.5	3.5	2.25	2.0

2. The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean.
3. The mean vector is recalculated each time a new member is added.

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.5, 2.16)	4	(5.0, 7.0)
4	1, 2, 3	(1.5, 2.16)	4, 5	(4.2, 5.0)
5	1, 2, 3	(1.5, 2.16)	4, 5, 6	(4.3, 5.0)
6	1, 2, 3	(1.5, 2.16)	4, 5, 6, 7	(3.8, 4.8)

Note:

```
1.5 = (1.5 + 1.0 + 2.0) / 3
2.16 = (1.0 + 2.0 + 3.5) / 3

3.8 = (5.0 + 3.5 + 4.5 + 2.5) / 4
4.8 = (6.0 + 4.0 + 5.0 + 4.5) / 4
```

Step 4: Check the result of the new clustering

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.5, 2.16)
Cluster 2	4, 5, 6, 7	(3.8, 4.8)

Step 5: Compare each individual's distance to the 2 clusters

But we cannot yet be sure that each individual has been assigned to the right cluster.

- So, we compare each individual's distance to its own cluster mean and to that of the opposite cluster. For example,
  - The distance between individual 1 and the centroid of Cluster 1
    - `sqrt((1.5 - 1.5)^2 + (2.16 - 1.0)^2) = 1.16`
  - The distance between individual 1 and the centroid of Cluster 2
    - `sqrt ((3.8 - 1.5)^2 + (4.8 - 1.0)^2) = 4.4`

Individual	Distance to mean (centroid) of Cluster 1: (1.5,2.16)	Distance to mean (centroid) of Cluster 2: (3.8,4.8)
1	1.16	4.4
2	0.5	3.9
3	1.4	2.2
4	5.2	1.69
5	2.7	0.85
6	4.1	0.72
7	2.5	1.33