Clustering:

1) Clustering is an unsupervised technique

2) The aim of clustering is to group similar objects together and dissimilar objects in different groups

K-Means Clustering :-

- One of the most popular clustering algorithms
- The aim of K-Means clustering is that given N data points, say $x_1, x_2, \ldots, x_N \in \mathbb{R}^d$ it tries to divide the data into k (specified) disjoint groups say $\{C_1, C_2, \ldots, C_k\}$ s.t. the following cost is minimized

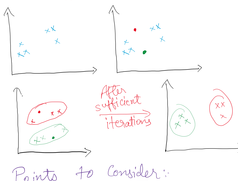
$$\text{cost}(k) = \sum_{i=1}^{N} d(x_i, C)^2$$

Each Cluster $C_1, C_2, \ldots, C_k$ can be represented by its center or mean say $c_1, c_2, \ldots, c_k$

So $d(x_i, C)$ is the distance of the ith point to its nearest cluster center.

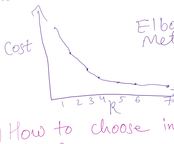
Lloyd's Algorithm for K-Means

1. Assign Initial Cluster Centers
2. Repeat till convergence
   - Assign each point to its nearest cluster center
   - Recalculate means of the new clusters as new cluster centers



Points to Consider:

i) How to choose k?



ii) How to choose initial centers?

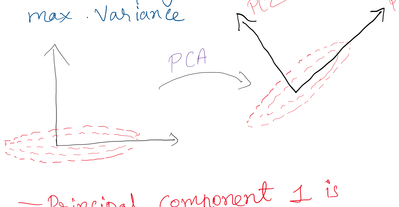
- Randomly
  or
- k-Means ++
  - Choose 1st center randomly
- Next center is with high probability farthest from 1st and so on for (k-1) iterations

iii) K-Means doesnot work when

PCA (Principal Component Analysis)

→ Unsupervised technique for obtaining variance information from high-dimensional data

→ Orthogonal Projection of data into possibly lower dimensions so that projected data has max. Variance



- Principal component 1 is the direction in which data has maximum variance

→ PC2 is the direction of 2nd highest variance and orthogonal to PC1 and so on...

→ PCA is used for dimensionality reduction.

→ Issue with PCA :-
The newly obtained principal components do not have explainability

→ How to obtain PCA

Step 1: Standardize your data (Z-score normalization)

Step 2: Find the covariance matrix of the data $\frac{1}{n-1}(A^T A)$

Step 3 :- Find the eigen values & eigen vectors of the data

The eigen vector corresponding to highest eigen value is 1st P.C. to 2nd highest eigen value is 2nd P.C. and so on....

Step 4: Project your original data on eigenvectors/PCs to get data with reduced dimension

The steps 1, 2, 3 can be done without EVD by actually taking SVD of the standardized data matrix too.

In Python direct function to obtain PCA.