1) K-Means Clustering (k: centroids):

A) Steps:

i) Try with diff k-values (centroids); k=3

ii) Initialize, k no. of centroids; pick 3 random centroids but far from each other, else use: KMean++

iii) Categorized all the points in a group to the nearest centroid

iv) Take average of all the 3 group respectively, i.e. 3 averages

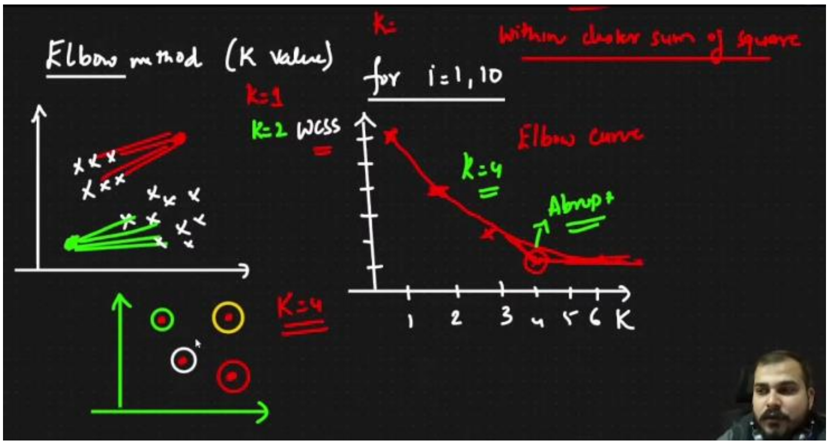
v) Compute & rearrange all the 3 centroids in the centre, in their respective group.

vi) All the pts. are there in their own location, so no more update!

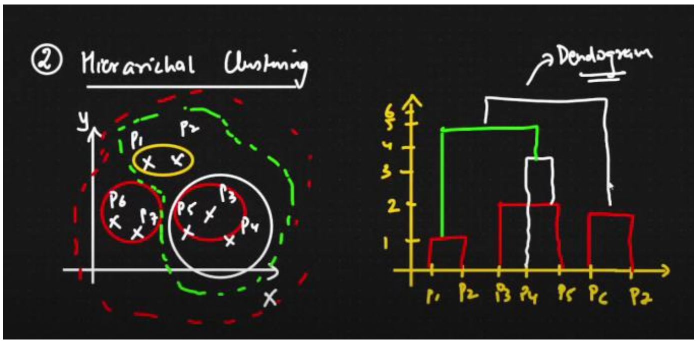
B) How to find the exact k-value?

--> Elbow method:

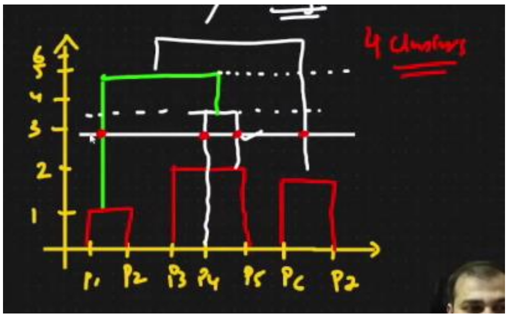
wcss: within cluster sum of square



2) Hierarichal Clustering: i) Bottom – Up, ii) Top – Down!



You need to find the longest vertical line that has no horizontal line passed through it!



3) Max time is taken by K-Means or Hierarichal clustering?

→ Hierarichal clustering!

(Dataset is small: Hierarichal clustering)

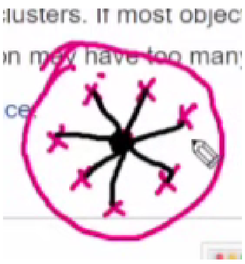
(Dataset is large: KMeans clustering)

4) How do we validate the cluster model if it's performing well or not?

→ Silhouette score (used in both KMean & Hierarichal)!

i)

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



i = centroid,

j = outer pts.

