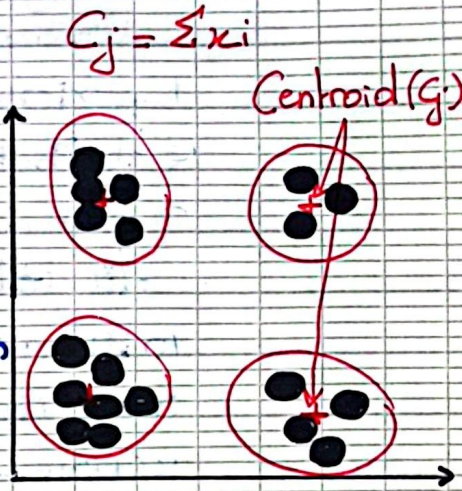
Clustering Methods.

→ Introduction to Clustering

. Groups together data samples that are similar
. A form of unsupervised learning
. A technique of data exploration or a method of looking for patterns or a structure in the data that are of interest.



→ Cluster Definition

. A cluster is represented by a single point known as a centroid (center)
. The centroid is the mean of all data points in a cluster

Major Clustering Approaches

→ Partitioning Approach

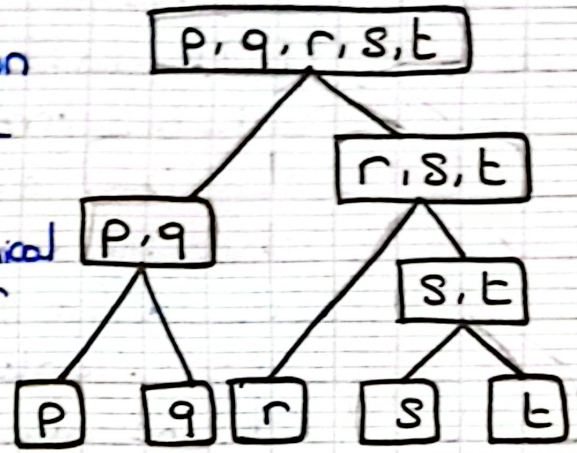
Suppose we have n objects that we want to group between K groups. ($K \leq n$)

↳ Each group contains at least one object
↳ Each object must belong to exactly one group.

→ Hierachical Clustering

. This method creates a hierarchical decomposition of the given set of data object

. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed



→ Model Based Approach

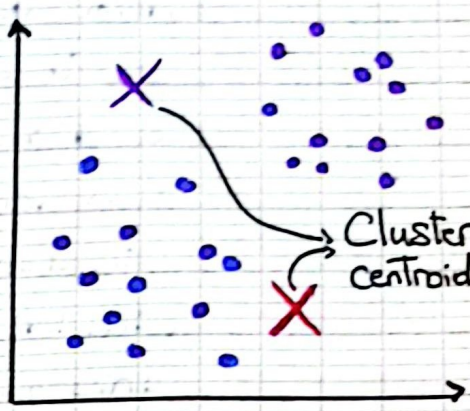
. In this method a model is hypothesized for each cluster to find the best fit of data for a given model.

. This method locates the clusters by clustering the density function and it reflects spatial distribution of the data points

. This method also provides a way to automatically determine the number of clusters based on standard statistic, taking outliers or noise into account.

. It therefore yields a robust clustering method

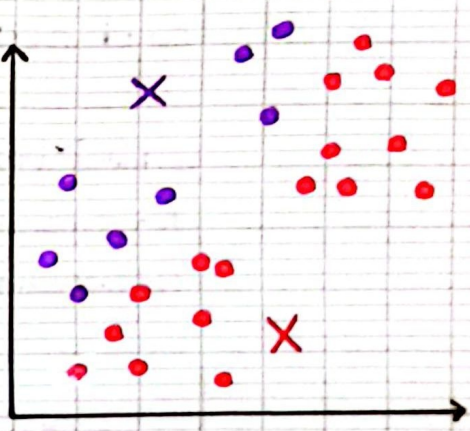
K-means Clustering



- One of the simplest unsupervised learning method to solve known clustering issues.
- It's a partition algorithm
- K-means clustering requires defining:
  - K = number of cluster (manually set)
  - Training set ($x_1, x_2, x_3, \ldots x_m$)

