

Clustering Using K-mean

Dr. Parikshit N. Mahalle

Cluster Analysis

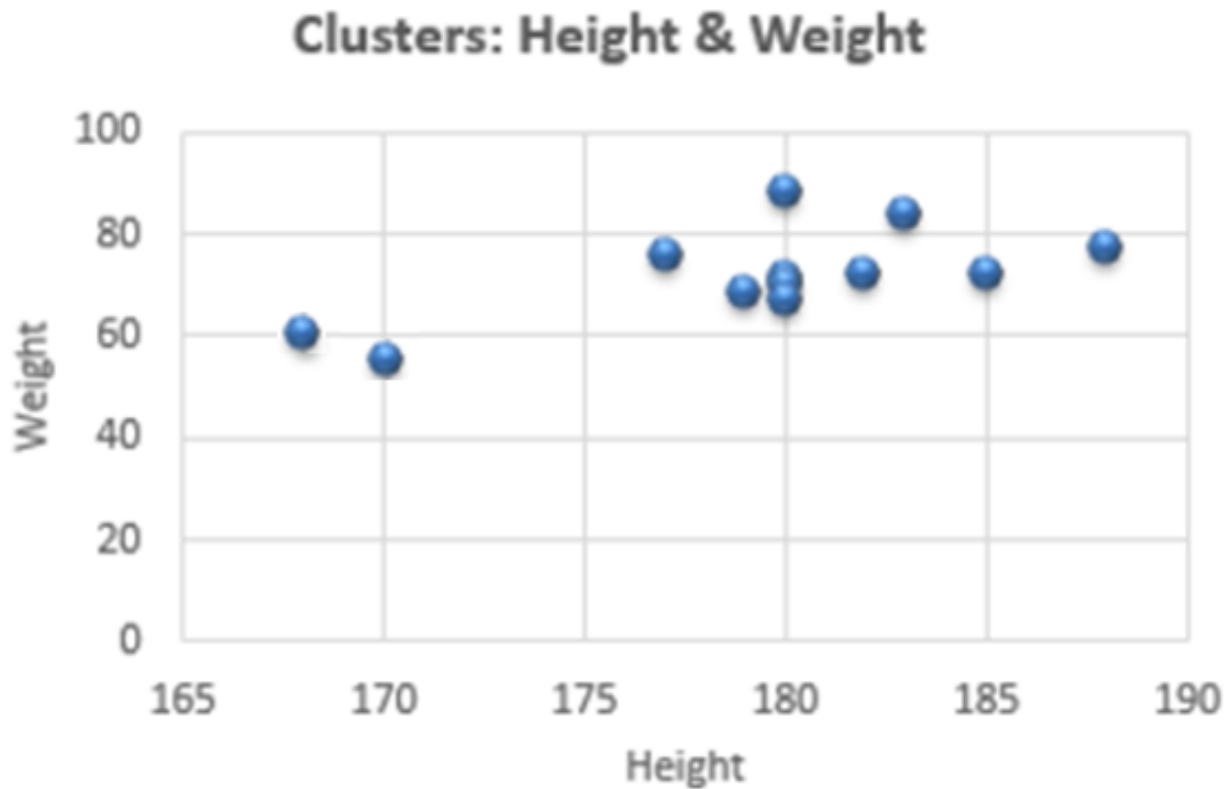
- **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).
- It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

How does K-Mean algorithm works to
Cluster Data into Different Groups?
Mathematically.....

K Means Clustering Algorithm

Data sample

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76



Plot of data sample

Data Sample

Step 1: Input

Height	Weight
185	72
170	56
168	60
179	68
182	72
188	77
180	71
180	70
183	84
180	88
180	67
177	76

Randomly Select Some Data Rows as cluster Centroids.

Step 2: Initialize cluster centroid

here $k = 2$

Cluster	Initial Centroid	
	Height	Weight
K_1	185	72
K_2	170	56

1. We Selected two rows because we are considering the value of k as 2.
2. Selecting First two rows just for easily understanding the problem.

Distance Metrics

- Euclidean distance
- Manhattan distance
- Minkowski distance

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^k |x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$

Using Euclidean Distance Formula

Step 3: Calculate Euclidean Distance (Square Root is There)

Euclidian Distance from Cluster 1	Euclidian Distance from Cluster 2	Assignment
$(185-185)^2+(72-72)^2$ =0	$(185-170)^2+(72-56)^2$ = 21.93	1
$(170-185)^2+(56-72)^2$ = 21.93	$(170-170)^2+(56-56)^2$ = 0	2

- **Step 4:** Move on to next observation and calculate Euclidean Distance

Height	Weight
168	60

Euclidean Distance from Cluster 1	Euclidean Distance from Cluster 2	Assignment
$(168-185)^2 + (60-72)^2$ =20.808	$(168-170)^2 + (60-56)^2$ = 4.472	2

- Since distance is minimum from cluster 2, so the observation is assigned to cluster 2. Now revise Cluster Centroid – mean value Height and Weight as Cluster Centroids. Addition is only to cluster 2, so centroid of cluster 2 will be updated
- Updated cluster centroids

Cluster	Updated Centroid	
	Height	Weight
K=1	185	72
K=2	$(170+168)/2$ = 169	$(56+60)/2$ = 58

- **Step 5:** Calculate Euclidean Distance for the next observation, assign next observation based on minimum euclidean distance and update the cluster centroids.

