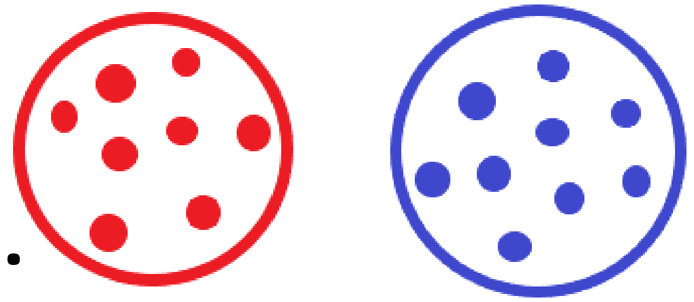
Clustering

The technique of grouping data points together. It is an assumption that data points of the same group posses similar properties or features, whereas data points from different groups will have high dissimilarity.

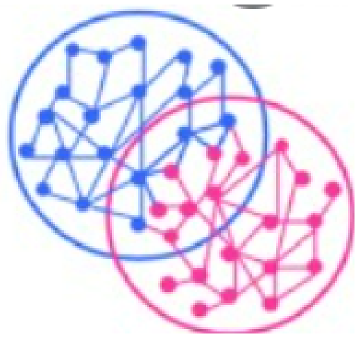
Clustering is an Unsupervised Learning Method and is commonly used for statistical data analysis in many fields. With clustering, we only try to investigate the structure of the data by grouping them.

Types of Clustering

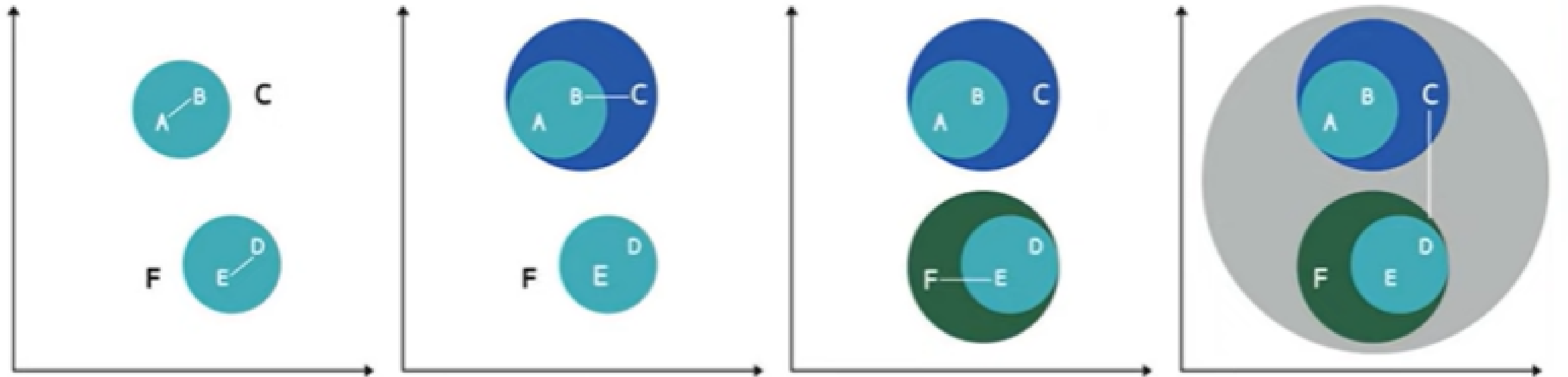
1->Exclusive Clustering: Exclusive Clustering is the hard clustering in which data point exclusively belongs to one cluster.



2-> Overlapping Clustering: Overlapping clustering is the soft cluster in which data point belongs to multiple clusters.



3-> Hierarchical Clustering: Hierarchical clustering is grouping similar objects into groups. This forms the set of clusters in which each cluster is distinct from another cluster and the objects within that each cluster is similar to each other