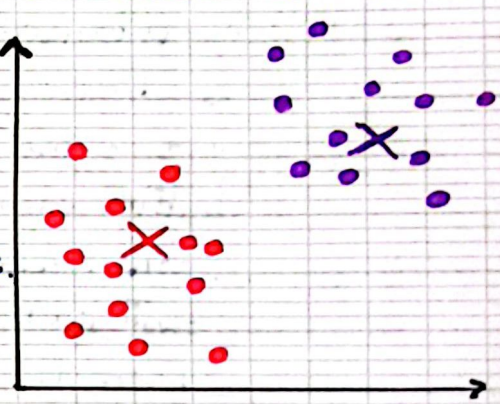
- Let's say we choose K = 2 for the example above
  - we begin by randomly initializing the cluster centroids
  - Then the process of the algorithm begins and executes iteratively until the clusters are learned.

- After random initialization, the algorithm goes trough all the data points, and the cluster centroid initialized in the previous step.
  - to compute the distance between the data points
- Depending upon the minimum distance from the purple cluster centroid or the red cluster centroid, a data point will group itself into that particular group



- We take the previous 2 cluster centroids and iteratively reposition them for optimization
- we take all the red dots, compute their average and move the current cluster centroid to this new location
- We repeat the same process for the purple cluster centroid
- This marks the end of ONE iteration of K-means

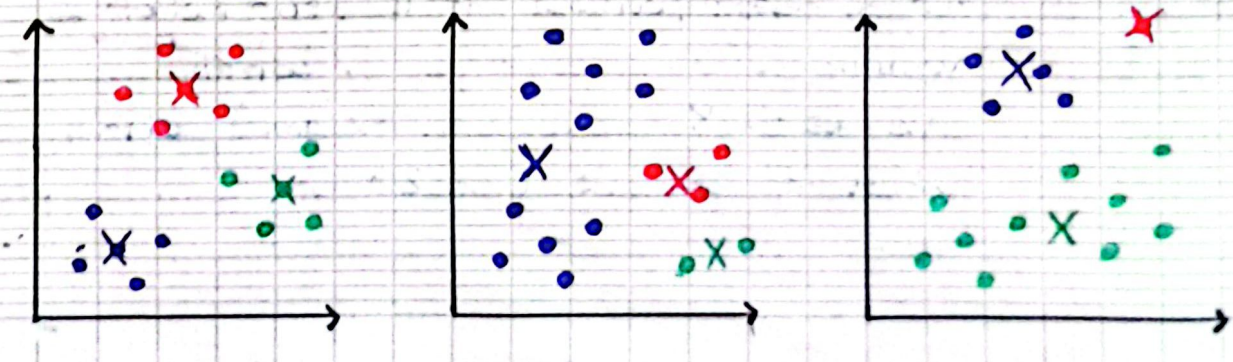
- We iteratively repeat the steps before until the cluster centroids stop changing their positions and become static.
- We consider the algorithm has converged



- At the end K-means clustering algo converges and devides the data points into 2 clusters.

→ K-Means Clustering : Optimization and Convergence

- K-means might end up converging with different solutions according to how clusters were initialized



→ K-Means Clustering : Steps

1. Randomnly initialize the cluster centroid
2. Determine cluster membership for each input
3. Re-compute cluster centroids
4. Repeat from step 2 until stopping criteria is met (i.e. centroids do not change across iterations)

→ Cluster Quality

A. Intra-cluster cohesion (compactness)

Measures how close the data points in a cluster are to the cluster centroid

- The sum of squared error is commonly used measure here, similar to the mean squared error but with the sum versus average $SSE = \sum (y_i - \hat{y}_i)^2$

B. Inter-cluster separation (isolation)

Measures how much clusters are far from each other. The more compact and isolated the better.

→ Determining the best K number of clusters

Elbow Method:

1. Run clustering algorithm (e.g. K-means clusters) for different values of K. For instance, by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WSS)

$$WSS = \sum_{i=1}^{\text{nb of clusters}} \sum_{x \in \text{Cluster}_i} \text{distance}(x, \text{Centroid}_i)$$