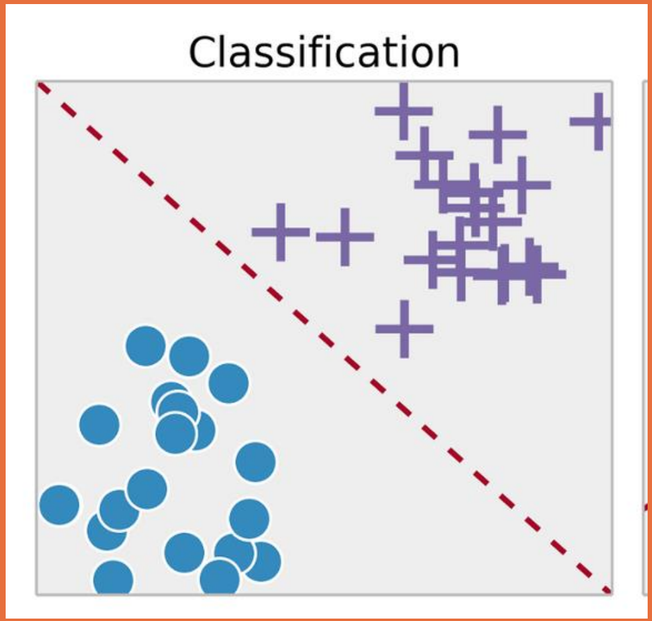
Machine Learning Fundamental Part 2

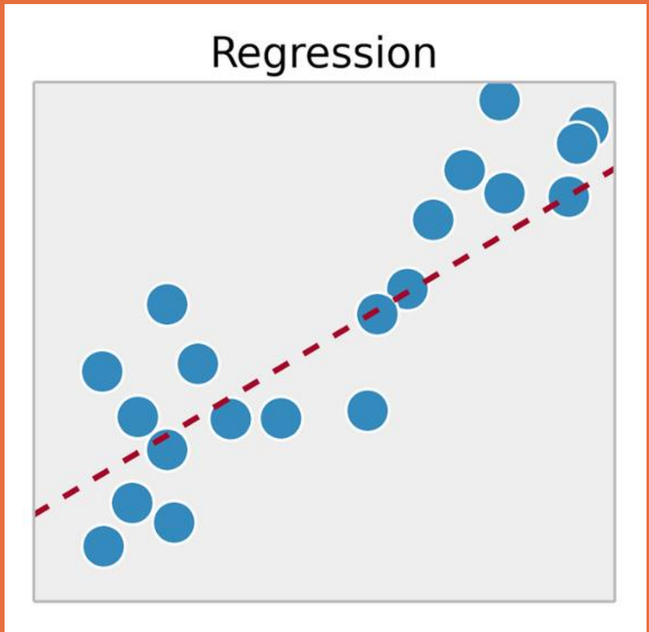
Prepared by Muhamad Irfan Fadhullahl

Today's Discussion

Outline of Topics

Supervised learning
Support Vector Machine
K Nearest Neighbour
Random Forest
Unsupervised Learning
K-Means Clustering
Hierarchial Clustering
Cross Validation
Model Performance and Selection





