

Partitioning Methods ***(k -Means)***

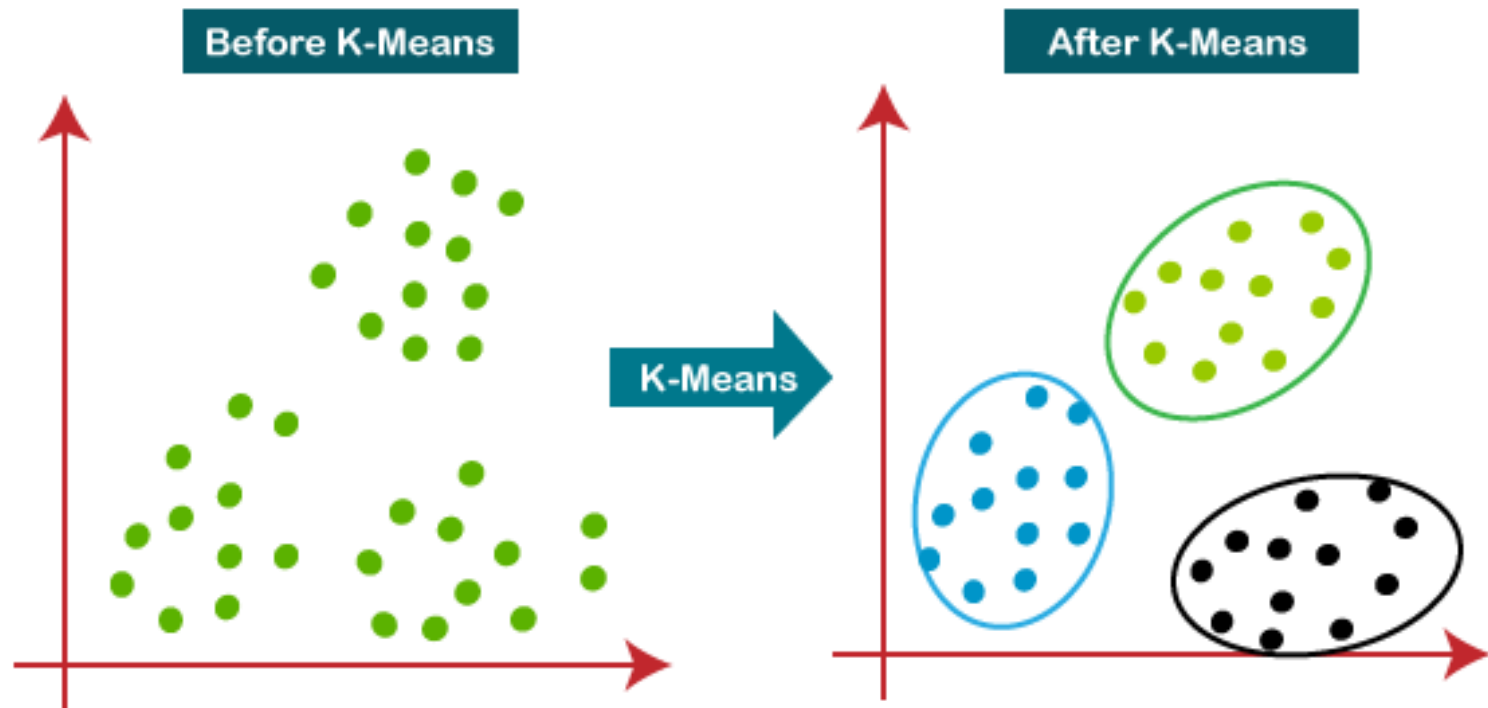
- The simplest and most fundamental version of cluster analysis is ***partitioning***, which organizes the objects of a set into several exclusive groups or clusters.
- Formally, given a data set, D , of n objects, and k , the number of clusters to form, a partitioning algorithm organizes the objects into k partitions ($k \leq n$), where each partition represents a cluster.
- The *most commonly* used partitioning methods are ***k-means*** and ***k-medoids***.