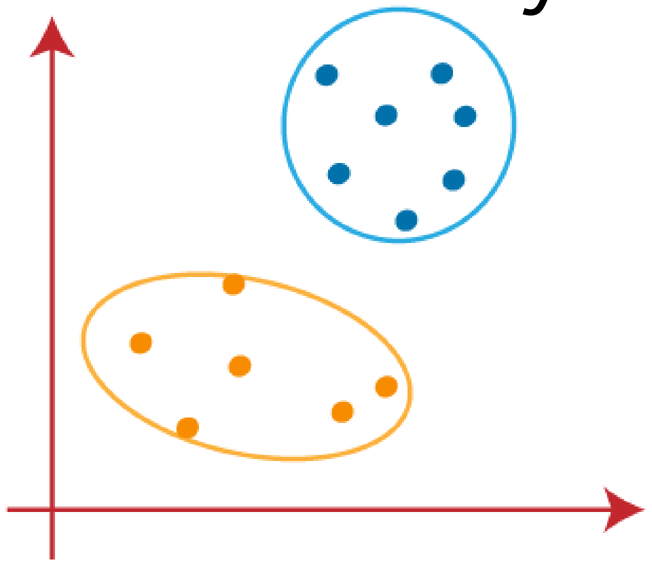


Introduction to K-Means Algorithm

K-Means Clustering: The algorithm which groups all the similar data points into a cluster is known as K-Means Clustering. This is an unsupervised machine learning algorithm. This contains no labeled data. K-Means is a centroid-based algorithm in which each group has a centroid.

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter ‘K’ in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible.



Working of K-Means Algorithm

- Step 1: First, we need to provide the number of clusters, K, that need to be generated by this algorithm.
- Step 2: Next, choose K data points at random and assign each to a cluster. Briefly, categorize the data based on the number of data points.
- Step 3: The cluster centroids will now be computed.
- Step 4: Iterate the steps below until we find the ideal centroid, which is the assigning of data points to clusters that do not vary.
- 4.1 The sum of squared distances between data points and centroids would be calculated first.
- 4.2 At this point, we need to allocate each data point to the cluster that is closest to the others (centroid).
- 4.3 Finally, compute the centroids for the clusters by averaging all of the cluster’s data points

When using the K-means algorithm, we must keep the following points in mind:

It is suggested to normalize the data while dealing with clustering algorithms such as K-Means since such algorithms employ distance-based measurement to identify the similarity between data points.

Because of the iterative nature of K-Means and the random initialization of centroids, K-Means may become stuck in a local optimum and fail to converge to the global optimum. As a result, it is advised to employ distinct centroids' initializations

K-Means Clustering Algorithm Applications

- To get relevant insights from the data we're dealing with.
- Distinct models will be created for different subgroups in a cluster-then-predict approach.
- Market segmentation
- Document Clustering
- Image segmentation
- Image compression
- Customer segmentation
- Analyzing the trend on dynamic data

Advantages and Disadvantages

- Advantage of K-Means Clustering :
- Fast and preferable to use with large datasets.
- Uses within cluster variance as a measure of similarity.
- K-Means Guarantees convergence.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

- Choosing Value of K manually.
- Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.
- K-means clustering gives varying results on different runs of an algorithm. A random choice of cluster patterns yields different clustering results resulting in inconsistency.
- Sensitive to scaling.Changing or rescaling the dataset either through normalization or standardization will completely change the final results.
- K-means algorithm can be performed in numerical data only

K-means Clustering

Steps:

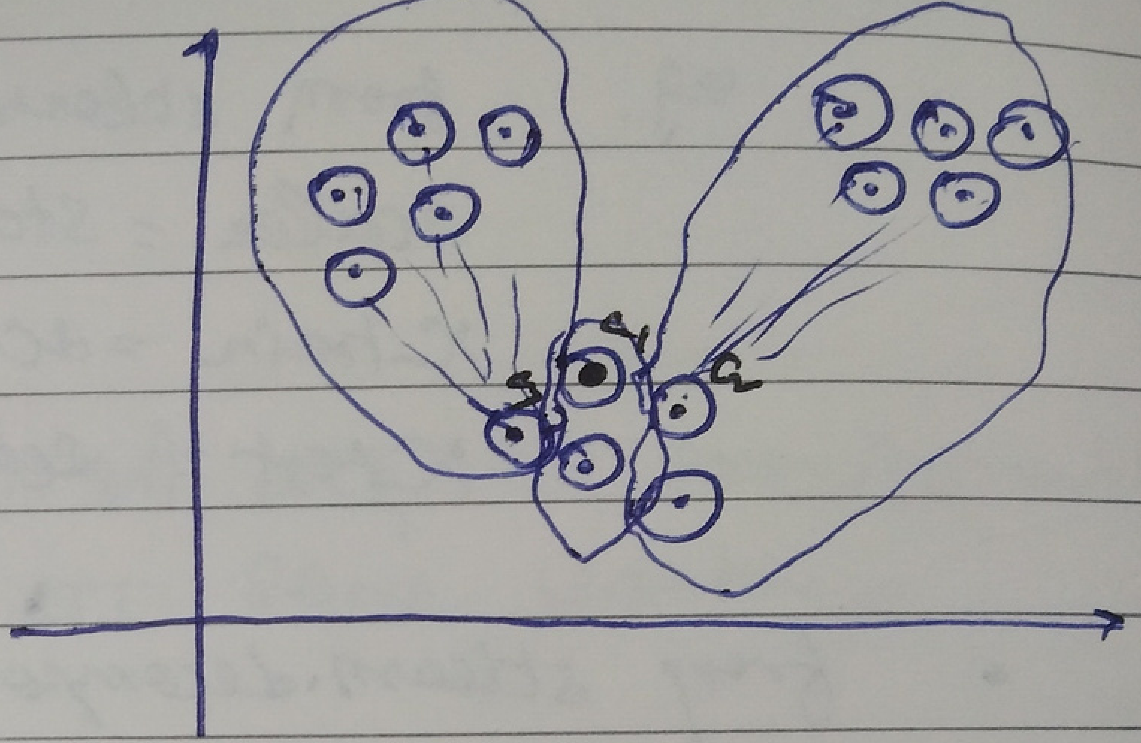
1) Decide K clusters

2) Initialize centroids

3) Assign cluster

4) Move centroids

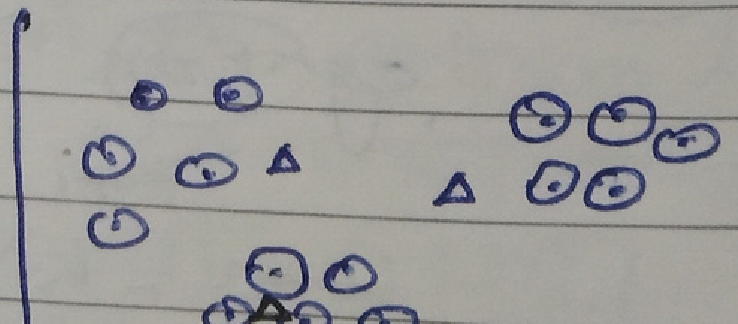
5) Finish



first take K=3 clusters then randomly take 3 points as centroid then find the distance of all other points from these three points, जिस centroid के पास होंगे ~~उस~~ 3 लेगा उसका cluster बना दें। (Eucledian distance)

Now, 3 clusters. अब ~~फिर~~ से centroid calculate करना है। By taking (mean of x, y) for each clusters

