Ch:- 7 Clustering

(1) K-means clustering algorithm.

- K-means clustering is an Unsupervised learning algorithm, which groups the unlabeled dataset into different clusters.

- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

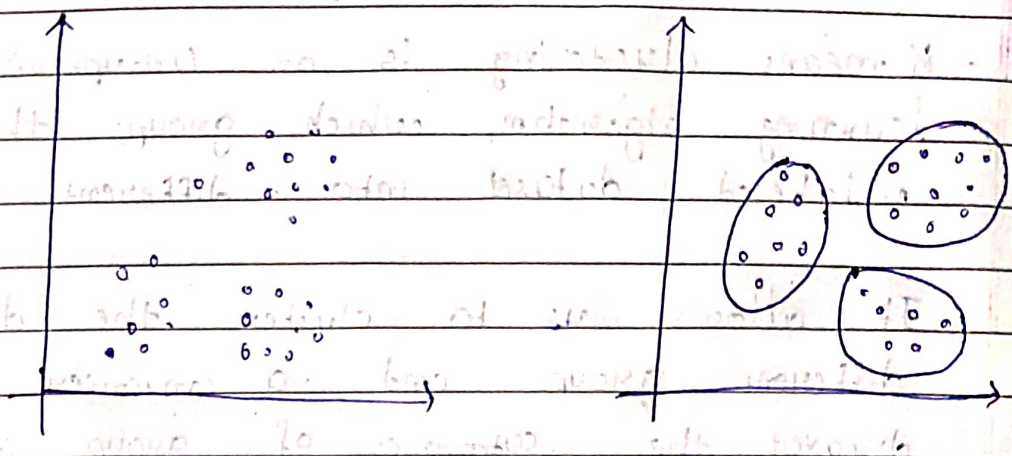
- It is a centroid-based algorithm, where each cluster is associated with a centroid.

- The algorithm takes the unlabeled 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 clustering algorithm mainly perform two tasks :-

- Determines the best value for K center points or centroids by an iterative process

- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.



Before k-Means

After k-Means

Example:

	Height	Weight		
			K_1	K_2
1	185	72		
2	170	56	(185, 72)	(170, 56)
3	168	60		
4	179	68		
5	182	72		
6	188	77		
7	180	71		
8	180	70		
9	183	84		
10	180	88		
11	180	67		
12	177	76		

K_1 - 1, 4, 5, 6, 7, 8, 9, 10, 11, 12

K_2 - 2, 3

Euclidean Distance

$$\sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

→ ED for 3rd row

$K_1 \rightarrow \sqrt{(168 - 185)^2 + (60 - 72)^2}$

$K_1 = 20.80$

$K_2 \rightarrow \sqrt{(168 - 170)^2 + (60 - 56)^2}$

$= 4.48$

New centroid calculati

$\left(\frac{168+170}{2}, \frac{60+56}{2}\right)$

$K_2 (= 169, 58)$

→ ED for 4th row

$(x_0, y_0) = (179, 68)$ $(x_c, y_c) = (185, 72)$

$K_1 = \sqrt{(179 - 185)^2 + (68 - 72)^2}$

$K_1 = 6.32$

New centroid

$\left(\frac{179+185}{2}, \frac{68+72}{2}\right)$

$K_1 = (182, 70)$

$K_2 = \sqrt{(179 - 169)^2 + (68 - 58)^2}$

$K_2 = 14.14$

Q = Hierarchical clustering

- It is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis

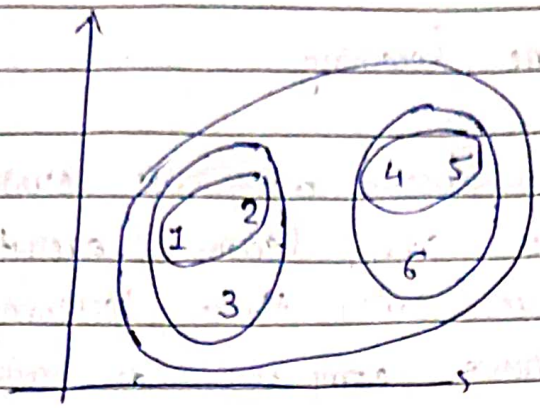
- In this algorithm, we develop the hierarchy of clusters in the form of a tree

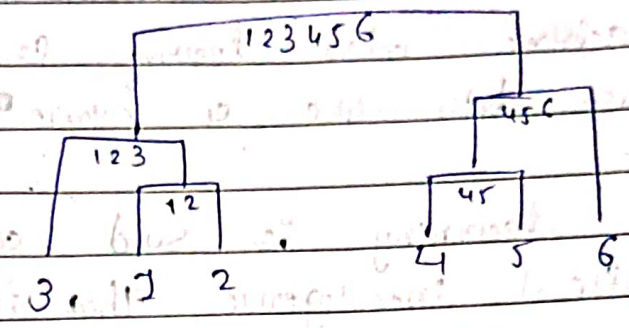
- It has two approaches :

→ Agglomerative :- It is a bottom-up approach in which the algorithm starts with taking all data points as single clusters and merging them untill one cluster is left

→ Divisive :- It is the reverse of the agglomerative algorithm as it is a top-down approach.

→ Agglomerative hierarchical clustering





→ ~~Agg~~ Divisive hierarchical clustering