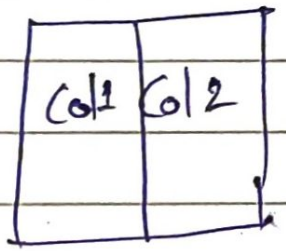
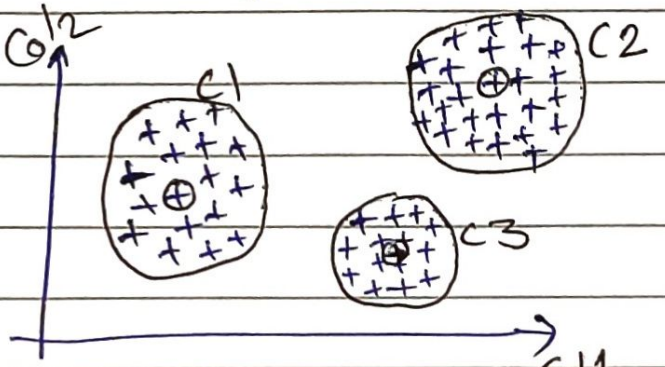
K-Means Clustering

Group them into "K" clusters

"K" no. of clusters.

For every cluster it will assign Centroid (to each cluster)





$C_1, C_2, C_3 \rightarrow$ clusters

Set of points in $C_1 \rightarrow S_1$

" " " " $C_2 \rightarrow S_2$

" " " " $C_3 \rightarrow S_3$

*) $D = S_1 \cup S_2 \cup S_3 \rightarrow$ This gives my Complete Data

D → Dataset

*) $S_1 \cap S_2 = \phi$
$S_2 \cap S_3 = \phi$
$S_1 \cap S_3 = \phi$
} Properly clustered → each datapoint belongs to only 1 cluster
→ not even single datapoint should belongs to 2 clusters

↓ Mutually exclusive.

K-Means

↳ K- Centroids

↳ K- Set of points $(S_1, S_2, \cdots S_K)$ → set1, set2, → set K

↳ $S_1 \cup S_2 \cup S_3 \cup S_4 \cdots \cup S_K = D$ → Dataset

↳ Mutually exclusive sets $(S_1, S_2, \cdots S_K)$

↳ $S_1 \cap S_2 = \phi$

$S_2 \cap S_3 = \phi$

$S_1 \cap S_3 = \phi$

Q) How to find Centroid.

Ans- $C_i = \frac{\text{Sum of Datapoints}}{\text{\# of Datapoints}} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

↓

nothing but Mean $C_i = \frac{\sum_{j=1}^{n} x_j}{n}$

↓

$C_i = \frac{1}{n} \sum_{x_j \in S_i} x_j$ → datapoints in Set 1.

↓

size of S_i

$n = |S_i|$

$$C_i = \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j$$

We can also use Median as well

↓

we are using Mean so, it will be impacted by outlier. Centroid

ML

Supervised

Regression Classification

Unsupervised Learning

Clustering

Hard Clustering

→ K means
→ Hierarchical

Soft clustering

GMM

Gaussian Mixture Model.

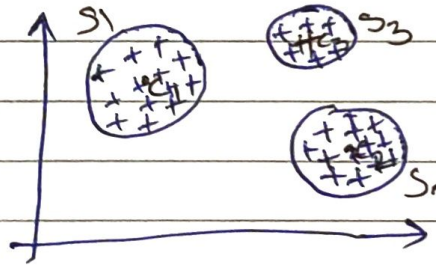
a point should belongs to only 1 cluster.

↳ It means intersection blw two clusters = ϕ

↳ $S_1 \cap S_2 = \phi$
↳ $S_2 \cap S_3 = \phi$
↳ $S_1 \cap S_3 = \phi$

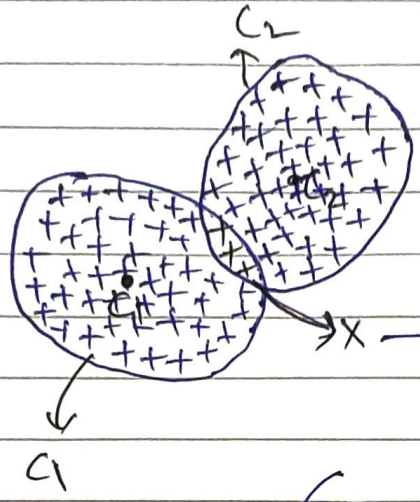
} → Mutually exclusive

Then it is known as Hard Clustering



→ Hard clustering.

Soft clustering :→ (Ex:- GMM)



→ Two clusters are overlapping
→ It has 3 intersection points
→ 3 datapoints are having/ belongs to both clusters.

X ⟶ P(X) → $C_1 \rightarrow 0.6$
P(X) → $C_2 \rightarrow 0.4$

(The probability of the datapoint that belongs to C_1 (centroid 1) (or) and C_2 (centroid 2)

↓

which is having high probability

↓

Then that data point belongs to high probability Centroid.

K-Means

↳ Hyperparameter → K → (Total no. of clusters)

(Note! · In Geo-metric Algorithms

↳ Linear, Logistic, SVM..... → In all these algorithms we are looking at optimization eqn.)

