K-Means Clustering Algorithm

K-mean clustering algorithm is an well known algorithm to segregate the data in different clusters of groups.

Algorithm : Group the m number of data present in A in K clusters.

Step 1: Randomly choose K data points as centroids.

Step 2: Find similarities between each data sample and the centroids. If the data point is found similar to a centroid then that data point is belong to the cluster corresponding to that centroid.

➢ Similarity can be measured by many formulas.

1. Correlation.
2. Euclidean Distance
3. Manhattan Distance

Here we followed Euclidean distance.

Step 3: Compute the centroids of the newly formed clusters by averaging technique.

Example

Objective: Apply K-Means algorithm to group 8 data elements in to 3 clusters, viz., cluster1, cluster2, cluster3.

1. Step 1: Store the data sets (8 data sample length 2).

A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8) A5 = (7,5), A6 = (6,4), A7 = (1,2), A8 = (4,9)

Iteration 1

2. Step 2: Choose Initial centroids (seeds) of three clusters as A1 (cluster1), A4 (cluster2) and A7 (cluster3).

Seed1 = A1 = (2,10), Seed2 = A4 = (5,8), Seed3 = A7 = (1,2),

3. Step 3: Calculate Euclidean distance between each data point with respect to the 3 seeds according to the following formula.

$$\text{d(a,b)} = \left((x_b - x_a)^2 + (y_b - y_a)^2\right)$$

1st Iteration Results

1. w.r.t A1: d(A1,seed1) = 0 d(A1,seed2) = 13 d(A1,seed3) = 65 A1 goes to cluster 1.
2. w.r.t A2: d(A2,seed1) = 25 d(A2,seed2) = 18 d(A2,seed3) = 10 A2 goes to cluster 3.
3. w.r.t A3: d(A3,seed1) = 72 d(A3,seed2) = 25 d(A3,seed3) = 53 A3 goes to cluster 2.
4. w.r.t A4: d(A4,seed1) = 13 d(A4,seed2) = 0 d(A4,seed3) = 52 A4 goes to cluster 2.
5. w.r.t A5: d(A5,seed1) = 50 d(A5,seed2) = 13 d(A5,seed3) = 45 A5 goes to cluster 2.
6. w.r.t A6: d(A6,seed1) = 52 d(A6,seed2) = 17 d(A6,seed3) = 29 A6 goes to cluster 2.
7. w.r.t A7: d(A7,seed1) = 65 d(A7,seed2) = 52 d(A7,seed3) = 0 A7 goes to cluster 3.
8. w.r.t A8: d(A8,seed1) = 5 d(A8,seed2) = 2 d(A8,seed3) = 58 A8 goes to cluster 2.

At the end of this step we have –

Cluster 1 : {A1}

Cluster 2 : {A3,A4,A5,A6,A8}

Cluster 3: {A2,A7}

Iteration 2

4. Step 4: Find the centroid (seed) of the newly formed clusters by averaging.

C1: (2,10)

C2: ((8+7+5+6+4)/5,(4+8+5+4+9)/5) = (6,6)

C3: ((2+1)/2,(5+2)/2) = (1.5,3.5)

Iteration 2

At the end of the iteration 2, we get

Cluster 1: (A1,A8)

Cluster 2: (A3,A4,A5,A6)

Cluster 3: (A2,A7) with centers

C1: (3,9.5)

C2: (6.5,5.25)

C3: (1.5,3.5)

Iteration 3

Iteration 3

At the end of the iteration 3, we get

Cluster 1: (A1,A4,A8)

Cluster 2: (A3,A5,A6)

Cluster 3: (A2,A7) with centers

C1: (3.6,6,9)

C2: (7,4.33)

C3: (1.5,3.5)