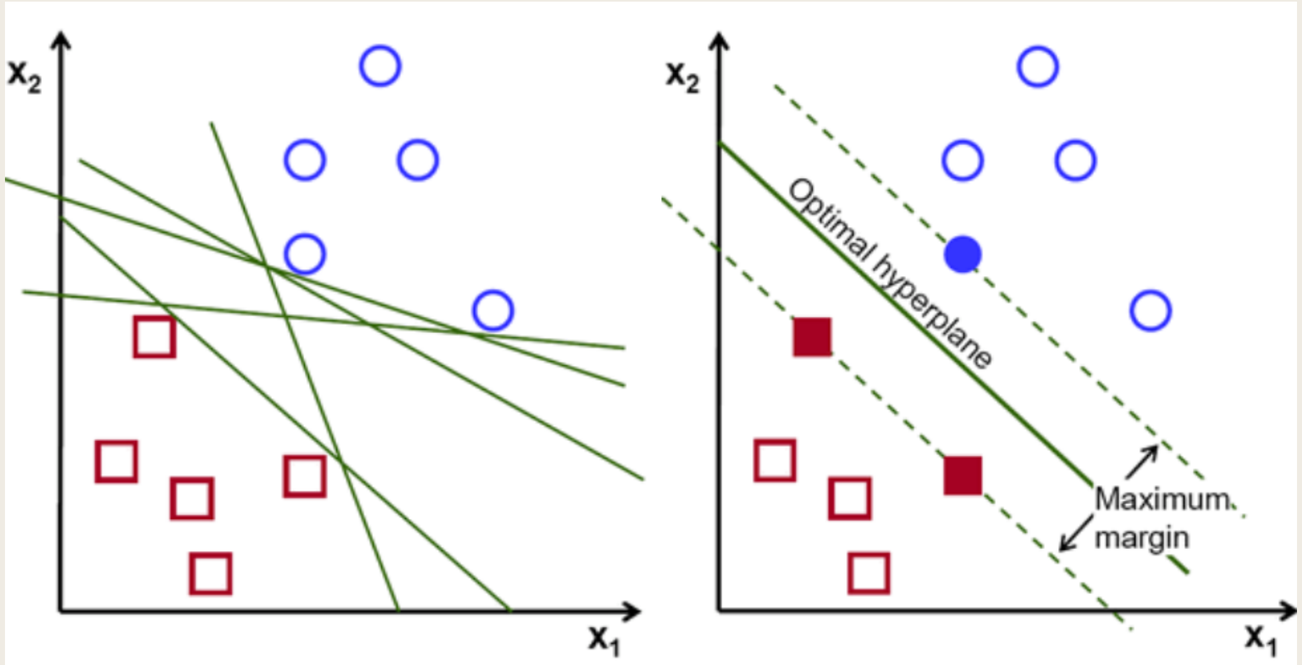
- Random Forest
- Support Vector Machine
- K Nearest Neighbours

Supervised Learning

04 Support Vector Machine

Explanation

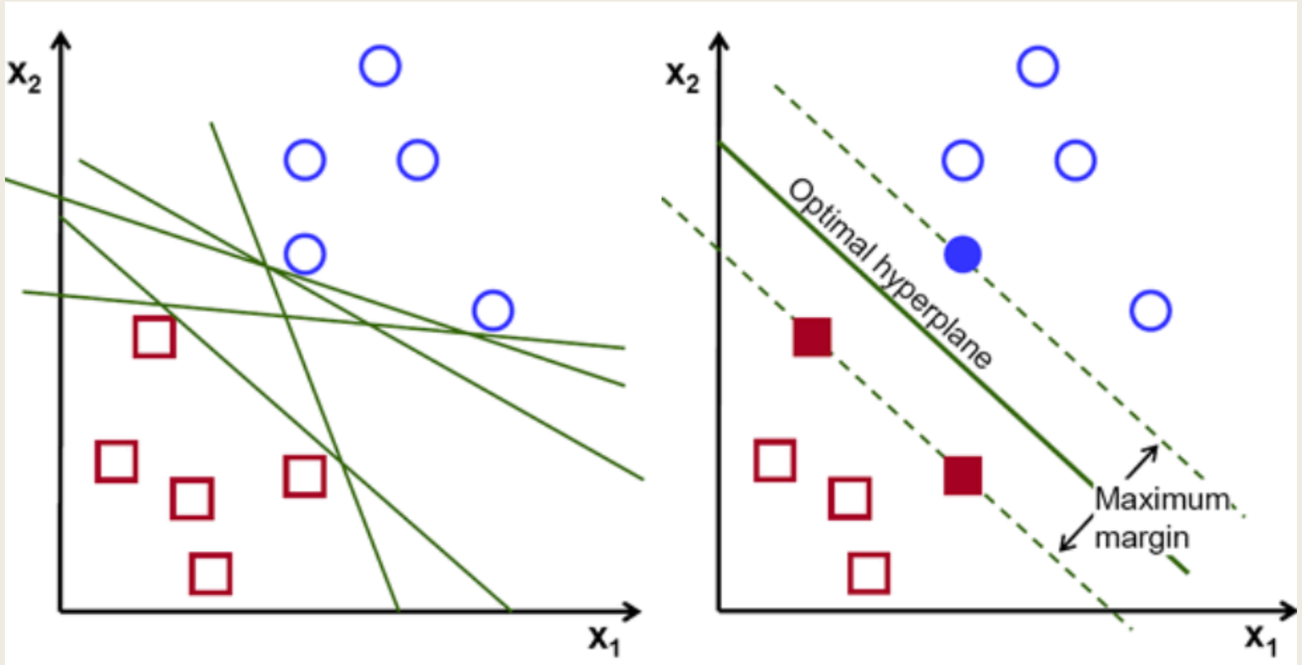
Support Vector Machines (SVMs) are a set of supervised learning methods which learn from the dataset and can be used for both regression and classification.



