Unsupervised Machine Learning

Clustering Algorithms

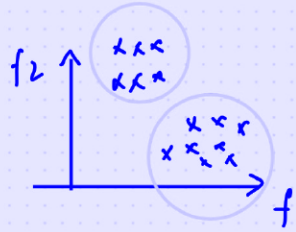
1. K Means Clustering
2. Hicranichal Clustering.
3. DBScan Clustering.

I/p — Dependent

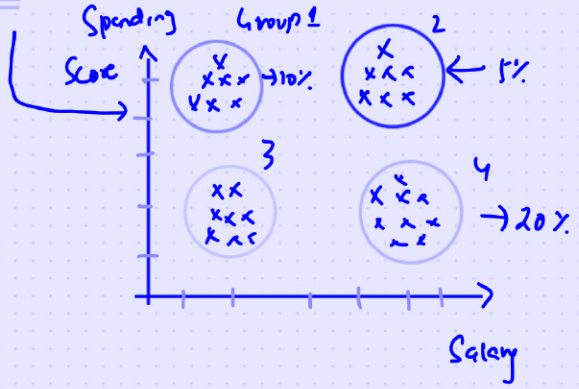
f_1 f_2 - - f_n o/p

Supervised ML

f_1	f_2	f_3	- -	f_n
-	-	-		-
-	-	-		-



Ecommerce → Iphone 20 → 1.5 lakhs

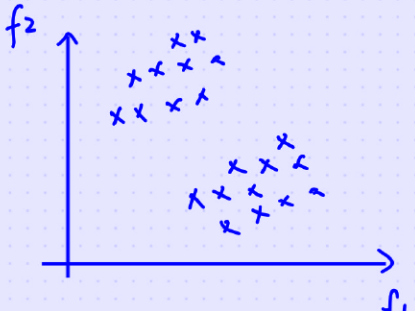


Dataset

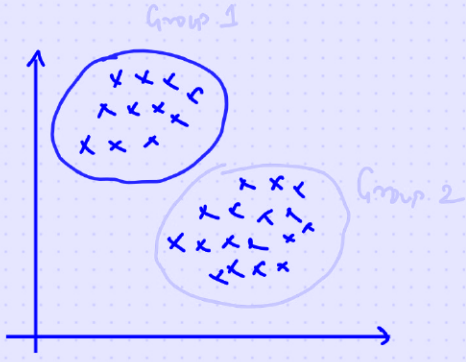
Salary	Spending Score (1-10)
-	-
-	-
-	-

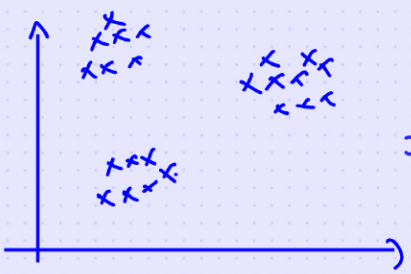
1. K Means Clustering

Geometric Intuition

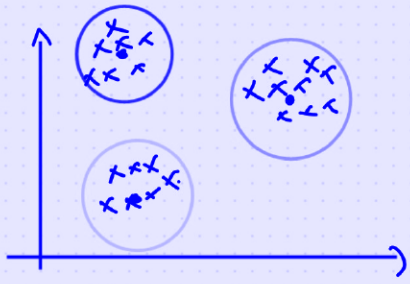


⇒ K Means

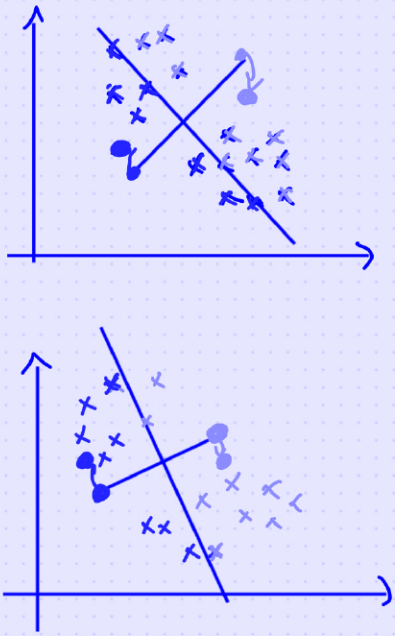




⇒ K Means



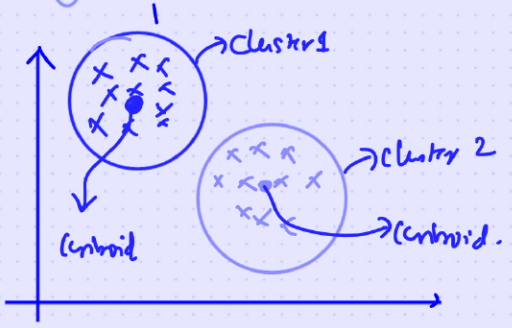
K Means Mathematical Intuition



Steps:

k = 2

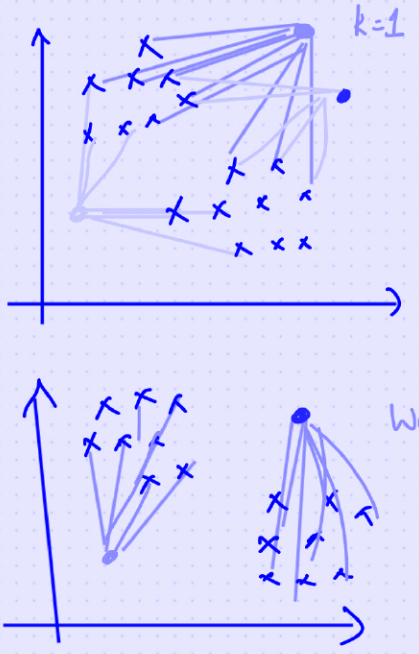
1. Initialize some K → centroids
2. Points that are nearest to the centroid → Group
3. Move the centroids → Mean



How do we select the K Value ?

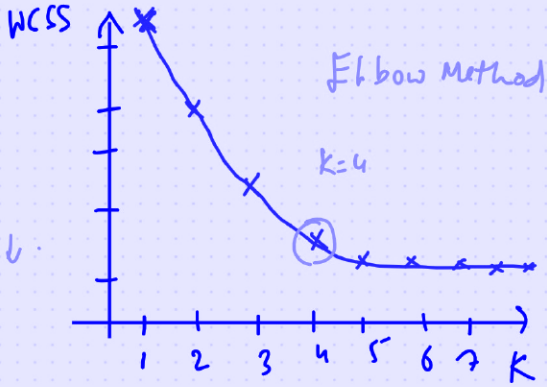
WCSS = Within Cluster Sum of Squares

Initialize K = 1 to 20

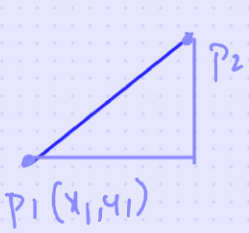


$$WCSS = \sum_{i=1}^{K} (\text{Distance between point to the nearest centroid})^2$$

WCSS ↓↓.



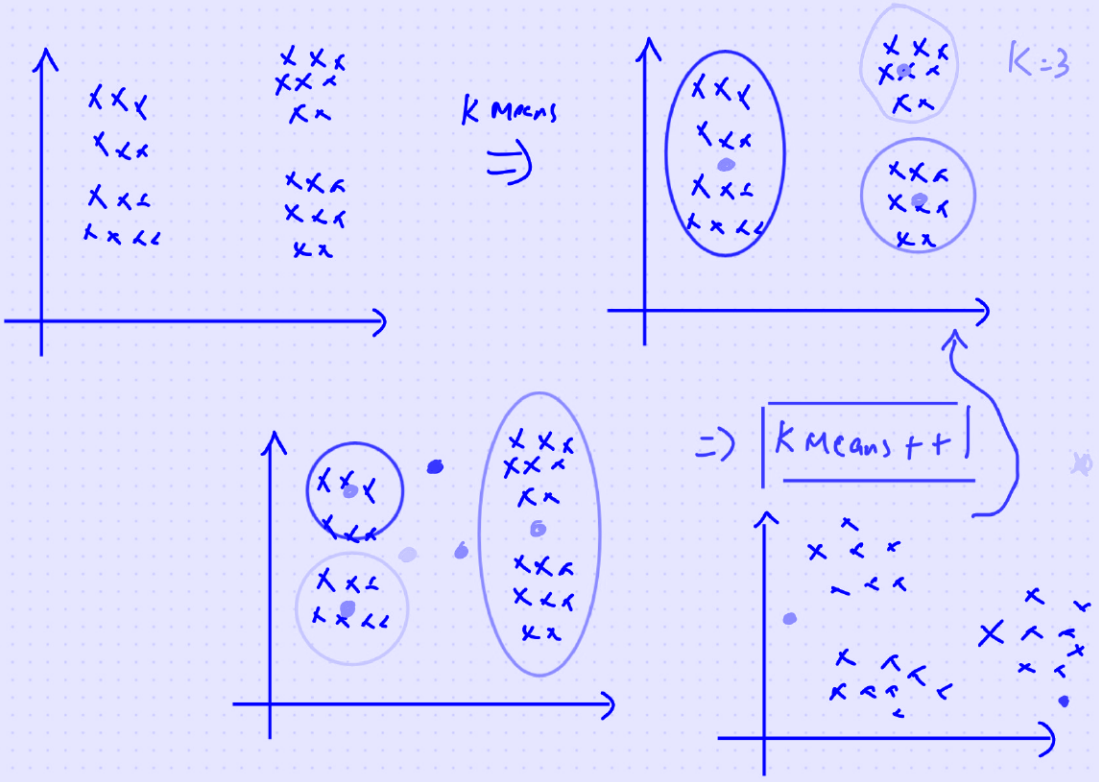
Euclidian Distance or Manhattan Distance



