Unsupervised ML → no output

① K Means clustering

→ Main target is to group together

→ K = centroids

→ In custom ensemble technique first we apply cluster algorithm

Steps → Initialization of centers should be very far, otherwise may be inaccurate

→ K means++

① We try with different K values

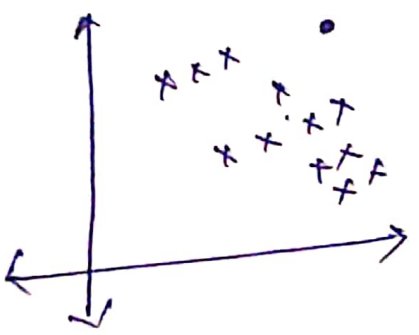
② Initialise K no. of centroids

③ Draw a straight line and ⊥ line, find the points which are near to a particular set.

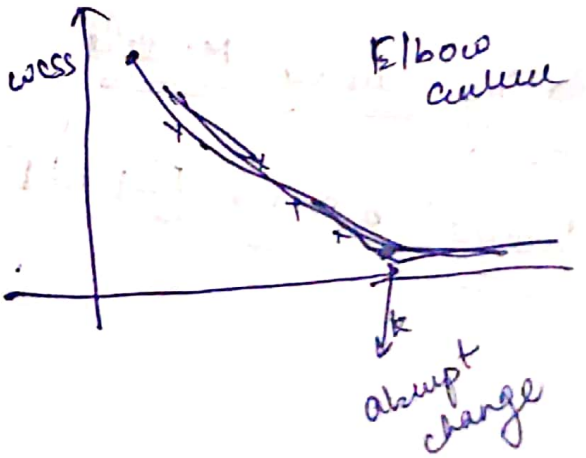
④ Compute the average of a particular set to update centroid.

For deciding K value

Elbow Method

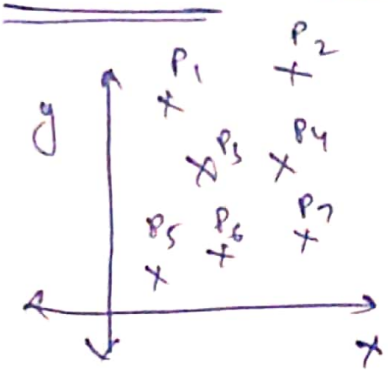


WCSS = within cluster sum of squares

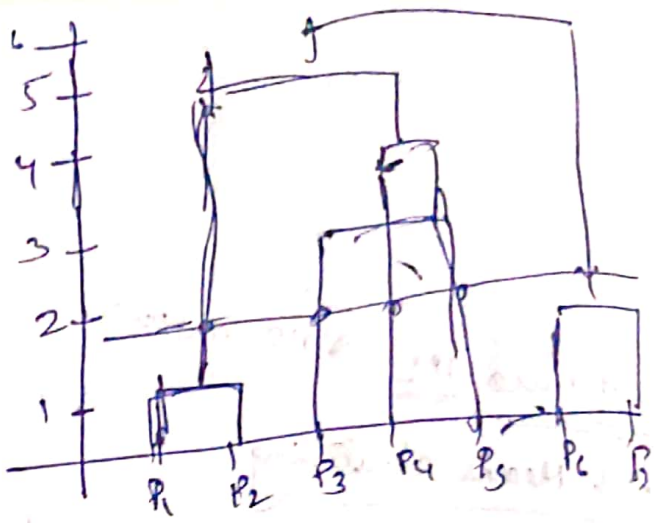


Elbow curve → complex model

② Hierarchial clustering



⇒



→ Make a dendogram

Steps

→ Join the 2 points with nearest distance

→ Repeat

No. of groups how to decide:- You need to find the longest vertical line that has no horizontal line passing through it

→ no. of intersections = no. of clusters.

→ Hierarchial clustering will take more time than K Means.

→ Hierarchial is suitable for small datasets.

Validating clustering models

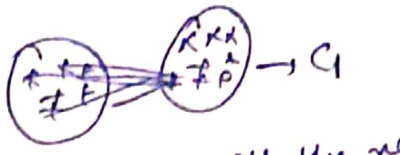
→ Silhouette score [-1 to 1]

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

i = centroid

j = remaining points

Average distance.



$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

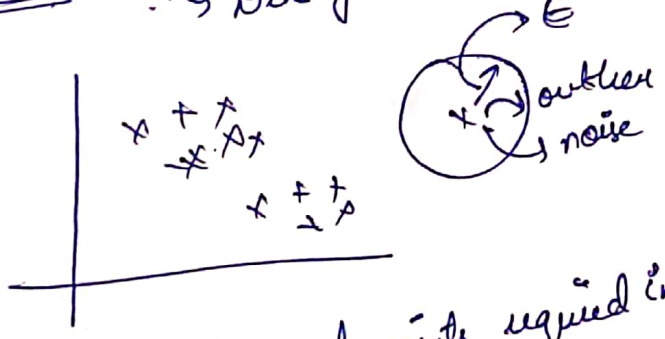
Distance between the nearest point in cluster 1 with the nearest cluster other points

→ If $b(i) >> a(i)$ then it is a good model

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

DBscan clustering [Min points, core points, Border points, Noise points]

→ Density based scan



Min points = The no. of points required inside the ϵ for that point to become a core point

→ If it has only 1 core point inside ϵ then it becomes a border point.

→ If it is a noise point, then that point can be neglected