- - -

- - - ⇒ unsupervised m

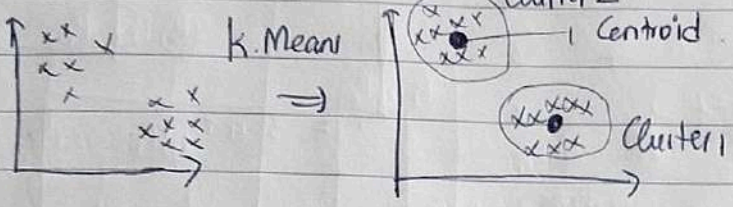
⇓

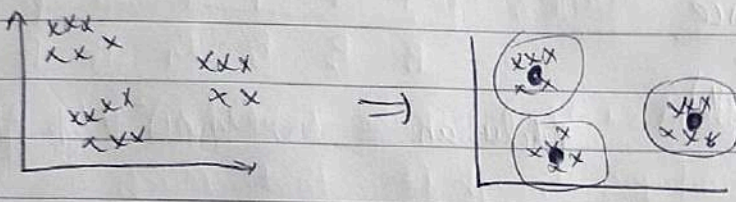
Group (or) Cluster

Unsupervised ML

1. K. Mean Algorithum
2. Hierarichal Clustering } Validate
3. DB Scan - Clustering
4. Silhoutte - Scoring

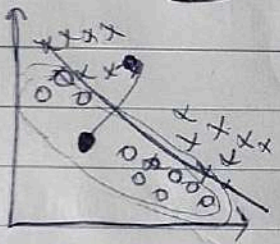
K Means Clustering Algorithum



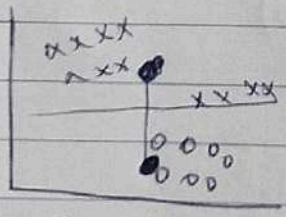


euclidian Distance
(or)
Manhanteen Distance

K=2



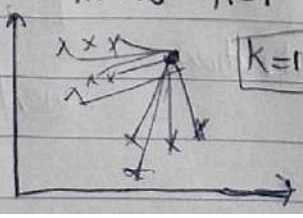
1. Initize some k → centroid
2. Points that are nearest to the centroids → Group
3. Move the centroids → Average



How do we select the k value?

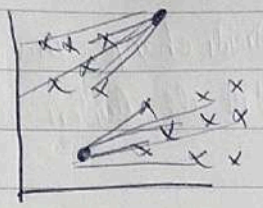
WCSS = within cluster sum of squares

Initial K=1 to 20

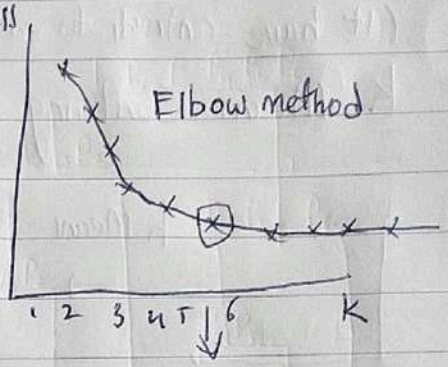


K=1 $WCSS = \sum_{i=1}^{n} (\text{distance b/w points to nearest centroid})^2$

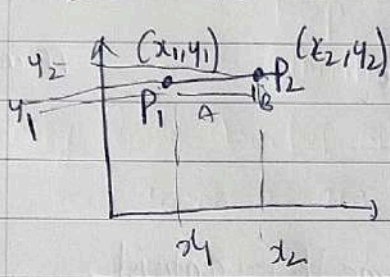
K=2



WCSS



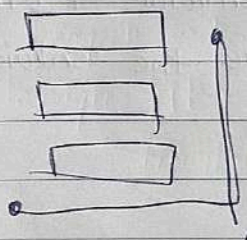
* Euclidian Distance



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

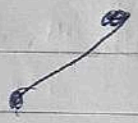
Manhatten dist $= |x_2-x_1| + |y_2-y_1|$

IRON MAN → US

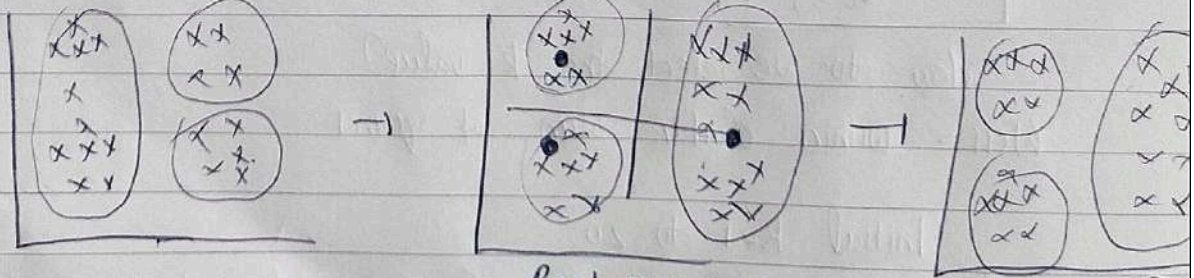


clutting

Air traffic



In this senario. K-means went wrong ..

