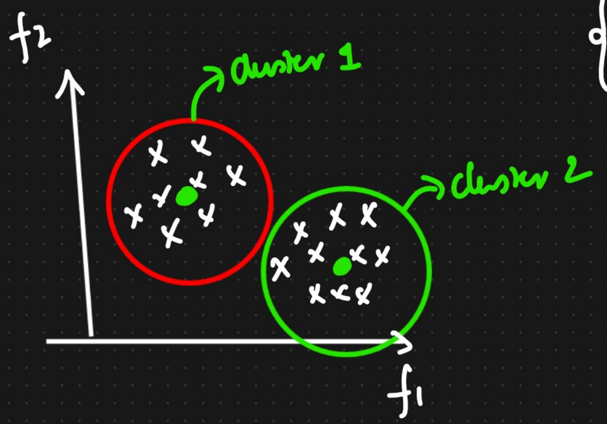
Day 6 — Machine Learning Algorithms

Unsupervised ML

① K Means Clustering
② Hicrarichal clustering
③ Silhoutte Score
④ DBScan Clustering

Agenda { ① Svm & SvR
② Xgboost
③ PCA }

K Means Clustering

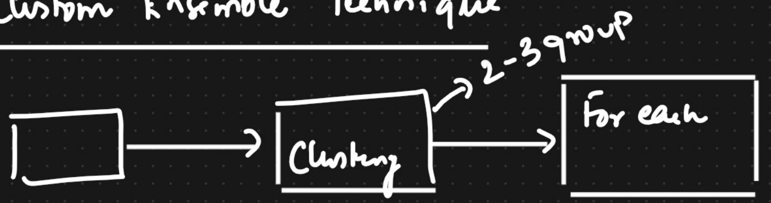


Unsupervised ML

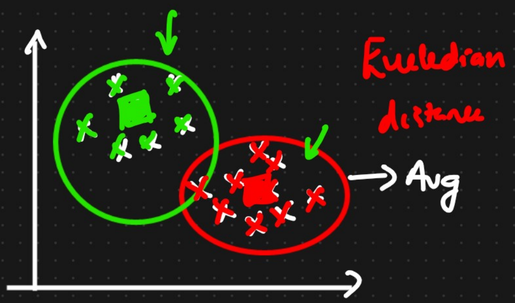
O/P	f1	f2
	—	—
	—	—
	—	—
	—	—

Clusters
⇓
{ Similar Kind of data }

Custom Ensemble Technique



K Means → K = centroids



K=2

High dimension

$\boxed{K=2}$

centroids

① We try K values ⇒ suitable K=2

② Initialize K number of centroids ✓

③ Compute the avg to update centroids ✓

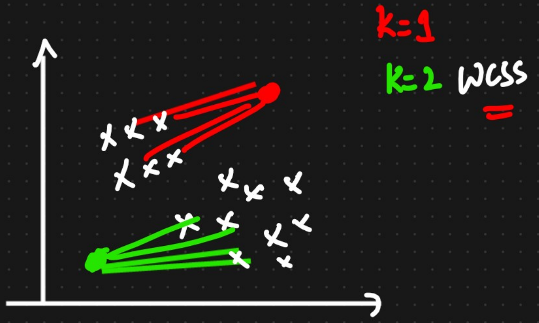
Validating

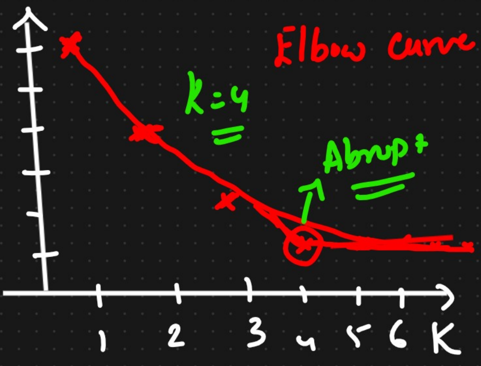
within cluster sum of square

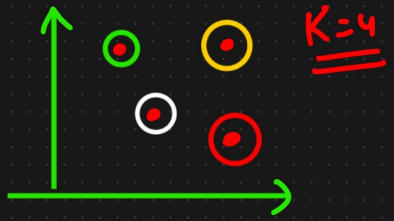
K=

Elbow method (K value)