Next Observation.

Height	Weight
179	68

Euclidean Distance Calculation and Assignment

Euclidain Distance from Cluster 1	Euclidain Distance from Cluster 2	Assignment
7.211103	14.14214	1

Update Cluster Centroid

Cluster	Updated Centroid	
	Height	Weight
K=1	182	70.6667
K=2	169	58

Continue the steps until all observations are assigned

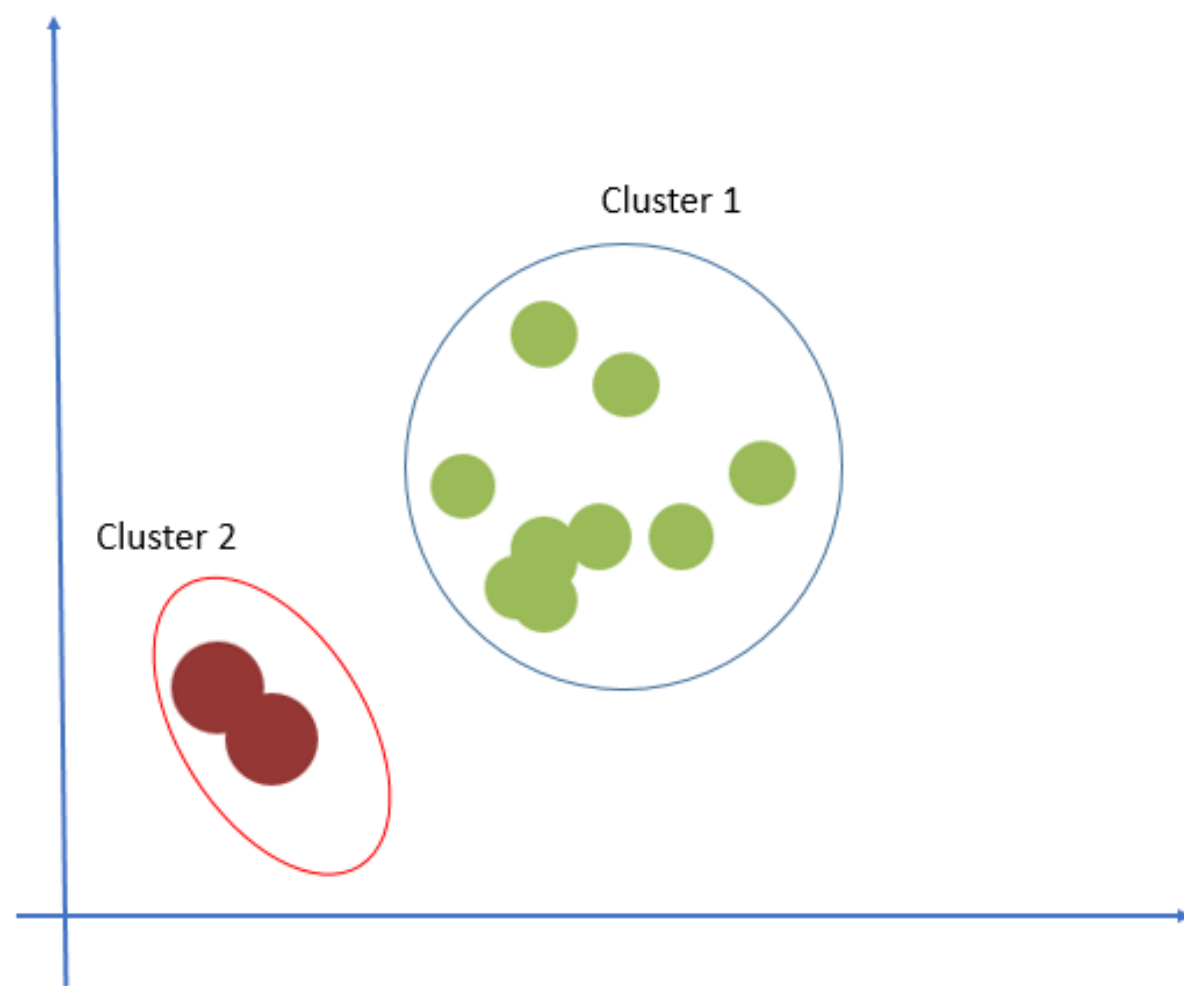
Final assignments

Height	Weight	Assignment
185	72	1
170	56	2
168	60	2
179	68	1
182	72	1
188	77	1
180	71	1
180	70	1
183	84	1
180	88	1
180	67	1
177	76	1

Cluster Centroids

Cluster	Updated Centroid	
	Height	Weight
K=1	182.8	72
K=2	169	58

This is what was expected initially based on two-dimensional plot.



K-means cluster-Algorithm

In the clustering problem, we are given a training set $x^{(1)}, \dots, x^{(m)}$, and want to group the data into a few cohesive "clusters." Here, we are given feature vectors for each data point $x^{(i)} \in \mathbb{R}^n$ as usual; but no labels $y^{(i)}$ (making this an unsupervised learning problem). Our goal is to predict k centroids **and** a label $c^{(i)}$ for each datapoint. The k-means clustering algorithm is as follows:

1. Initialize **cluster centroids** $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

For every i , set

$$c^{(i)} := \arg \min_j \|x^{(i)} - \mu_j\|^2.$$

For each j , set

$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

}

Our Example Final Outcome

