Unsupervised Machine Learning

⇒ Class of machine learning technique used to find patterns in the data

⇒ The data given unsupervised learning algorithm is not labelled.

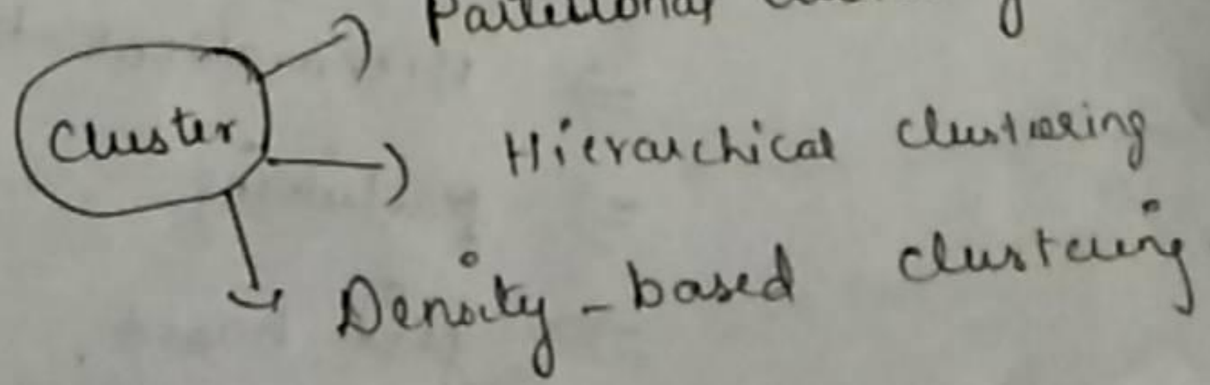
⇒ which means the only input variables (x) are given with no corresponding output variables

⇒ The most prominent methods of unsupervised learning are cluster analysis & principal Component analysis

Clustering:

The data is divided into several groups with similar traits.

Three popular categories of clustering algorithm.



Why clustering?

find hidden relationship between the datapoints in the dataset

Ex:
- Marketing: Characterstic and discover customer segment
- Biology: Classify among dif types of Plants & animals
- Library: It is used in clustering dif books on the basis of topic & information

Clustering method:

four types of clustering methods

⇒ density based
⇒ Hierarchical based
⇒ partioning
⇒ grid based.

Partitioning methods :

⇒ find mutually exclusive clusters of **Spherical** shape.

⇒ **distance** based

⇒ May use **mean or medoid** (etc)

to represent cluster center

⇒ Effective for **small to medium** size data.

K-means cluster :

The K-means algorithm defines the centroid of cluster as the mean value of the points within the cluster.

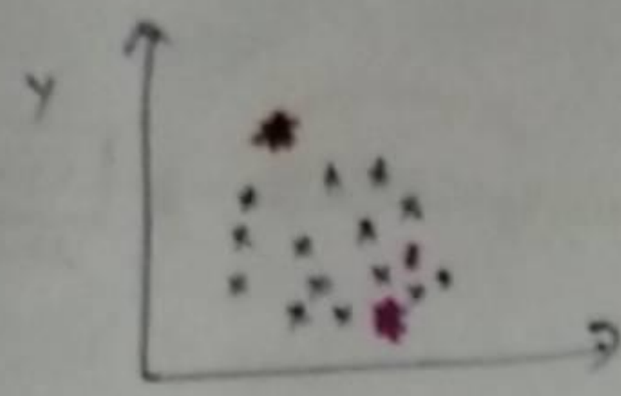
⇒ K-value = 2 (centroids)

⇒ initialize K-value (the centroid randomly)

⇒ distance find → Euclidean / manhattan

⇒ Select the group & find average

⇒ After that centroid moved again until when the point can't change.



