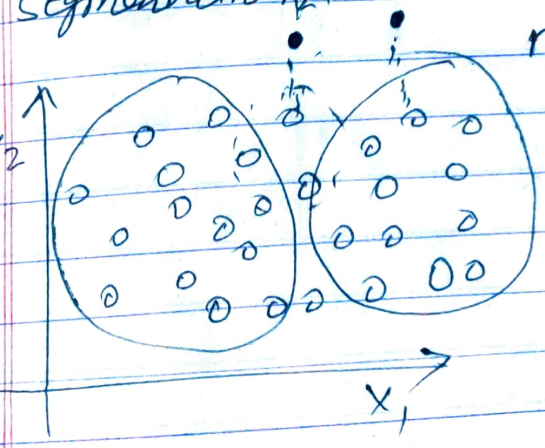
Clustering

In clustering we divide datasets in certain groups & train that dataset.
(based on similarity)

① K-Mean ② Hierarchical clustering ③ DBSCAN

K-mean

App^n - Cust. segmentation, Data Analysis, Anamoly detection, Search engine optimization, speech / image segmentation.



n groups.

WCSS
Within cluster
Sumation of Square.

We find K using WCSS.

K = 2 drop take 2 random points.

X_1	X_2

Associate each object with dataset centroid shifts such that clusters are formed.

after this if point D_1 comes it will calculate distance to centroid of each cluster which one is less that cluster D_1 is from.

HOW to take K value?

Skmean
kmean ++
Mini batch kmean.

WCSS = Within Cluster Summation of Squa

n clusters.

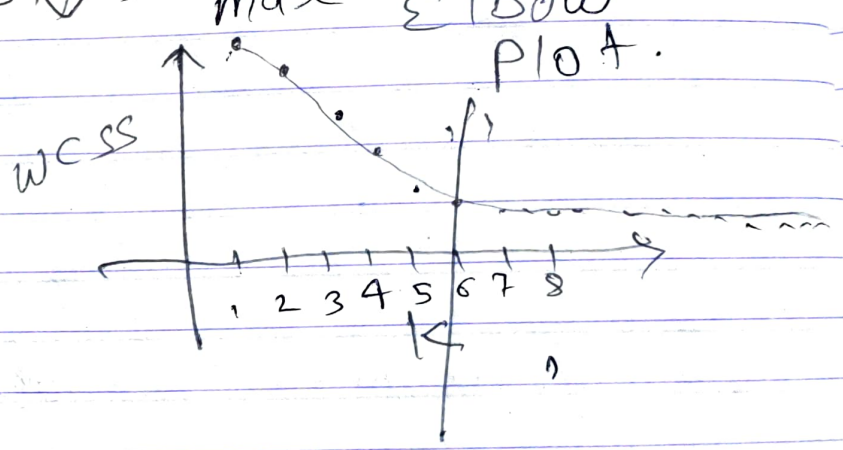
$$WCSS = \sum\sum (C_i - P_i)^2$$

↳ inertia ⟹ wcss

$K = n \Rightarrow WCSS = 0$

$K = 1 \Rightarrow WCSS = max$

$K \uparrow$ WCSS $\downarrow$.



Drawback Kmean : Allocation of centroid.

(To overcome Kmean ++)

dispersion

distance b/w two or points.

1. Always take centroid (minimum).

2. Triangle rule.

3. Provide dataset into batches (Minibatch Kmean)

* By

H

* By default Kmeans++ is used in code.

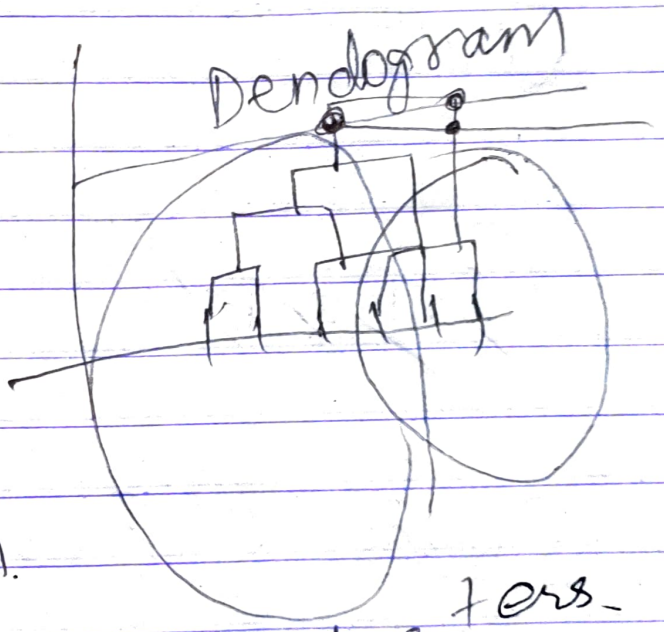
Hierarchical

- Agglomerative. – n data points
  ↳ n clusters.
- Divissive
  ↳ ex-Kmean.

→ Hierarcical clustering

Bottoms up approach.

more distance
vertical
⇒ Take that
line
for cluster
formation.
(mean of it).



2 clusters.
on dendogram
approach.

DBSCAN (Best Clustering).

Density Based Specia Spatial clustering of Appⁿ with Noise.

hyper parameter.

Epsilon - radius of circle (for a data point)

Min Points - min no. of datapoints inside circle to make cluster.

core points - point taken to form cluster.

Border points - point which is part of some core point cluster but can't form any cluster.

Noise - not inside any cluster or can't form any cluster. (outliers).

any cluster

~~Eps1 → bad~~

Drawback: (1) Eps very high, Eps very Low

(2) Min point very high.

Validation of clustering.

Clusters are supposed to be created in such a way such that it must be having very high inter classing similarity & very low intraclass similarity.

Rand index
$= \frac{\text{Total Agree}}{\text{Total Agree} + \text{Total Disagree}}$

order status1@lenovo.com.

Jaccard coefficient $= \left(\frac{SS}{SS+SD+DS}\right)$ SS → ground truth.

SS → 1
⇒ for a point agreeing to ground truth.

Entropy

Purity

Silhouette coefficient.

cohesion

seperation.

DD → 0
⇒ Point dont belong to ground truth as well as algo.

SD →

Ground Truth	Cluster Prediction: S	Cluster Prediction: D
S	SS	SD
D	DS	DD

$$\text{Rand index} = \frac{SS + DD}{SS + DS + SD + DD}$$

$$\text{Entropy} = -P_i \log P_i$$

$$= -\sum P_i \log(P_i)$$

Purity = Total % age of dataset correctly placed in cluster.

= Purity of cluster i

$P_i = \max(P_{ij})$

↳ ith class in jth cluster

$$P(\text{Cluster}) = \sum \frac{m_i}{n} P_i$$

(whole)

Not part of syllabus to learn

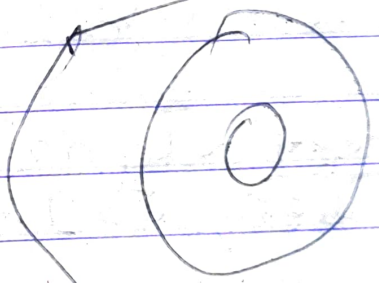
Silhoutte score $S(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$

$a(x)$ → average distance of dataset from all other pts in same cluster

(should be high.)

$b(x)$ → avg. dist. of x from all points in another cluster

$$Sc = \frac{1}{N} \sum S(x)$$



$a(x) \not> b(x)$

wrong