05 Support Vector Machine

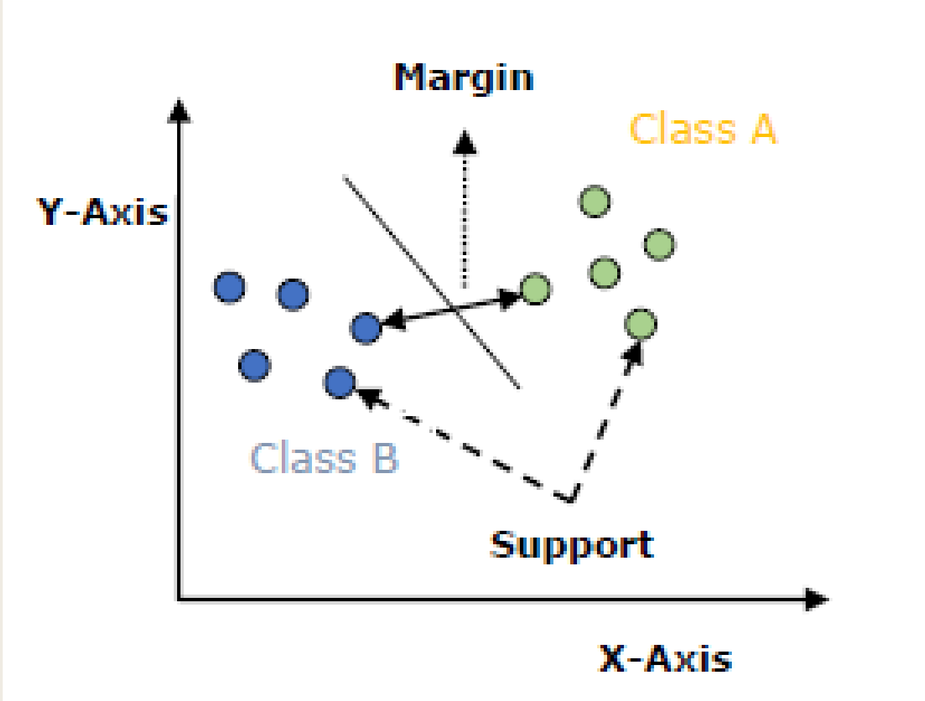
Explanation

Support Vector Machines (SVMs) are a set of supervised learning methods which learn from the dataset and can be used for both regression and classification.



06 SVM Concept

- Support Vectors – Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.
- Hyperplane – As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
- Margin – It may be defined as the gap between two lines on the closet data points of different classes.



07

SVM Kernels

Linear Kernel

It can be used as a dot product between any two observations.

$$K(x, x_i) = sum(x * x_i)$$

Polinomial Kernel

It is more generalized form of linear kernel and distinguish curved or nonlinear input space.

$$k(X, Xi) = 1 + sum(X * X_i)\text{^}d$$

Radio Basis Function Kernel

RBF kernel, mostly used in SVM classification, maps input space in indefinite dimensional space.

$$K(x, xi) = exp(-gamma * sum(x - xi\text{^}2))$$

Advantages

SVM classifiers offers great accuracy and work well with high dimensional space. SVM classifiers basically use a subset of training points hence in result uses very less memory.