→ This is my data points

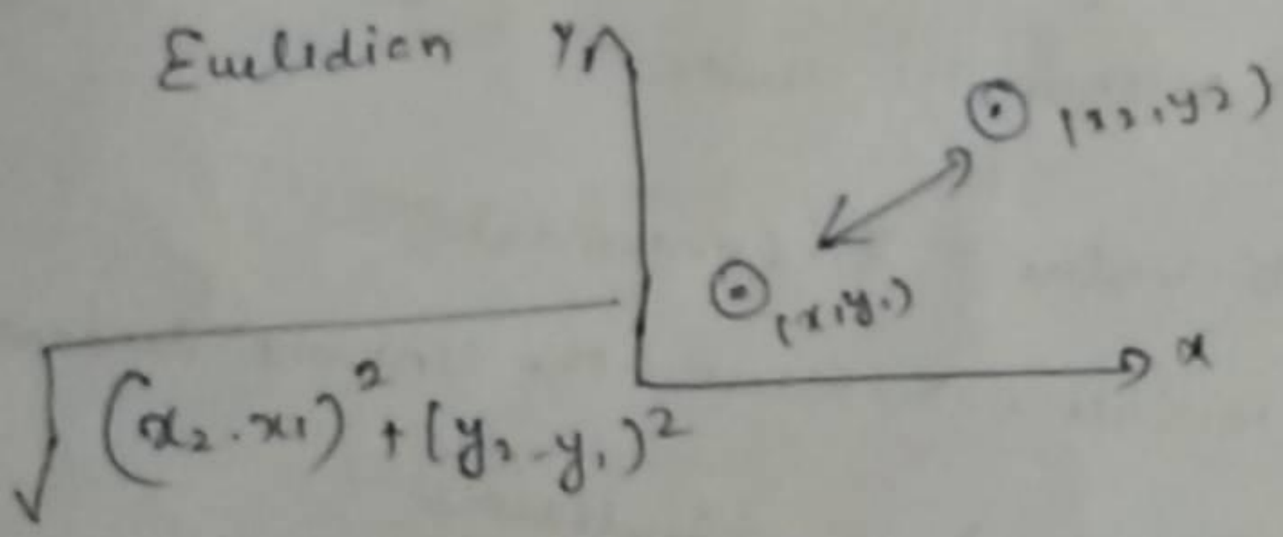
→ Randomly select centroids

→ my randomly cluster is blue, Red...

Okay, After Select random centroids we will find distance b/w actual points and Randomly Select centroids

How find distance ? use Euclidean, Manhattan...

Euclidian



How select K value?

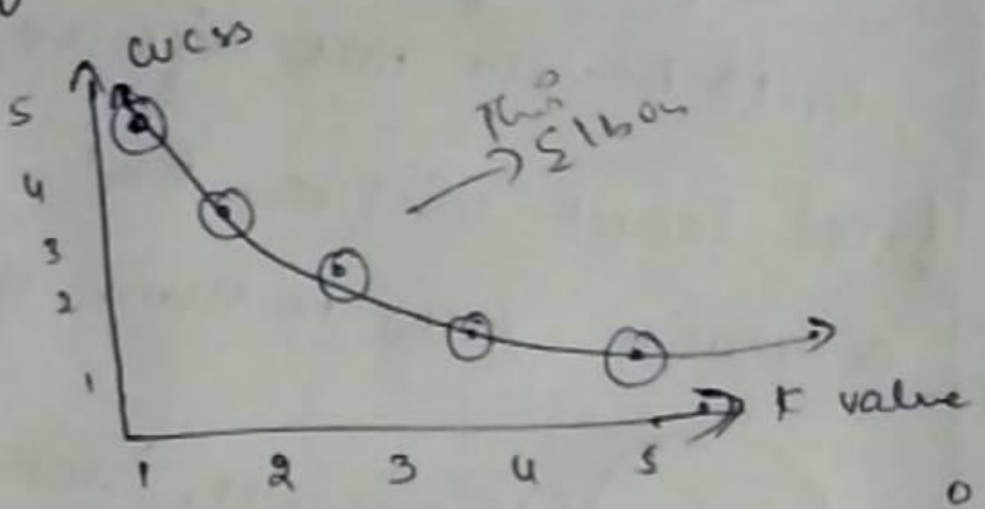
Elbow

Elbow method?

Now I am run K = 1 to 20 through looping

when my K=1 my wcss is **high**

I mean, s



Which is sudden decrease that is correct and best value of K

Okay, what is wcss

within cluster sum of square

$$\sum_{i=1}^{n} (c_i + x_i)$$

c_i → centroid

x_i → datapoints.

Hyperparameter tuning :

① n_clusters → no of clusters that we want (elbow)

② init → initial cluster centroids

③ n_init → no of times K-means alg will be run with different centroid seeds. final result will be best o/p of n-init consecutive runs in terms of inertia

④ init (K-means++, 'random', or ndarray)

K-means++ (default) → Select inital cluster in smart way to speed up convergence

Random → choose K rows at intial centroids

ndarray (n_clusters, n_features)

⑤ max_iter (maximum no of iterations of K-means algorithm for a single run)