***k-means* Clustering Algorithm**

- K-Means Clustering is an *Unsupervised Learning* algorithm, which groups the ***unlabelled*** dataset into different ***clusters***. Here K defines the number of pre-defined clusters that need to be created in the process, as if $K=2$, there will be two clusters, and for $K=3$, there will be three clusters, and so on.
- It is an *iterative algorithm* that divides the ***unlabelled dataset into k different clusters*** in such a way that each data point belongs only one group that has similar properties.
- It is a *centroid-based algorithm*, where each cluster is associated with a ***centroid***.

How does the K-Means Algorithm Work?

- The algorithm takes the *unlabelled dataset* as input, divides the dataset into *k-number of clusters*, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.



K-Means Algorithm:

The working of the *K-Means algorithm* is explained in the below steps:

Step-1: Select the number **K** to decide the number of clusters.

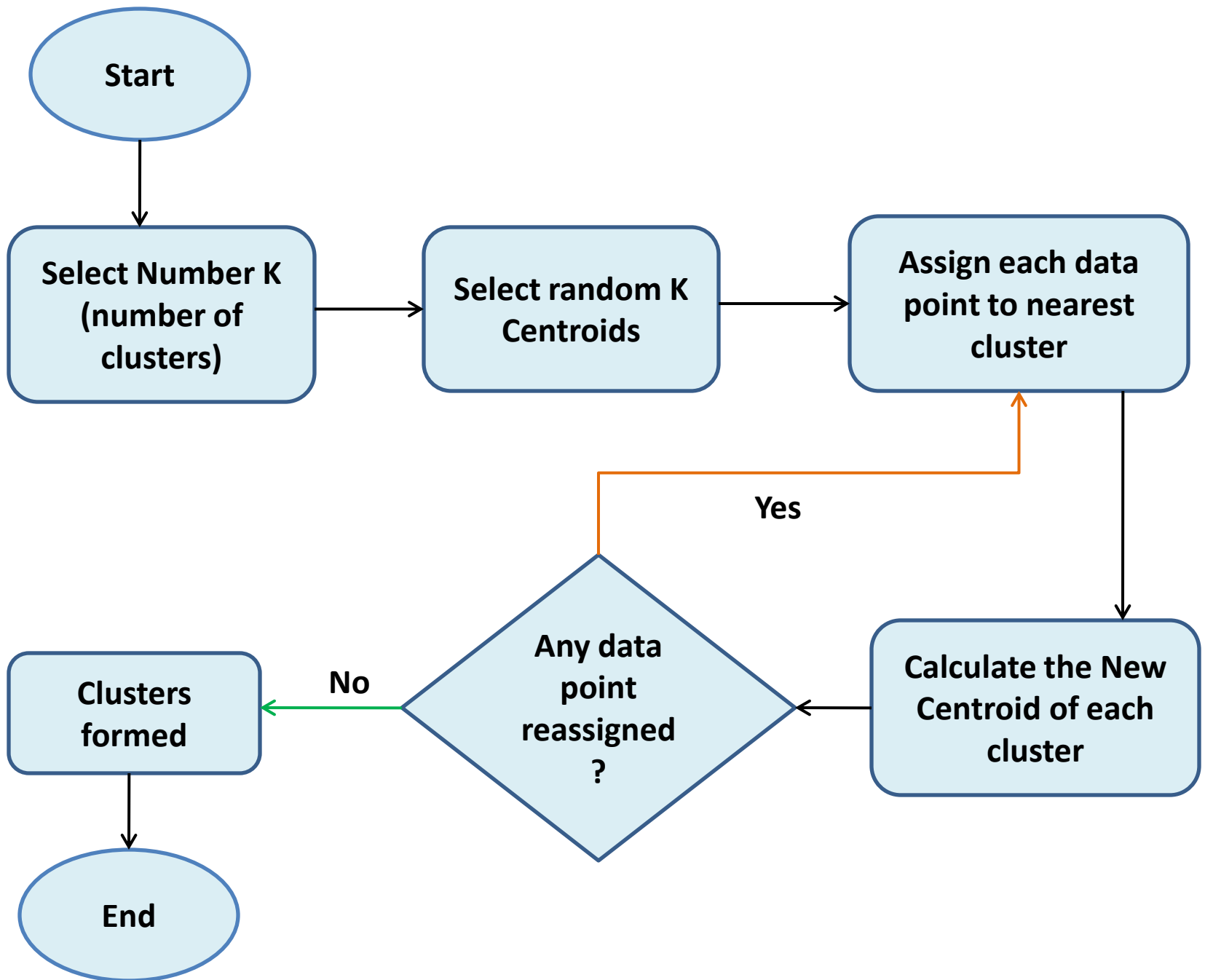
Step-2: Select random *K points or centroids*. (It can be other from the input dataset).