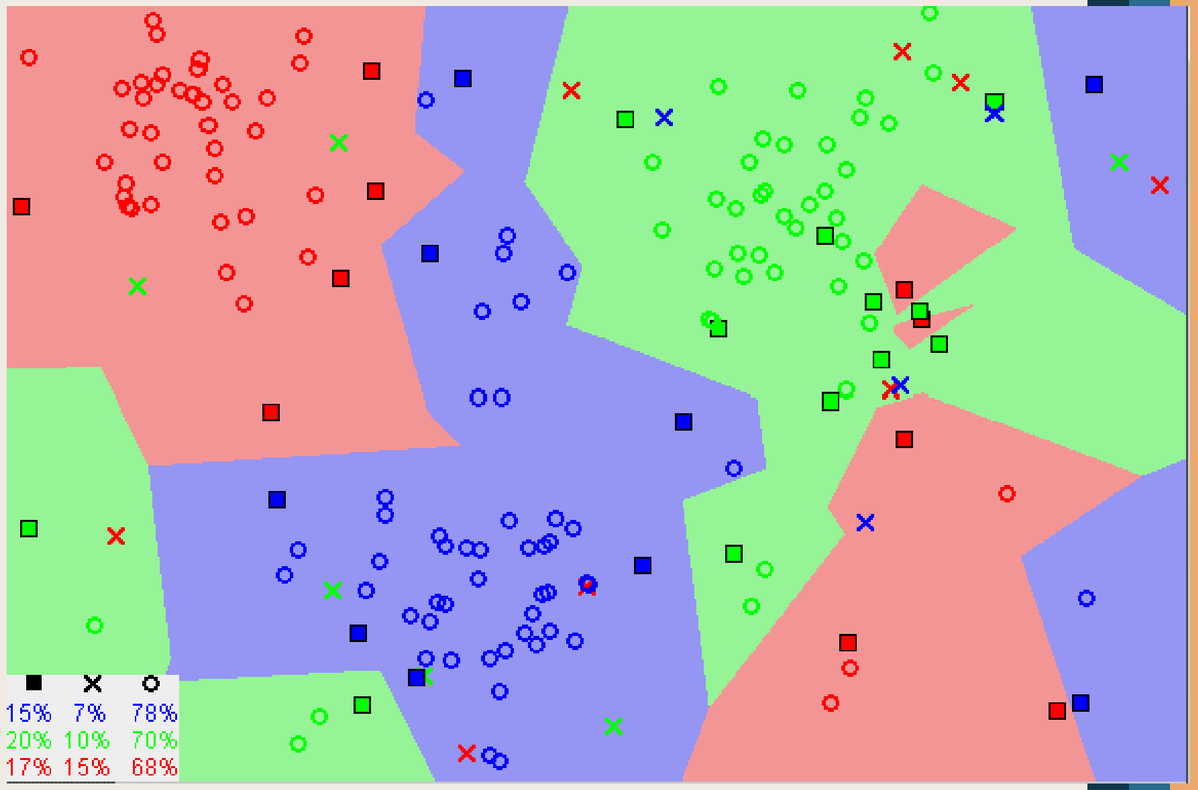
Disadvantages

They have high training time hence in practice not suitable for large datasets. Another disadvantage is that SVM classifiers do not work well with overlapping classes.

K- Nearest Neighbours

Explanation

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.



- Lazy learning algorithm − It does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − It doesn’t assume anything about the underlying data.