$$ED = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

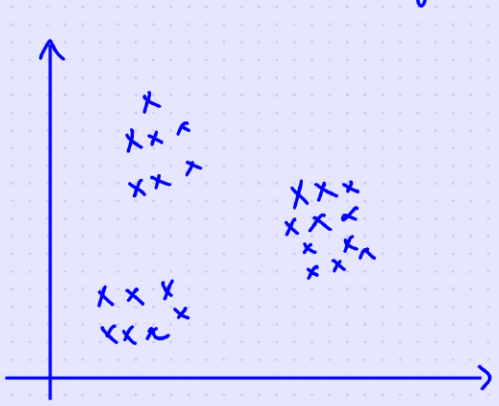
$$MD = |x_2 - x_1| + |y_2 - y_1|$$

Random Initialization Trap (K Means ++)

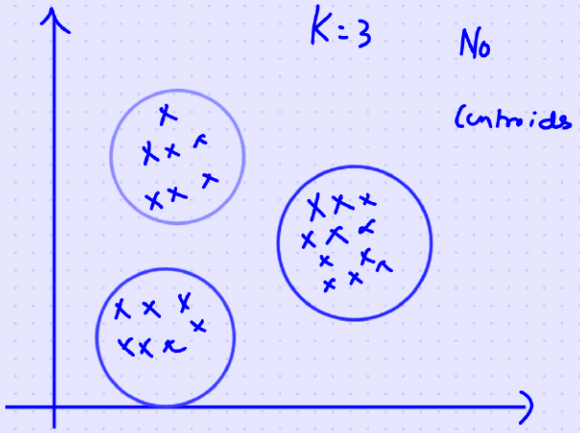


K Means ++ Initialization Technique

Hierarichal Clustering



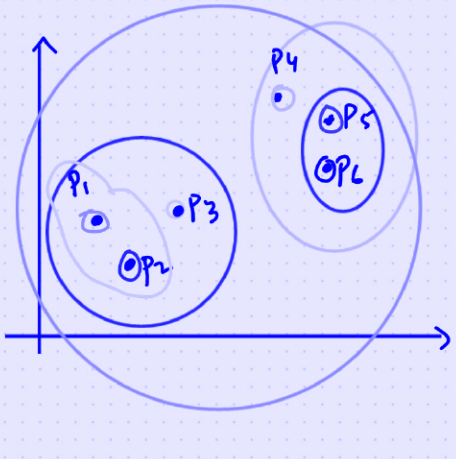
Heirarichal Clustering



HC

① Agglomerative
② Divisive
} => Geometric Intuition

Steps

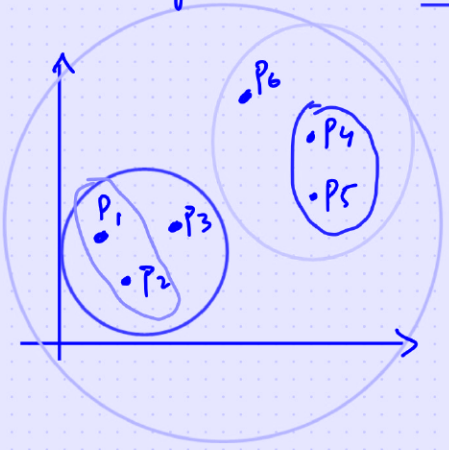


① For each point initially we will consider it as a seperate cluster

② Find the nearest point and creat a new cluster

③ Keep on doing the same process [step 2] until we get a single cluster.