K-means → distance → Eludiean distance (In order to perform for K-means)

→ As of now there is no optimization Equation for K-means

(Actually there iare optimization Eqn for K-means but the issue is it will take very high Computational power for optimization Equation.
↳ As of Now reaseavch is going on but not yet proposed

↓

Because of this reason (no optimization Eqn for K-means) we will use Lloyd's Algorithm.

↓

That's why K-means is also known as Lloyd's Alg

Task!→ Need to form K-clusters Such that intraCluster distance is Low and interCluster Distance is High

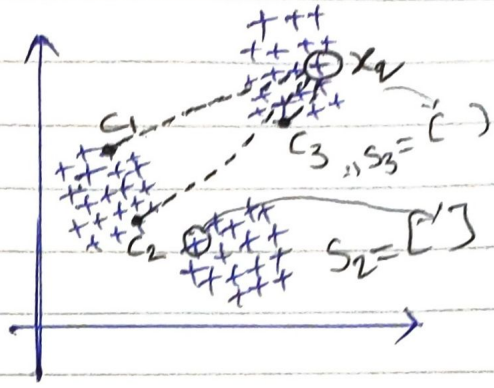
1st step!- Randomly pick "K" data points from the Dataset (D) and Call them as $C_1, C_2, C_3, \ldots\ldots, C_k$ (Centroids)

(This is Initialization step)

→ Intialization Step

C_1, C_2, C_3 → These 3 centroids are randomly choosen.

(This are not the real centroids, So, we need to update the Centroid)



Step:2 :→

For each point x_i in the Dataset → Select the nearest the Centroid C_j

↳ Then add x_i to set S_j.

How to select is calculate distance from x_q to C_1 and x_q to C_2 and x_q to C_3

↳ Get the Centroid which is having minimum distance

↳ In the above example $x_q \in C_3$ (Centroid 3)

↳ So, it belongs to Set 3 (S_3)

→ In this way we will calculate distance from every point in the dataset to centroids and we will store it into the sets (it belongs to which set then

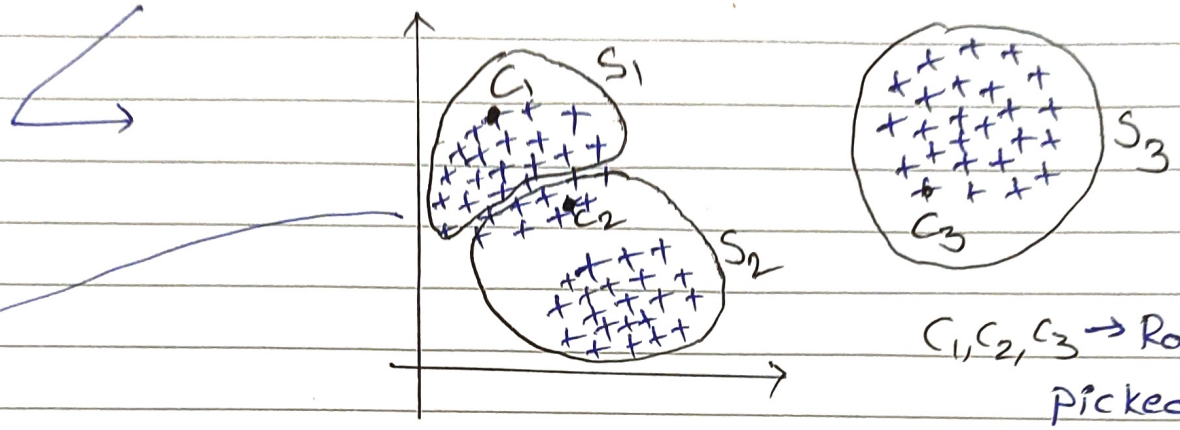
↳ means storing it in respective set

In Simple:→

For each point x_i in the Dataset

↳ select the nearest centroid C_J

↳ Add x_i to set S_J



C_1, C_2, C_3 → Randomly picked Centroids

(This sets are formed according to the distance (nearest Centroid)

(This is called Assignment Step → bcoz assigning each datapoint to corresponding set a/c to distance (nearest Centroid)

Step-3:→

↳ Recalculate C_J (centroid) for each set S_J.

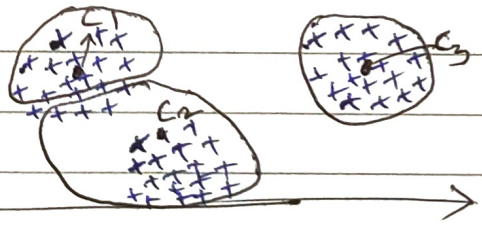
(This step is called Re Computation of Centroid.

$$C_J = \frac{1}{|S_J|} \sum_{S_J \in x_i} x_i$$

→ Sum of datapoints.

↳ no. of data points

Centroids are updated



⇒ Centroids positions are changed.

Step-4 :→

Repeat Step-2 & Step-3 until Convergence.

Centroids won't change.

(i.e sets won't change)

(and this clusters are Mutually exclusive)

$$S_1 \cup S_2 \cup S_3 = 0$$

$$S_1 \cap S_2 = \phi$$

$$S_2 \cap S_3 = \phi$$

$$S_1 \cap S_3 = \phi$$