Advantages

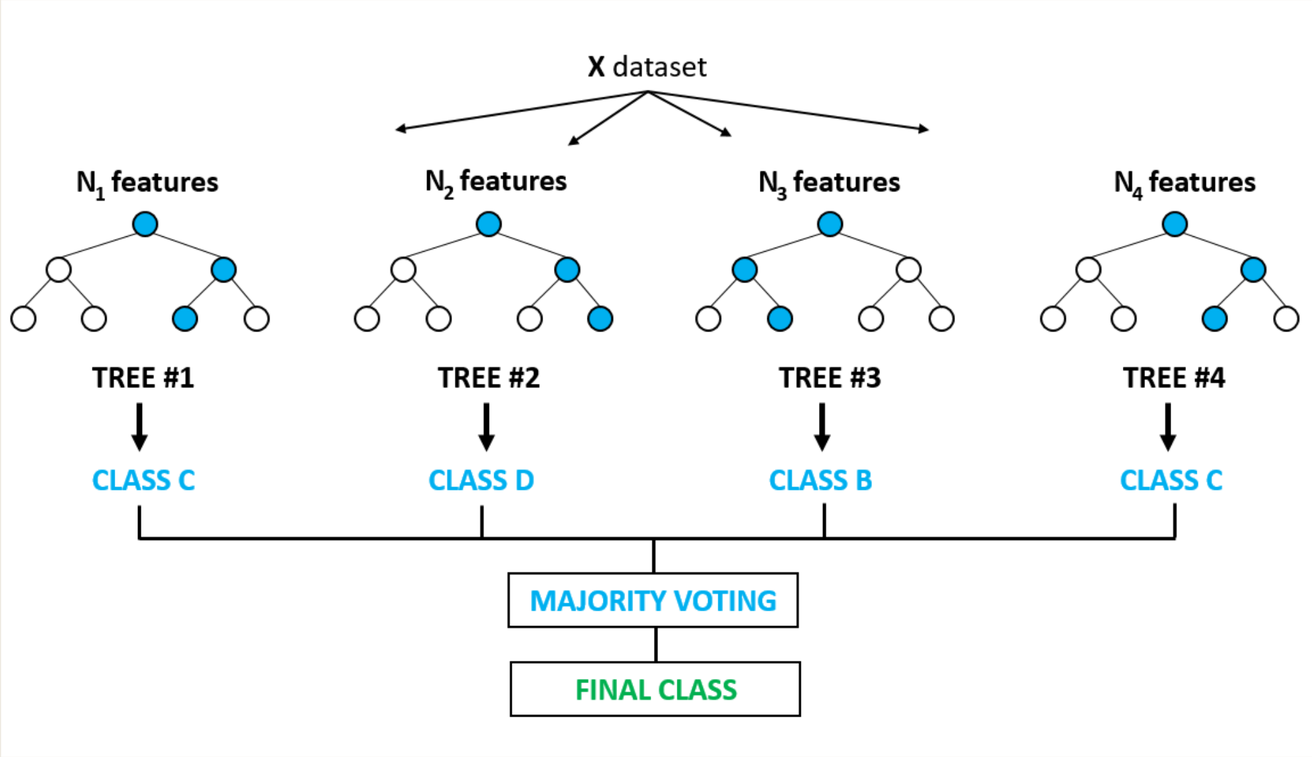
- t is very simple algorithm to understand and interpret.
- It is very useful for nonlinear data because there is no assumption about data in this algorithm.
- It is a versatile algorithm as we can use it for classification as well as regression.
- It has relatively high accuracy but there are much better supervised learning models than KNN.

Disadvantages

- It is computationally a bit expensive algorithm because it stores all the training data.
- High memory storage required as compared to other supervised learning algorithms.
- Prediction is slow in case of big N.
- It is very sensitive to the scale of data as well as irrelevant features.

Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.



Bagging Method

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression.