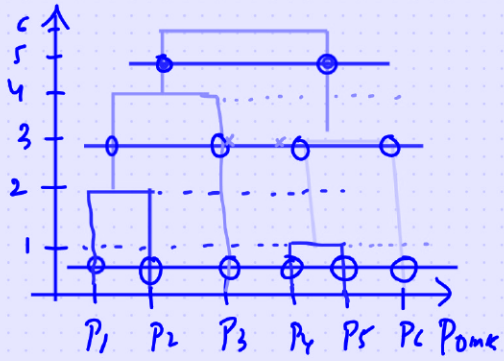
How many clusters? K=2



Threshold

Euclidian Dist

Dendogram



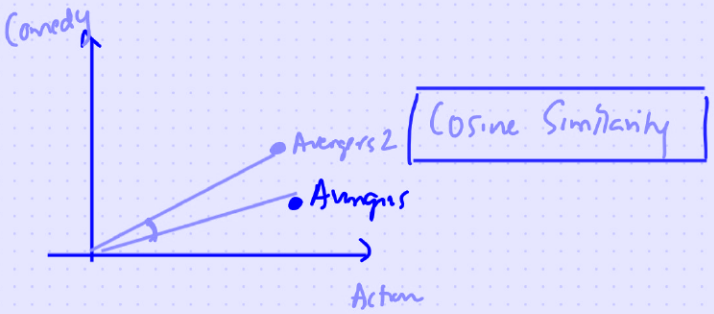
④ Select the longest vertical line such that no horizontal line passed through it.

K Means Vs Hierarchal Clustering

Scalability And Flexibility

① Dataset size
- Huge => K Means
- Small => Hierarchal Clustering

② Types of Data
- Numerical data → K Means or Hierarchal
- Variety of data → Hierarchal



Cosine Similarity

Silhouette Clustering

① a(i)

For data point $i \in C_I$ (data point i in the cluster C_I), let

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j)$$

be the mean distance between i and all other data points in the same cluster, where $|C_I|$ is the number of points belonging to cluster i, and d(i,j) is the distance between data points i and j in the cluster C_I (we divide by $|C_I| - 1$ because we do not include the distance d(i,i) in the sum). We can interpret a(i) as a measure of how well i is assigned to its cluster (the smaller the value, the better the assignment).

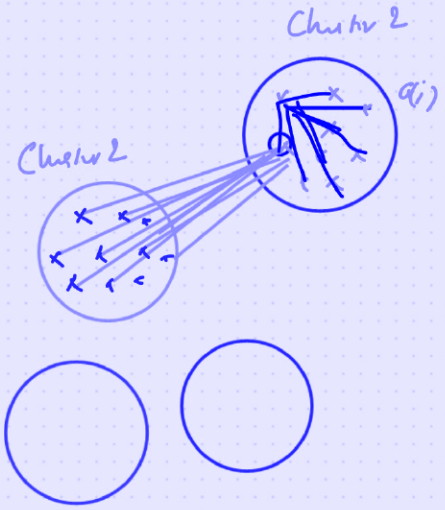
② b(i)

We then define the mean dissimilarity of point i to some cluster C_J as the mean of the distance from i to all points in C_J (where $C_J \neq C_I$).

For each data point $i \in C_I$, we now define

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$$

to be the *smallest* (hence the min operator in the formula) mean distance of i to all points in any other cluster (i.e., in any cluster of which i is not a member). The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.



③ Silhouette Score

small ← a_i/b_i

$a_i > b_i$ ⇒ Good cluster

$a_i < b_i$ ⇒ bad cluster

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$ ⇒ -1 to 1

and

$$s(i) = 0, \text{ if } |C_I| = 1$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

From the above definition it is clear that

$$-1 \leq s(i) \leq 1$$