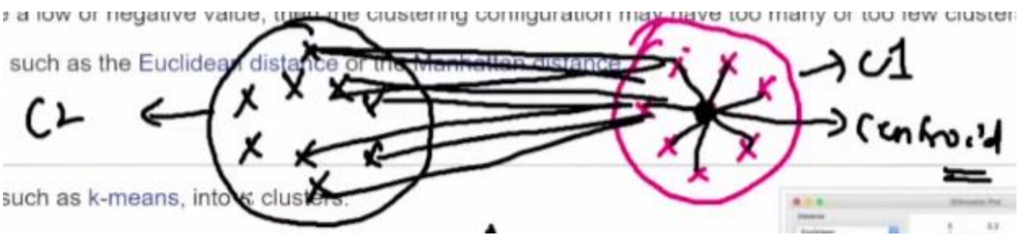
a(i):

Steps:

a) See each pts distance from the centroid, one by one!
b) Later, average it!

ii) b(i):

a) finds the nearest cluster
b) compare the distance of the pts with the nearest cluster and the earlier taken clusters, 1 pt at once with every pts of the other cluster!
For example:



a low or negative value, then the clustering configuration may have too many or too few cluster
such as the Euclidean distance or the Manhattan distance
such as k-means, into k clusters.
C2
C1
Centroid

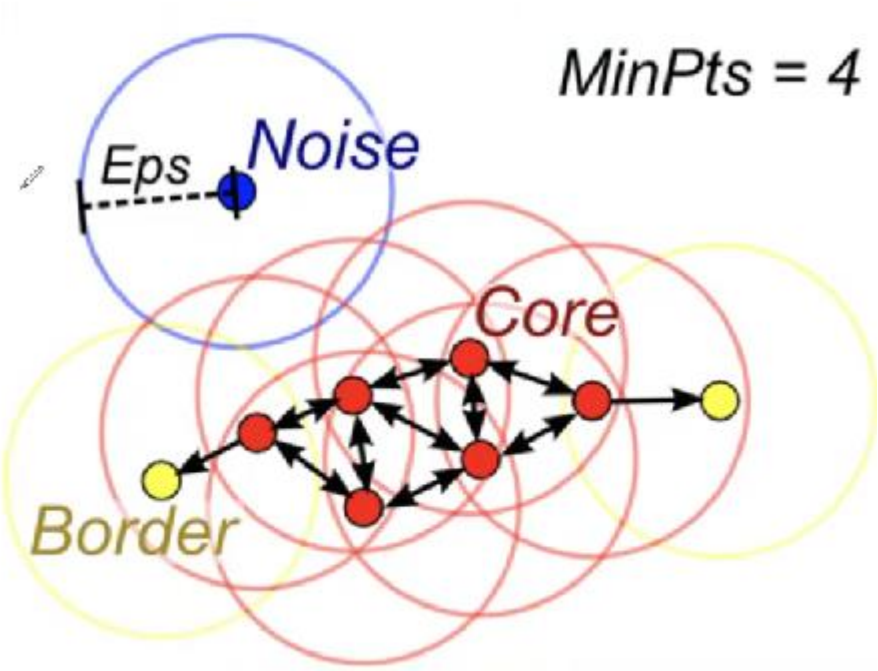
c) Later, take the average!

iii) Will a(i) >> b(i) or b(i) >> a(i), to have a good model?

➔ b(i) >> a(i)

iv) values in the silhouette clustering will be between -1 to +1,
nearer to +1: good model & vice versa!

5) DBSCAN:



i) eps: a line from which the circle is made!

ii) min pts.: 3 pts

iii) core pts.: should have at least 3 pts

iv) border pts.: should have 1 core pt

v) noise pts.: has nothing, an outlier

6) Bias & Variance:

Bias:

Training dataset:

i) performs well: low bias,

ii) not perform well: high bias!

Testing dataset:

i) good prediction: low variance,

ii) bad prediction: high variance!

Ideal scenario: Low bias & Low variance!