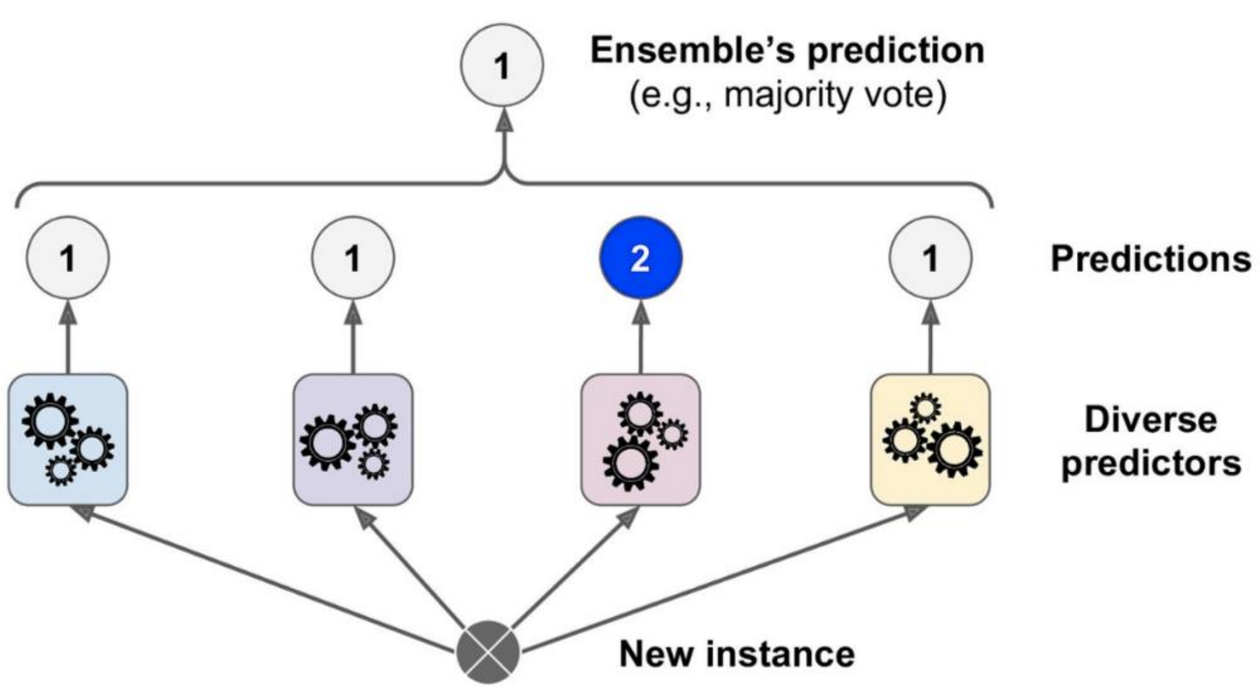


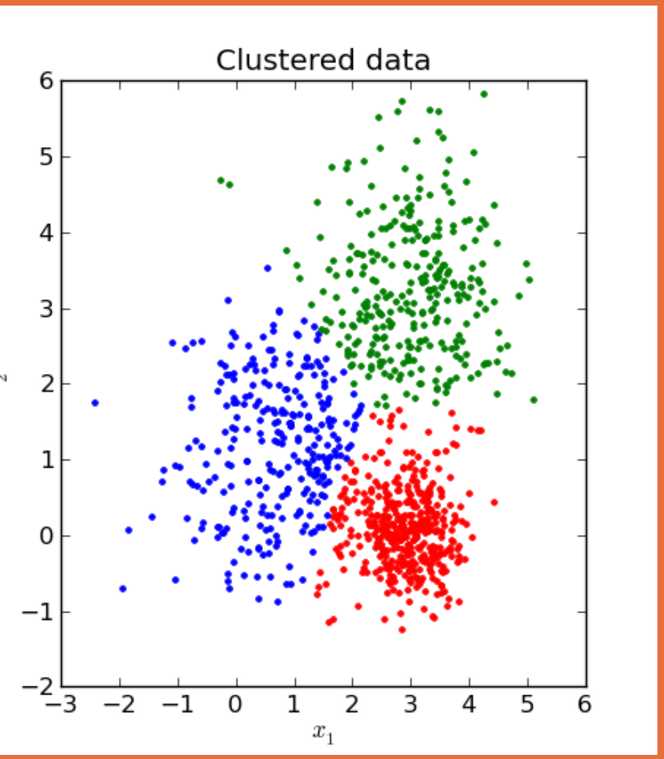
Figure 7-2. Hard voting classifier predictions