for i=1,10



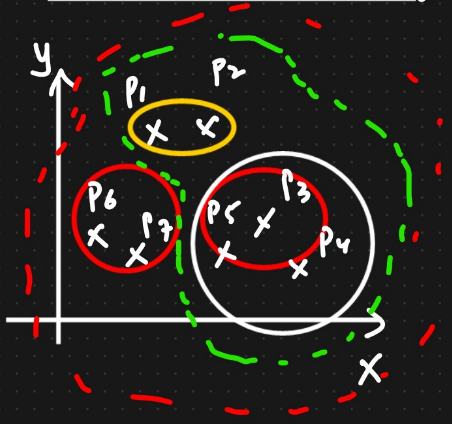


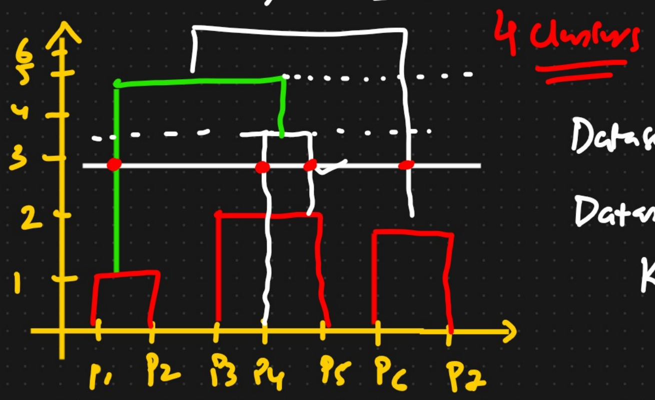


② Hierarichal Clustering

You need find the longest vertical line that has no horizontal line passed through it.

Dendogram

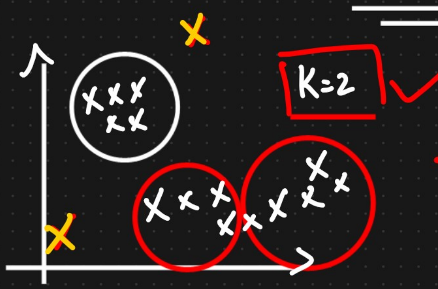




Dataset is small
Dataset is large } Kmeans

Max Time is taken by KMeans or

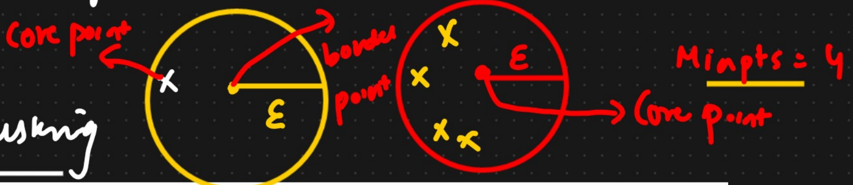
Hierarichal clustering ?? ✓
↳ Max Time

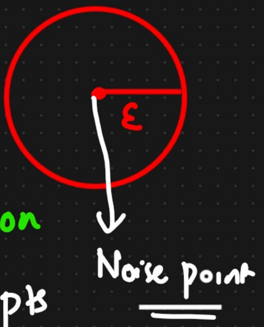


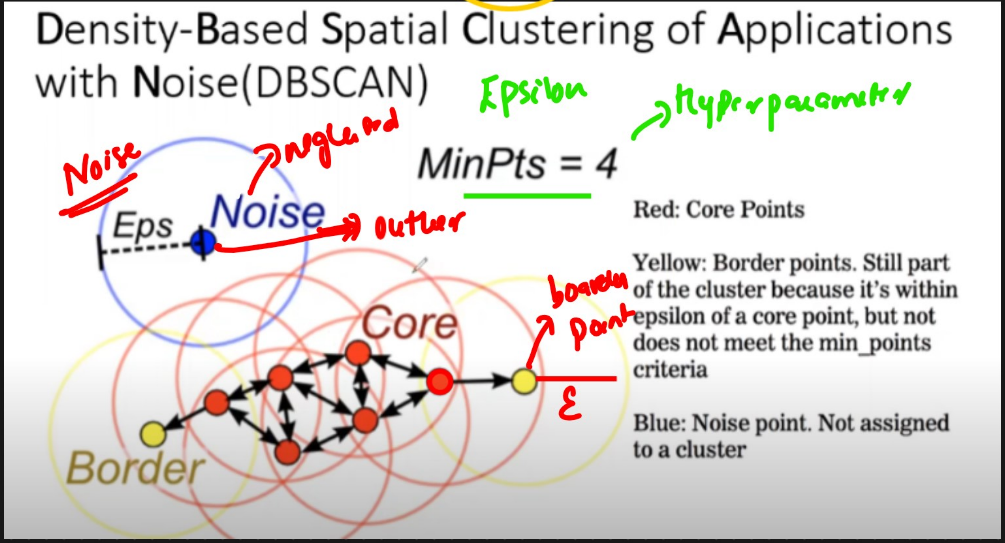
⇒ { KMeans ++ }

Validate Clustering Models

DBScan Clustering







Density-Based Spatial Clustering of Applications with Noise(DBSCAN)

