Unsupervised Machine Learning

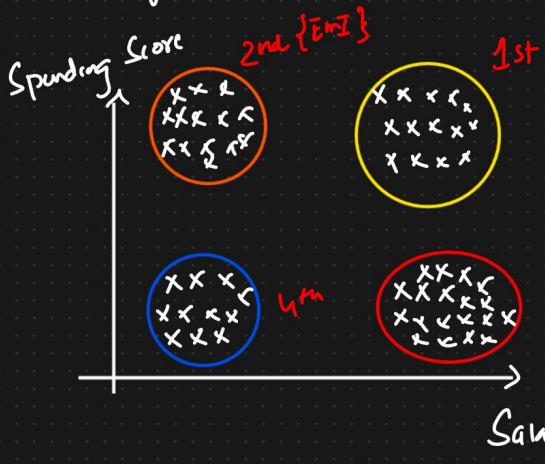
Clustering [grouping]

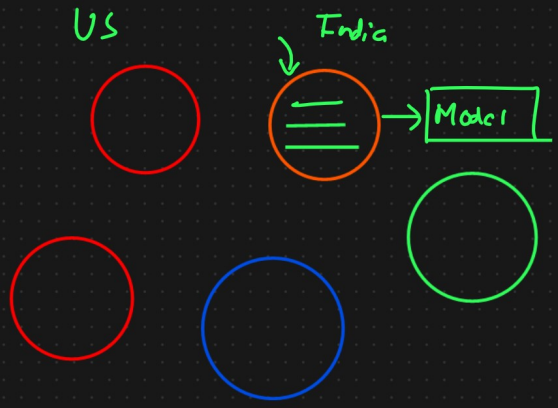
① K Mean Clustering
② Hierarchal clustering
③ DBScan clustering

⇒ Performance Metrics Validate clustering algorithm.