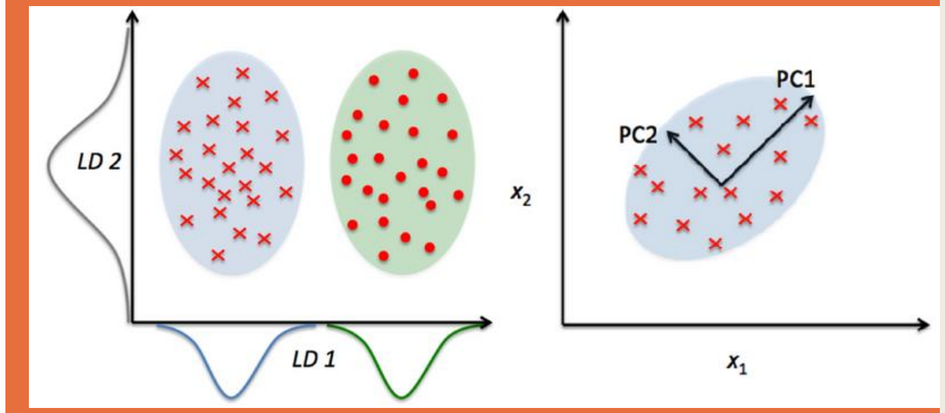
Advantages Random Forest

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance then single decision tree.
- Random forests are very flexible and possess very high accuracy.

Disadvantages Random Forest

- Complexity is the main disadvantage of Random forest algorithms.
- Construction and prediction of Random forests are much harder and time-consuming than decision trees.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.





Unsupervised Learning

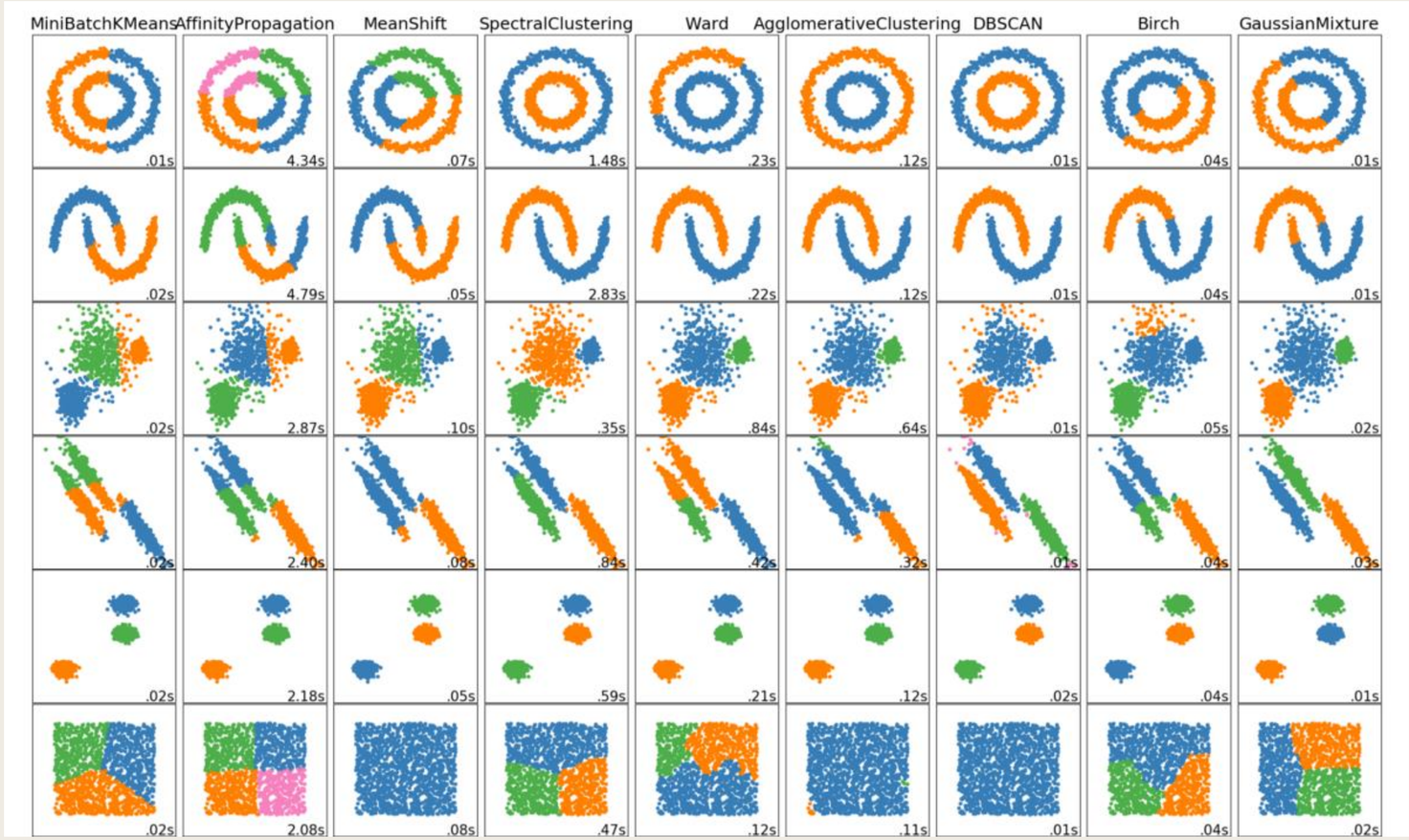
- K-Means Clustering
- Hierarchial Clustering