Step-3: Assign data point to their closest cluster centre according to the similarity or distance (Euclidean distance).

Step-4: Calculate the *centroid or mean* of all objects in each cluster.

Step-5: Repeat 3 and 4 steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.



How to Compute the Distance?

The k-means algorithm enables us to clusters or groups the data points by ***calculating the distance*** from centroids to existing data points.

- To calculate the distance we use any of the following are ***Distance Measures***.
 - ***Euclidean distance***
 - ***Manhattan distance***

Euclidean distance:

- Euclidean distance can be simply explained as the ***ordinary distance*** between two points.
- Mathematically it computes the ***root of squared differences*** between the coordinates between two objects.
- The ***Euclidean distance*** between two points or tuples, say, **$\mathbf{X} = (x_1, x_2, \dots, x_n)$** and **$\mathbf{Y} = (y_1, y_2, \dots, y_n)$** is

$$dist(X, Y) = dist(Y, X) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

Manhattan distance:

- Manhattan Distance metric is generally used when we are interested in the total distance travelled by the object instead of the displacement.
- This distance is calculated by **summing the absolute difference** between the coordinates of the points in n-dimensions.
- The **Manhattan distance** between two points or tuples, say, $\mathbf{X} = (x_1, x_2, \dots, x_n)$ and $\mathbf{Y} = (y_1, y_2, \dots, y_n)$ is

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \text{dist}(\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^n |X_i - Y_i|$$

Examples to calculate distances:

Let $X = (1, 2)$ and $Y = (3, 5)$ represent two data points.

- The ***Euclidean distance*** between the two data points is

$$\begin{aligned} \text{dist}(X, Y) &= \sqrt{(1 - 3)^2 + (2 - 5)^2} \\ &= \sqrt{(-2)^2 + (-3)^2} \\ &= \sqrt{4 + 9} \\ &= \sqrt{13} = 3.61 \end{aligned}$$

- The ***Manhattan distance*** between the two data points is

$$\begin{aligned} \text{dist}(X, Y) &= |1 - 3| + |2 - 5| \\ &= 2 + 3 = 5 \end{aligned}$$

Example: *k-means* algorithm

Suppose we want to group the visitors to a website using just their ***age*** as follows:

16, 17, 20, 21, 22, 23, 29, 36, 41, 43, 44, 45, 61, 62

Solution:

Step 1:

First, Initialize the number of clusters i.e. K value.

Let Us $K = 2$

Step 2:

Next, Select random ***K=2*** *points or centroids*.

i.e.

$$\text{centroid } (c1) = 17$$

$$\text{centroid } (c2) = 23$$

Step 3:

Now, calculate the distance from each data point to each centroid.