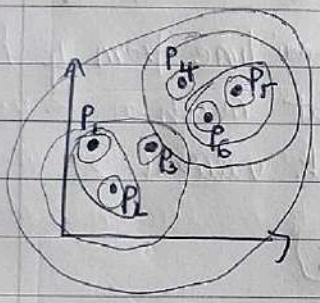


Random intization centroid → went wrong

K mean ++ → Inistization Techinque → far far distance with each other

Hierarichal Clustering Intuition

① Agglomerative } Geometric
② Divisive



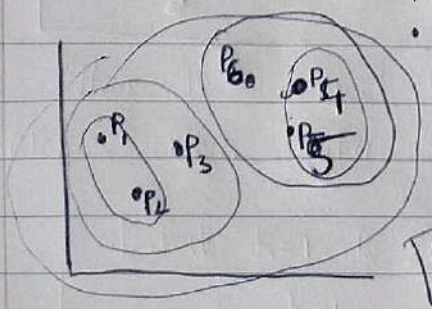
Steps:-

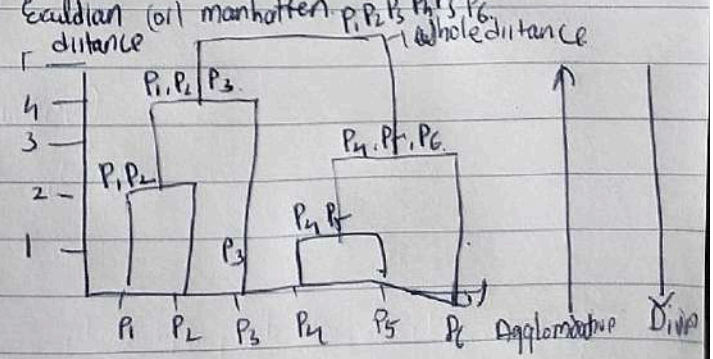
① For each data point initially will consider it as a seperate ~~clars~~ cluster. ($P_1, P_2, P_3, P_4, P_5, P_6$)

② We find the nearest point and create a new cluster

③ Keep on doing the same proces until get a single cluster.

Dendogram (we combine all cluster we have know how many cluters we have? to take) Euclidian (or) manhatten distance



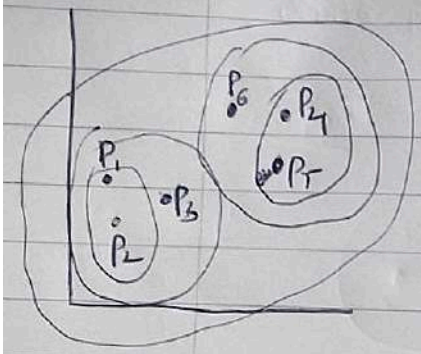


Threshold = 4

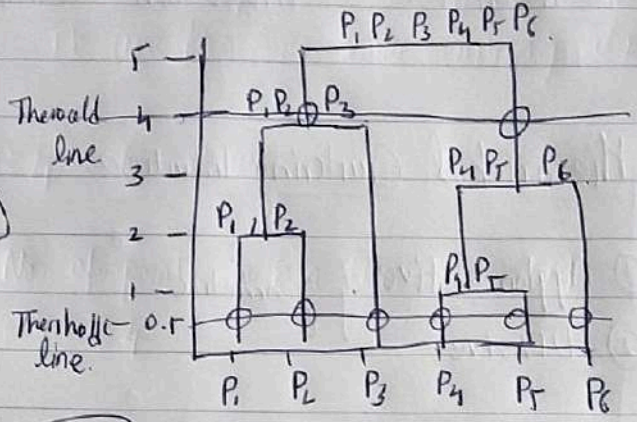
Dendogram

K = clusteer if K=2 we are having 2 clusters

Cosine similarity (or) Euclid distance



K=2



Keep Threshold ex → 4 → make line from 4 how many lines touches (or) goes through it that much as k value. see above 2 are touches so our k value is 2

if Threshold is 0.5 k value is 6 K=6.

Select the logest vertical line such that no horizontal line passes through it

Threshold (such as y axis)

K-mean Vs Hierarchal clusting

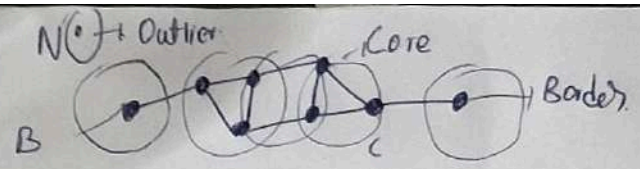
Scalability & Flexilibity

① Dataset size → Huge → K-Means.
Small → Hierenthal cluter (Dendogram onot see cla

② K-mean → Numerical data.
Hierarchal clusting → Numerical and other also (variety of data) } Important

③ Centroid → Ellow methode → No of centroids

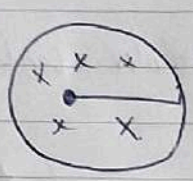
Core point
Border point
Outlier



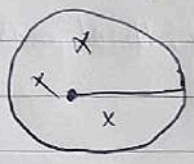
it.

DBSCAN Clustering → (Data haveing noise it can be handle will) it use for none liner clustring alg

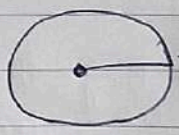
Core point → minimum=4 ∈=radius

①  → No. of points within the radius (∈) should be greter 7/4.

Border point → minimum=4.

 → No. of data points with in this radius will be less than minimum points.

Outlier

 → no other datapoints are exists