⇓

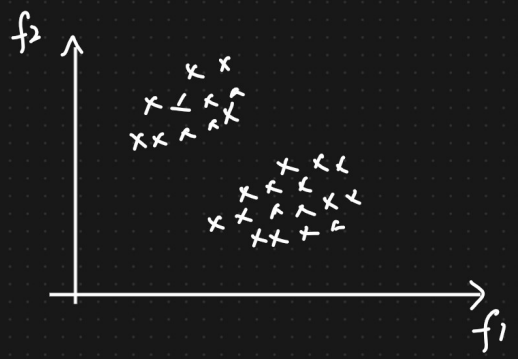
Silhouette Score



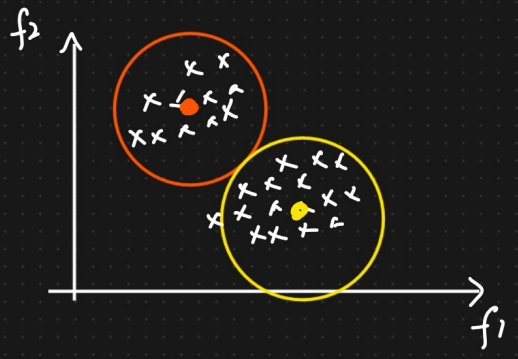


① K Means Clustering

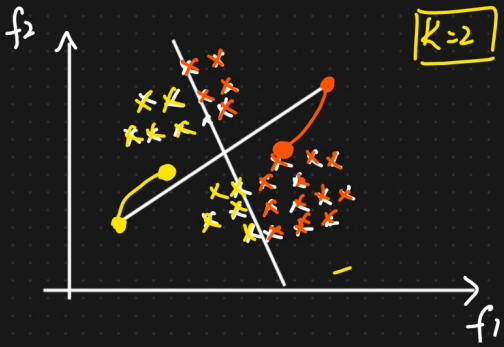
Geometric Intuition



KMeans ⇒

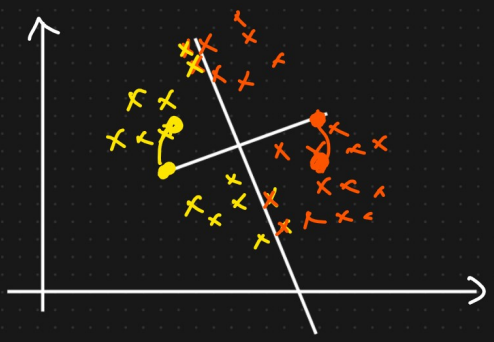


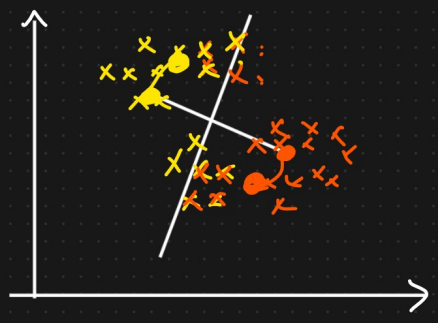
K Means Mathematical Intuition

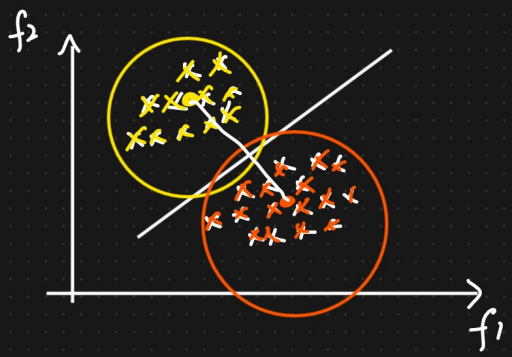


Steps. K centroids

① Initialize some K value
② Label all the points based on distance to the Nearest Centroid
③ Move the centroid







How do we select the K value [Elbow method] ⇒ Knee locator

WCSS = Within Cluster Sum of Square.

