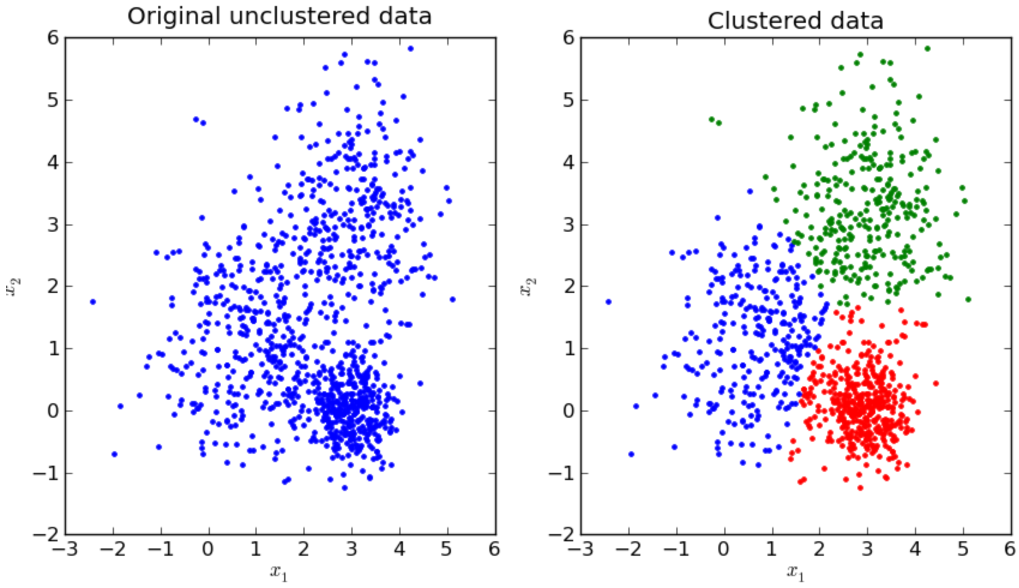
K Means Clustering

Unsupervised Learning: Unsupervised learning refers to the use of artificial intelligence (AI) algorithms to identify patterns in data sets containing data points that are neither classified nor labeled. It allows the system to identify patterns within data sets on its own. In unsupervised learning, an AI system will group unsorted information according to similarities and differences even though there are no categories provided. Unsupervised learning is a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data.

Clustering: Cluster analysis, or clustering, is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Objects in same group / cluster are more similar.


Original unclustered data

Clustered data

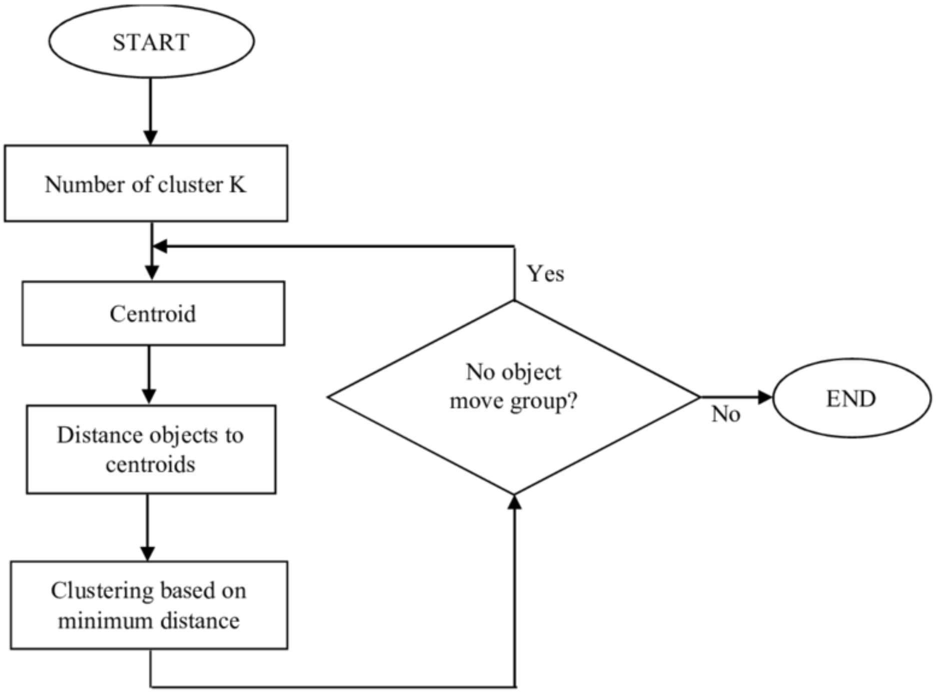
K Means Clustering: K Means Clustering is one of the simplest and most popular unsupervised machine learning algorithms. Usually, it takes k as input and partitions the set into k subsets (clusters), and thus learns to which group an individual sample belongs. 'Distance' is measured with respect to the mean value of the positions of samples in a cluster, called 'center of gravity' or 'centroid'. Comparatively lower intra-cluster 'distance' than inter-cluster.

