K Means

→ Unsupervised method of categorising data into different clusters i.e. groups.

→ K here represents the number of groups that are to be found in data.

→ K means is a hard and Flat clustering method, what we mean by hard clustering is that every data point here is not present in multiple clusters making the cluster unique. Thus, here the clusters do not overlap each other.

→ By reason of saying that K Means is a Flat clustering method is that, its not hierarchical and every cluster is not a part of some other cluster.

How algorithim work?

→ Suppose we have a dataset. First we start taking an arbitrary number of K. Let's say K=2. Now to form two groups from set of data, the algorithim chooses two random points as centroid and computes euclidean distances from centroid to all other data points.

→ The algorithim after measuring the distance of all data points from the centroid associate with each data point with centroid based on its proximity

→ K means algorithim works in iterations as now it updates the centroid by taking the mean of all data points assigned to each centroid's cluster.

→ It repeats above mentioned process, again compute, update the position of centraid until no data points changes the cluster upon updating of the centroid. This is known as point of convergence.

Mathematical Calculation →

For example, we want to form two clusters (K=2) with a dataset having two features X and Y have 13 data points (samples).

Data points	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
X	1	2	3	4	5	6	7	8	7	8	8	9	9
Y	2	2	3	1	2	3	5	9	7	8	6	10	11

Step 1 – Now we have to select random data point as centroid.
For example, we select Datapoint 3 as Centroid 1 (Cluster label = 1) and Datapoint 7 as Centroid 2 (Cluster label = 2)
These two data points act as our intial random centroids.

Step 2 → Second step is to calculate the distance from each of these two centroid to all data points. Here we use euclidean distance as distance metric

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Euclidean distance from centroid

Data points			Centroid 1.			Centroid 2			Cluster
D	x	y	x	y	Distance from Centroid 1	x	y	Distance from Centroid 2	Assign
D1	1	2	3	3	2.2	7	5	6.7	1
D2	2	2	3	3	1.4	7	5	5.8	1
D3	3	3	3	3	0.0	7	5	4.5	1
D4	4	1	3	3	2.2	7	5	5.0	1
D5	5	2	3	3	2.2	7	5	3.6	1
D6	6	3	3	3	3.0	7	5	2.2	2
D7	7	5	3	3	4.5	7	5	0.0	2
D8	8	9	3	3	7.8	7	5	4.1	2
D9	7	7	3	3	5.7	7	5	2.0	2
D10	8	8	3	3	7.1	7	5	3.2	2
D11	8	6	3	3	5.8	7	5	1.4	2
D12	9	10	3	3	9.2	7	5	5.4	2
D13	9	11	3	3	10.0	7	5	6.3	2

→ We selected D3 (3,3) as first Centroid & D7 (7,5) as Centroid 2.

~~For~~ Now calculate euclidean distance between each data point & centroid. For example, first sample (D1), we see that distance between D1 and Centroid 1 is less (2.2) when compared to Centroid 2 (6.7). Therefore we can assign this data to Centroid 1 i.e, cluster label - 1

→ Similarly, we do this for all the samples and assign them to a cluster based on their proximity to the centroid.

Step 3 – We update the centroid which we do by taking the mean of all data points assigned to each centroid's cluster. Therefore for finding the updated Centroid 1, we take the mean of all the data points that form Cluster 1. We do the same to find Centroid 2.

Updated Centroid,	X	Y
Centroid 1 =	$\frac{1+2+3+4+5}{5} = 3$	$\frac{2+2+3+1+2}{5} = 2$
Centroid 2 =	$\frac{6+7+8+7+8+8+9+9}{8} = 7.8$	$\frac{3+5+9+7+8+6+10+11}{8} = 7.4$

Updated Centroid		
Centroid	x	y
Centroid 1	3	2
Centroid 2	7.8	7.4

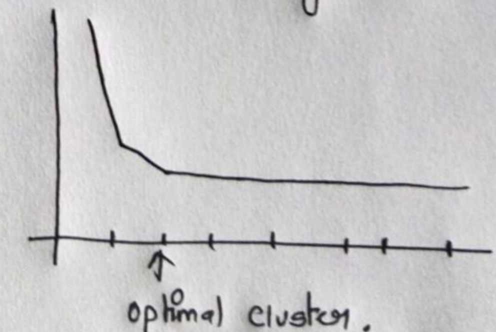
We again calculate the distance from each of these updated two centroids to all the data points and compare the distance to assign clusters to the data points.

step 4 → As mentioned earlier, algorithim keeps on reiterating the above process until it reaches the point of convergence, which means until no cluster labels are ~~reeig~~ reassigned on updating the centroid. when we calculate and compare the distances we find that we have reached the point of convergence as no clusters labels are updated anymore and hence we can stop the process and come up with final set of clusters.

Determining the number of K –

① Profiling approach → This is a method where we idenfify the characterstics of each segment and this method can be used if we have very good domain ~~language~~ knowledge. Here we take multiple values of K. We then analyse each of these clusters for various values of K and the segment that provide us with most meaningful result is chosen as final value.

② Elbow method – In this method, we compute the average distance of the data points from the centroid. With the increase in the number of centroids, the average distance between data points and their centroid decreases. We can use multiple values of K and plot them on a graph as shown below and look for the value of K where the slope decreases and average distances level out.



Other method like for validating value of K include Silhouette coefficient where the value of K provides the largest coefficient is considered.

Pre processing required for K-means Clustering →

① Outlier treatment → Because it is a distance based technique, outlier treatment is essential.

② Missing value treatment.

③ Rescaling Data → Distance based algorithm needs to be rescaled so that metric/scale remain same for all.

④ Dimensionality reduction → Reduce the number of features which are unnecessary and can make the output of Kmeans less meaningful.

→ K means is heavily dependent on number of predetermined K & its output results are hugely affected.