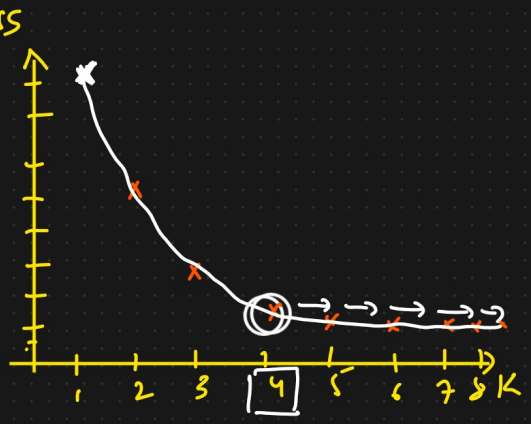
Initialize K: 1 to 10

K=1

K=2

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

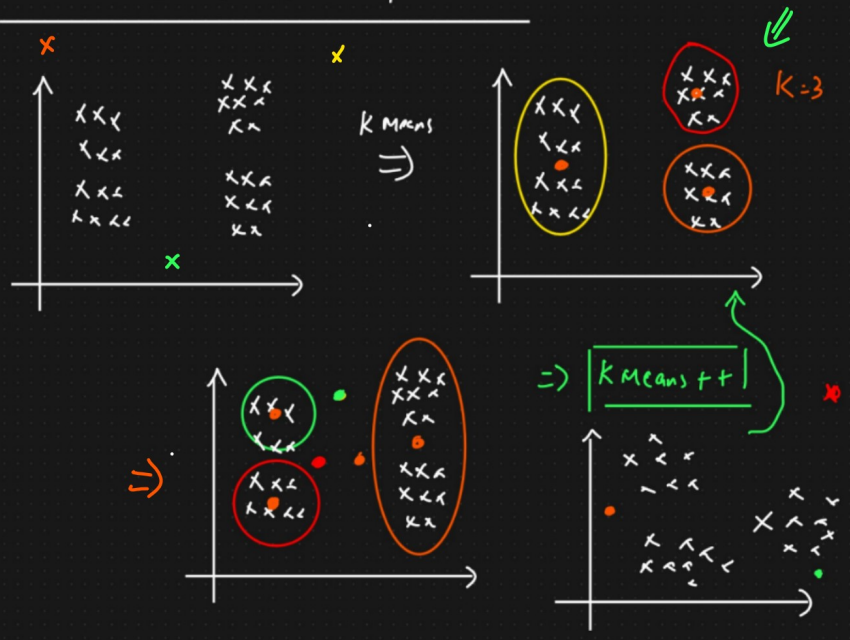
WCSS



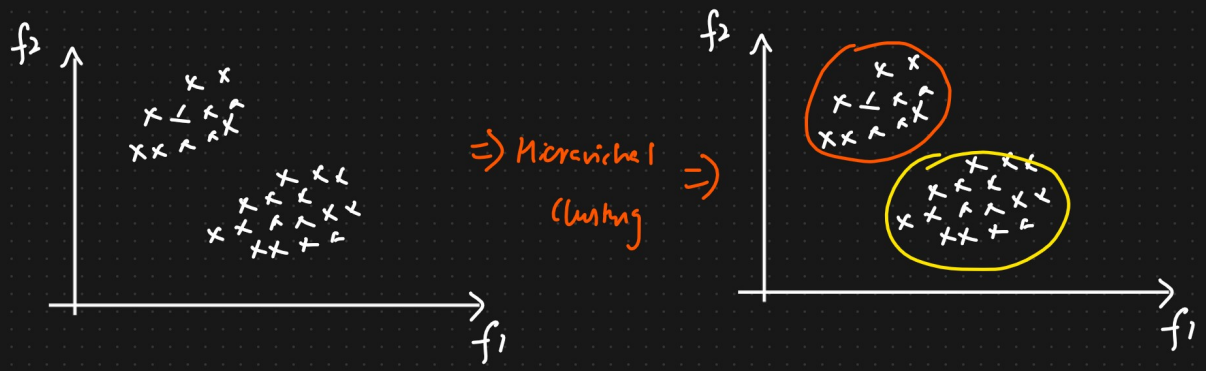
Euclidian Distance

K ⇒ Number of centroids.

Random Initialization Trap (K Means ++) ⇒

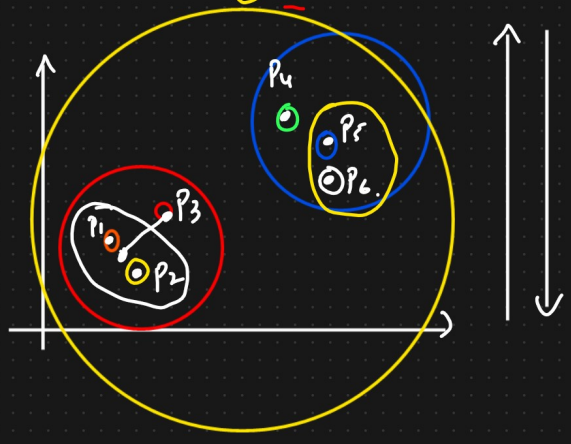


② Hierarichal Clustering [Agglomerative Clustering]



