K-MEAN Clustering

Dr. Fizar Ahmed

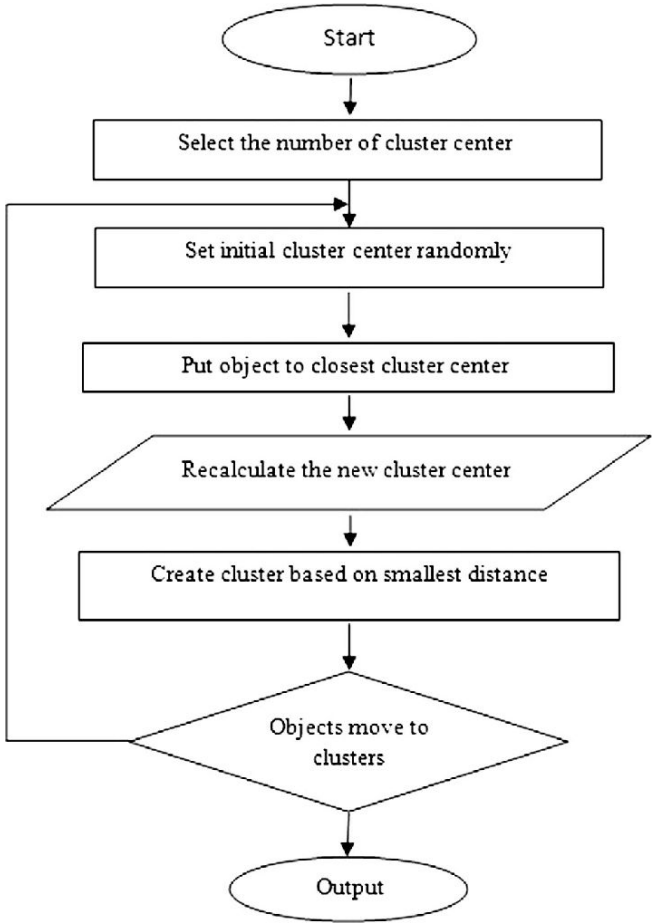
What is k-mean clustering

K-mean clustering is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

The way kmeans algorithm works is as follows:

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Flowchart of k-means clustering algorithm



Application of K-MEAN Clustering

Apply K-Mean Clustering for the following data sets for two clusters. Tabulate all the assignments.

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

- Given k = 2

Initial Centroid

Cluster	X	Y
k1	185	72
k2	170	56

Calculate Euclidean distance using the given equation.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Initial Centroid

Cluster	X	Y
k1	185	72
k2	170	56

- Calculate Euclidean distance using the given equation.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Cluster 1 (185,72)= $\sqrt{(185-185)^2+(72-72)^2} = 0$

Distance from Cluster 2 = $\sqrt{(170-185)^2+(56-72)^2}$

(170,56) = $\sqrt{(-15)^2+(-16)^2}$

= $\sqrt{255+256}$

= $\sqrt{481}$

= 21.93

Calculate Euclidean distance using the given equation.

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

Distance from Cluster 1 $= \sqrt{(185-170)^2+(72-56)^2}$

(185,72) $= \sqrt{(15)^2+(16)^2}$

$= \sqrt{255+256}$

$= \sqrt{481}$

$= 21.93$

Cluster 2 (170,56)$= \sqrt{(170-170)^2+(56-56)^2} = 0$

Cluster	Centroid		
	X	Y	ASSIGNMENT
k1	0	21.93	1
k2	21.93	0	2

- Calculate Euclidean distance for the next dataset (168,60)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 1} = \sqrt{(168-185)^2+(60-72)^2}$$

(185,72)

$$= \sqrt{(-17)^2+(-12)^2}$$

$$= \sqrt{283+144}$$

$$= \sqrt{433}$$

$$= 20.808$$

- Calculate Euclidean distance for the next dataset (168,60)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 2} = \sqrt{(168-170)^2+(60-56)^2}$$

(170,56)

$$= \sqrt{(-2)^2+(-4)^2}$$

$$= \sqrt{4+16}$$

$$= \sqrt{20}$$

$$= 4.472$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(168,60)	20.808	4.472	2

Update the cluster centroid.

Cluster	X	Y
k1	185	72
k2	= (170 + 168)/ 2 = 169	= (60+56)/ 2 = 58

- Calculate Euclidean distance for the next dataset (179,68)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 1} = \sqrt{(179-185)^2+(68-72)^2}$$

(185,72)

$$= \sqrt{(-6)^2+(-4)^2}$$

$$= \sqrt{36+16}$$

$$= \sqrt{52}$$

$$= 7.211103$$

- Calculate Euclidean distance for the next dataset (179,68)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 2} = \sqrt{(179-169)^2+(68-58)^2}$$

(169,58)

$$= \sqrt{(10)^2+(10)^2}$$

$$= \sqrt{100+100}$$

$$= \sqrt{200}$$

$$= 14.14214$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(179,68)	7.211103	14.14214	1

Update the cluster centroid.

Cluster	X	Y
k1	= 185+179/2 =182	= 72+68/2 =70
k2	169	58

- Calculate Euclidean distance for the next dataset (182,72)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 1} = \sqrt{(182-182)^2+(72-70)^2}$$

(182,70)

$$= \sqrt{(0)^2+(2)^2}$$

$$= \sqrt{0+4}$$

$$= \sqrt{4}$$

$$= 2$$

- Calculate Euclidean distance for the next dataset (182,72)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 2} = \sqrt{(182-169)^2+(72-58)^2}$$

(169,58)

$$= \sqrt{(13)^2+(14)^2}$$

$$= \sqrt{169+196}$$

$$= \sqrt{365}$$

$$= 19.10$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(182,72)	2	19.10	1

Update the cluster centroid.

Cluster	X	Y
k1	= 182+182/2 =182	= 70+72/2 = 71
k2	169	58

- Calculate Euclidean distance for the next dataset (188,77)

$$\text{Distance } [(x,y), (a,b)] = \sqrt{(x-a)^2+(x-b)^2}$$

$$\text{Distance from Cluster 1} = \sqrt{(188-182)^2+(77-71)^2}$$

(182,71)

$$= \sqrt{(6)^2+(6)^2}$$

$$= \sqrt{36+36}$$

$$= \sqrt{72}$$

$$= 8.4852$$

Dataset	Euclidean Distance		
	Cluster 1	Cluster 2	ASSIGNMENT
(188,77)	8.4852	26.87	1

Updated cluster centroid.

Cluster	X	Y
k1	= 182+188/2 = 185	= 71+77/2 = 74
k2	169	58

Final Assignment

Dataset No	X	Y	Assignment
1	185	72	1
2	170	56	2
3	168	60	2
4	179	68	1
5	182	72	1
6	188	77	1

Thank you