MinPts = 4

Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

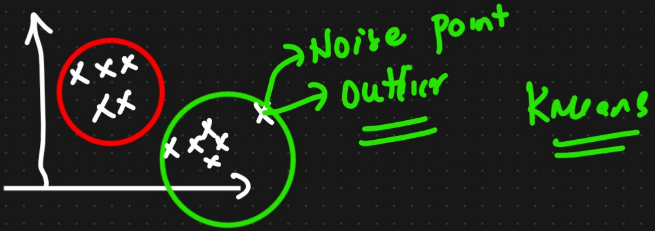
Ⓐ Epsilon

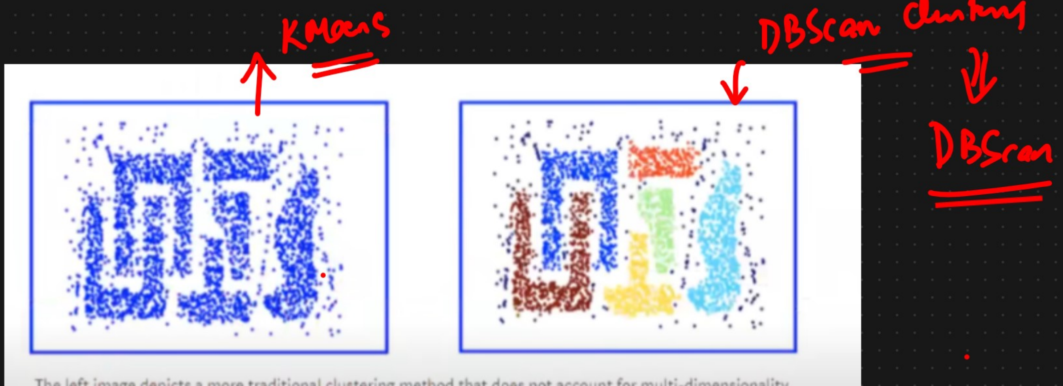
① Min pts

② Core points

③ Border points

④ Noise point



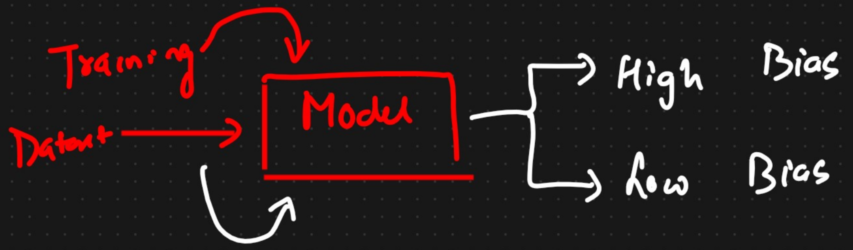


The left image depicts a more traditional clustering method that does not account for multi-dimensionality. Whereas the right image shows how DBSCAN can contort the data into different shapes and dimensions in order to find similar clusters.

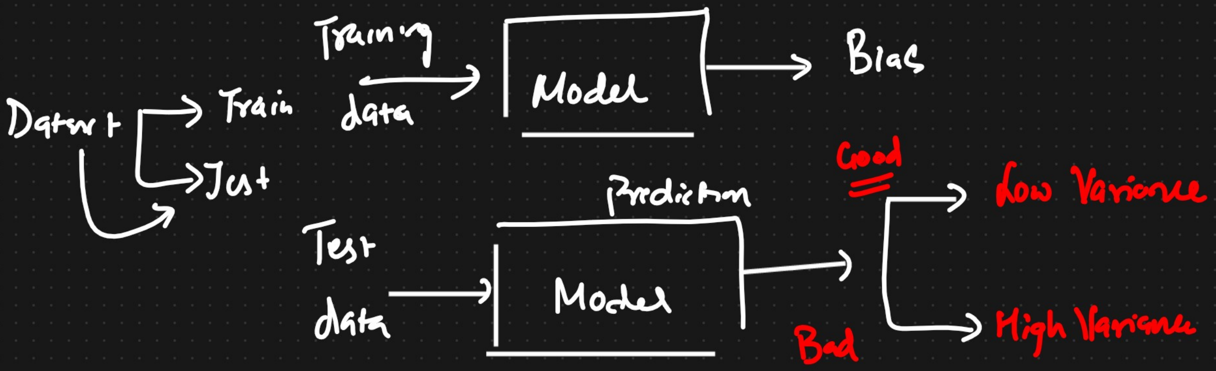
Defn of Bias And Variance

Training Dataset = 90%
Test Dataset = 70%
} ⇒ Overfitting
⇓
{Low Bias ✓}
{High Variance ✓}

Bias : It is a phenomenon that skews the result of an algorithm in favor or against an idea.
↳ Training dataset



Variance : Variance refers to the changes in the model when using different portions of the training or test data



Model 1	Model 2	Model 3
Train Acc = 90%	Train Acc = 60%	Train Acc = 90%
Test Acc = 75%	Test Acc = 55%	Test = 92%

Low Bias
High Variance

High Bias
High Variane

Low Bias
Low Variane

<u>Generalized Model</u>