HC

① Agglomerative
② Divisive
⇒ Geometric Intuition

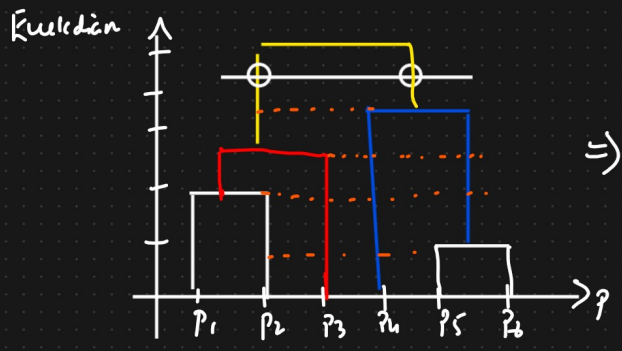


Steps

① For each point we will consider it as a separate cluster

② Find the nearest point and creak a new cluster

Euclidian Distance



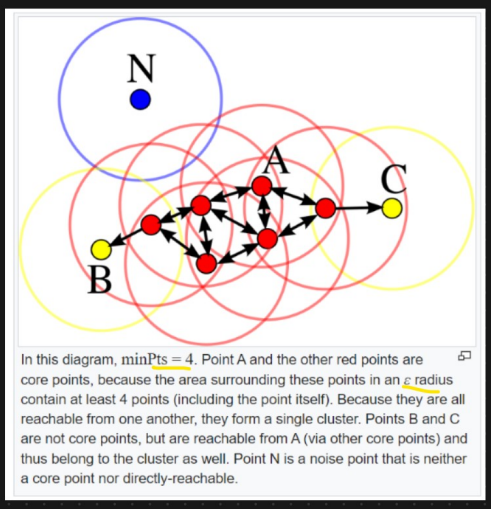
⇒ Dendogram.

k Means Vs Hierarchical Clustering

Scalability and Flexibility

① Dataset size → Huge ⇒ k Means
→ Small ⇒ Hierarchical Clustering

③ DBScan Clustering



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

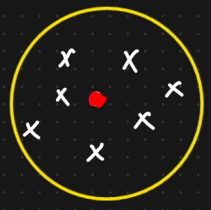
● → Core points
● → Border points
● → Noise/outliers
} Non linear Clustering.

Hyperparameter

① minpts = 4 ② ∈ = radius

Core point ←

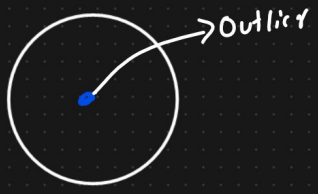
① No. of points within the ∈ ≥ minpts



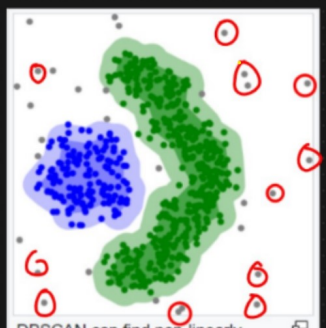
② Border points

No. of data points within the radius is less than minpts = 4

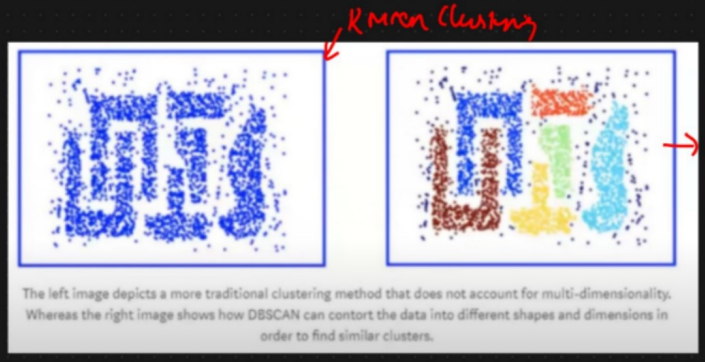
③ Outliers [DBScan is robust to outliers].



Some Examples after we apply DBScan Clustering



DBSCAN can find non-linearly separable clusters. This dataset cannot be adequately clustered with k-means or Gaussian Mixture EM clustering.



The left image depicts a more traditional clustering method that does not account for multi-dimensionality. Whereas the right image shows how DBSCAN can contort the data into different shapes and dimensions in order to find similar clusters.

Silhouette Score [Validate the clustering model].

① First step

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 C_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).

a(i)

②

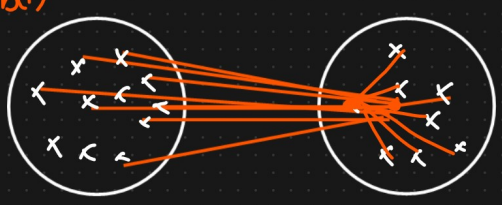
We then define the mean dissimilarity of point i to some cluster C_J as the mean of the distance from i to all points in C_J (where $C_J \neq C_I$).

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

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

b(i) > a(i) ⇒ Clustering is good.

b(i)



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

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

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

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