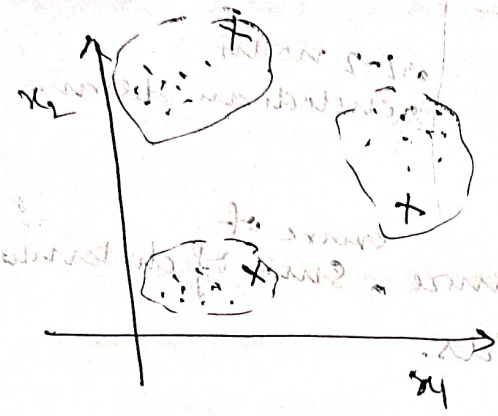
K-means Clustering

→ Unsupervised. Clustering.

Process



1. Choose K points randomly as centroids.
2. Assign class wrt nearest centroids.
3. Find new centroids according to cluster.

$$x_{1cg} = \frac{\sum m_i x_{1i}}{\sum m_i}$$ Similarly $$x_{2cg} = \frac{\sum m_i x_{2i}}{\sum m_i}$$

4. Replace old centroids with new ones.
5. Iterate these (2-4) until stopping criteria is reached.

EX- Stopping Criteria - Change in centroids < Certain threshold.

NOTE - CURSE OF DIMENSIONALITY

→ In increased dimensional spaces, distance b/w points change rapidly on changing same amount. Thus quality of prediction decreases.

→ Thus similar to KNN;

↑ dimension ↑ performance ↓

Objective / cost-function

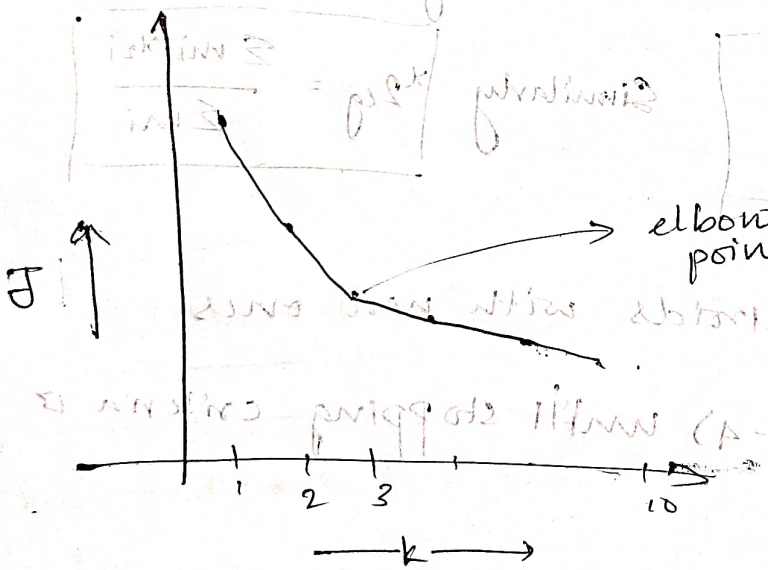
$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \| x_i^j - c_j \|^2$$

- L-2 norm
- euclidian distance
- square of

→ This says that we want to minimize sum of distances of data points from cluster centroids.

FINDING - K

↳ Elbow Method:



→ slope before elbow > slope after elbow

→ Decrease in cost function is highest at 3. Thus K=3 is the optimal value.

→ PROBLEM

→ Random Initialization

↳ Initial points are chosen randomly.

↳ very close or very far.

↳ This may lead to wrong results.

→ Solution

(i) Run K-Means multiple times & choose the one with lowest J value and has occured most no of times.

(ii) K-Means ++

Intuition: After choosing first point, some kind of probability distribution is used to make a selection of a point far from original point. This is because, far away points create better clusters.