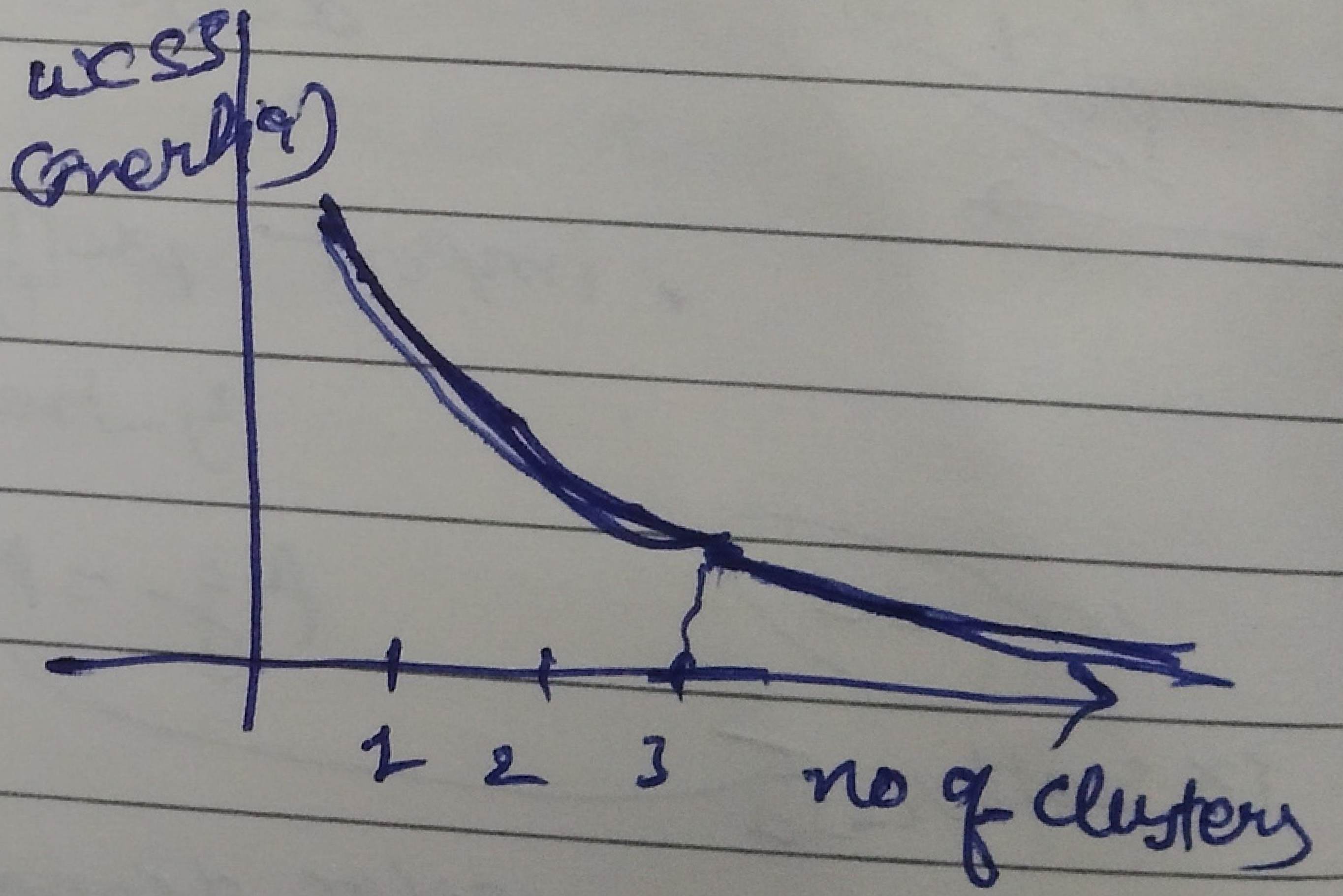
If ~~the~~ last & new centroid are same then stop, otherwise repeat



How to decide? कि कितने clusters बनेंगे ?

By **Elbow Method** ★

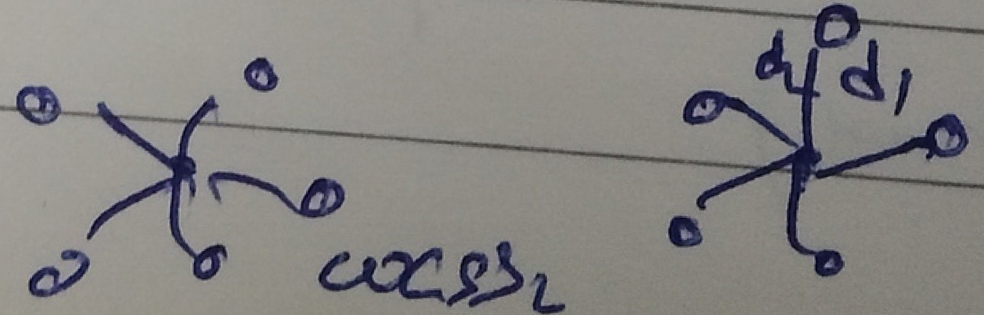
Elbow curve



★

wcss (with in cluster sum of square distance)
also known as inertia

सभी clusters के centroid के square का sum का distance



$d_1^2 + d_2^2 \dots = wcss_1$

$wcss_1 + wcss_2 \dots = wcss$

- First Assum 1 cluster → then find wcss
  for 2,3-----

say for 1 clusters $wcss_1$
-"- 2 $wcss_2$
-"- 3 $wcss_3$

$wcss_1 > wcss_2 > wcss_3$ --- True

Let total 50 points तो max 50 clusters बना सकते हैं।
so total wcss ⓪ आयेगा यदि 50 cluster बने तो
So जितने ज्यादा clusters उतना कम wcss

- Now How to find with the help of graph? By elbow point

जहां पर graph थोड़ा stablize होता जाता है।

यहाँ तक तेजी से dist reduce होता जा रहा फिर बाद में कम होता गया