Let us assume ***Distance 1*** is the distances from each data point (X_i) to ***centroid 1***(C_1)

$$\text{Distance 1} = |X_i - C_1|$$

Distance 2 is the distances from each data point (X_i)
to ***centroid 2***(C_2)

$$\textbf{Distance 2} = |X_i - C_2|$$

Iteration 1:

The Initial centroids are

$$\text{centroid } (c1) = 17$$

$$\text{centroid } (c2) = 23$$

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
16	17	23	1	7	1	16.5
17	17	23	0	6	1	
20	17	23	3	3	2	37.25
21	17	23	4	2	2	
22	17	23	5	1	2	
23	17	23	6	0	2	
29	17	23	12	6	2	
36	17	23	19	13	2	
41	17	23	24	18	2	
43	17	23	26	20	2	
44	17	23	27	21	2	
45	17	23	28	22	2	
61	17	23	44	38	2	
62	17	23	45	39	2	

Step:4

The ***Updated centroids*** are

centroid (c1)

=mean (16,17)

= 16.5

centroid (c2)

= mean (20, 21, 22, 23, 29, 36, 41, 43, 44, 45, 61, 62)

= 37.25

Iteration-2

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
16	16.5	37.25	0.5	21.25	1	19.83
17	16.5	37.25	0.5	20.25	1	
20	16.5	37.25	3.5	17.25	1	
21	16.5	37.25	4.5	16.25	1	
22	16.5	37.25	5.5	15.25	1	
23	16.5	37.25	6.5	14.25	1	
29	16.5	37.25	12.5	8.25	2	45.12
36	16.5	37.25	19.5	1.25	2	
41	16.5	37.25	24.5	3.75	2	
43	16.5	37.25	26.5	5.75	2	
44	16.5	37.25	27.5	6.75	2	
45	16.5	37.25	28.5	7.75	2	
61	16.5	37.25	44.5	23.75	2	
62	16.5	37.25	45.5	24.75	2	

Step:4

The ***Updated centroids*** are

centroid (c1)

$$= \text{mean} (16, 17, 20, 21, 22, 23)$$

$$= \mathbf{19.83}$$

centroid (c2)

$$= \text{mean} (29, 36, 41, 43, 44, 45, 61, 62)$$

$$= \mathbf{45.12}$$

Iteration-3

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
16	19.83	45.12	3.83	29.12	1	21.14
17	19.83	45.12	2.83	28.12	1	
20	19.83	45.12	0.17	25.12	1	
21	19.83	45.12	1.17	24.12	1	
22	19.83	45.12	2.17	23.12	1	
23	19.83	45.12	3.17	22.12	1	
29	19.83	45.12	9.17	16.12	1	
36	19.83	45.12	16.17	9.12	2	47.42
41	19.83	45.12	21.17	4.12	2	
43	19.83	45.12	23.17	2.12	2	
44	19.83	45.12	24.17	1.12	2	
45	19.83	45.12	25.17	0.12	2	
61	19.83	45.12	41.17	15.88	2	
62	19.83	45.12	42.17	16.88	2	

Step:4

The ***Updated centroids*** are

centroid (c1)

$$= \text{mean} (16, 17, 20, 21, 22, 23, 29)$$

$$= 21.14$$

centroid (c2)

$$= \text{mean} (36, 41, 43, 44, 45, 61, 62)$$

$$= 47.42$$

Iteration-4

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
16	21.14	47.42	5.14	31.42	1	21.14
17	21.14	47.42	4.14	30.42	1	
20	21.14	47.42	1.14	27.42	1	
21	21.14	47.42	0.14	26.42	1	
22	21.14	47.42	0.86	25.42	1	
23	21.14	47.42	1.86	24.42	1	
29	21.14	47.42	7.86	18.42	1	
36	21.14	47.42	14.86	11.42	2	47.42
41	21.14	47.42	19.86	6.42	2	
43	21.14	47.42	21.86	4.42	2	
44	21.14	47.42	22.86	3.42	2	
45	21.14	47.42	23.86	2.42	2	
61	21.14	47.42	39.86	13.58	2	
62	21.14	47.42	40.86	14.58	2	

- *No change* between *iterations 3 and 4* has been noted. By using clustering, 2 groups have been identified ***16-29*** and ***36-62***.
- The resultant clusters are

Cluster 1: 16, 17, 20, 21, 22, 23, 29

Cluster 2: 36, 41, 43, 44, 45, 61, 62