For a given k, initially k objects are selected randomly as centroids.

Two major repeated steps:

- Data assignment step: Each data point is assigned to its nearest centroid.
- Centroid update step: Centroids are recomputed involving the current data points.

Termination criteria: No data point changes its cluster; the sum of the distances is minimized; some maximum number of iterations is reached.


START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

Example of K Means Clustering

Object	Attribute 1 (X): weight index	Attribute 2 (Y): pH value
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

We have four samples. They are (1,1), (2,1), (4,3) and (5,4). Suppose, **K = 2**. Hence, have to group the samples into **two clusters**.

Initially, randomly choose **two centroids** for **two clusters**. **Centroid $_1$ = (1,1)** for Cluster $_1$ and **Centroid $_2$ = (2,1)** for Cluster $_2$

Data assignment step:

Calculate distances for all samples from both centroids. (Formula: Euclidean Distance)

Sample	Distance from Centroid $_1$	Distance from Centroid $_2$	Determined Cluster
(1,1)	**0**	1	Cluster $_1$
(2,1)	1	**0**	Cluster $_2$
(4,3)	3.61	**2.83**	Cluster $_2$
(5,4)	5	**4.24**	Cluster $_2$

Centroid update step: Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Centroid $_1$ = $\frac{1}{1}, \frac{1}{1}$ = 1,1

Centroid $_2$ = $\frac{2+4+5}{3}, \frac{1+3+4}{3} = \frac{11}{3}, \frac{8}{3}$

Data assignment step:

Calculate distances for all samples from both centroids. (Formula: Euclidean Distance)

Sample	Distance from Centroid $_1$	Distance from Centroid $_2$	Determined Cluster
(1,1)	**0**	3.14	Cluster $_1$
(2,1)	**1**	2.36	Cluster $_1$
(4,3)	3.61	**0.47**	Cluster $_2$
(5,4)	5	**1.89**	Cluster $_2$

Centroid update step:

Centroid $_1$ = $\frac{1+2}{2}, \frac{1+1}{2}$ = 1.5, 1

Centroid $_2$ = $\frac{4+5}{2}, \frac{3+4}{2} = \frac{9}{2}, \frac{7}{2}$

Data assignment step:

Calculate distances for all samples from both centroids. (Formula: Euclidean Distance)

Sample	Distance from Centroid $_1$	Distance from Centroid $_2$	Determined Cluster
(1,1)	**0.5**	4.30	Cluster $_1$
(2,1)	**0.5**	3.54	Cluster $_1$
(4,3)	3.20	**0.71**	Cluster $_2$
(5,4)	4.61	**0.71**	Cluster $_2$

In consecutive iterations, cluster for any sample has not changed. Hence, clustering has been done successfully.

Object	**Attribute 1 (X): weight index**	**Attribute 2 (Y): pH value**	**Cluster**
Medicine A	1	1	Cluster $_1$
Medicine B	2	1	Cluster $_1$
Medicine C	4	3	Cluster $_2$
Medicine D	5	4	Cluster $_2$