K Means Clustering

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by ‘K’ in K-means.

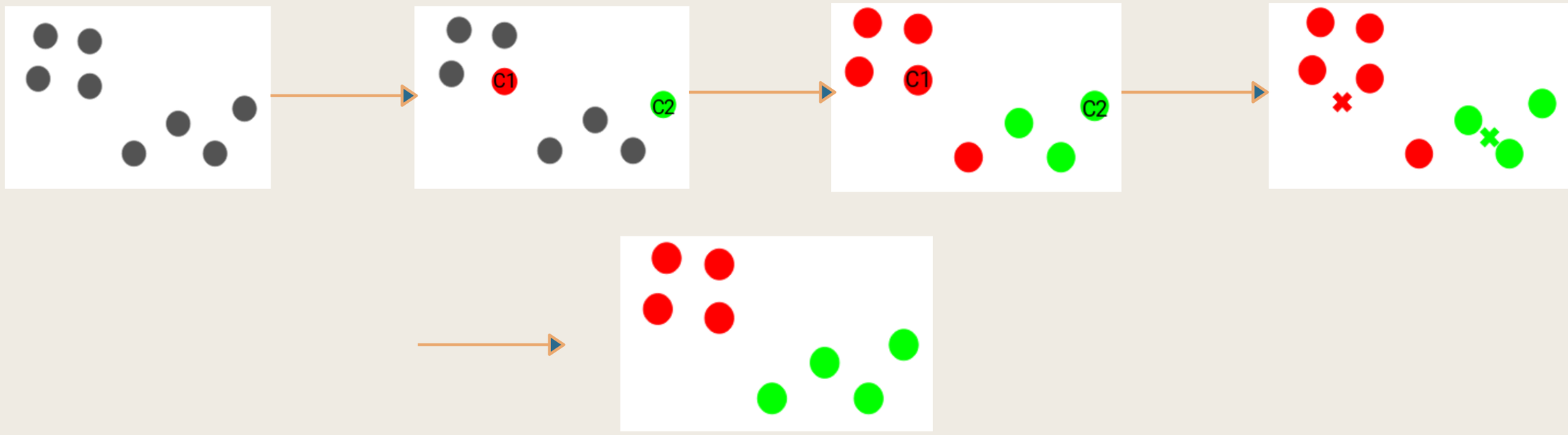
Main Goal

- To get a meaningful intuition from the data we are working with.
- Cluster-then-predict where different models will be built for different subgroups.



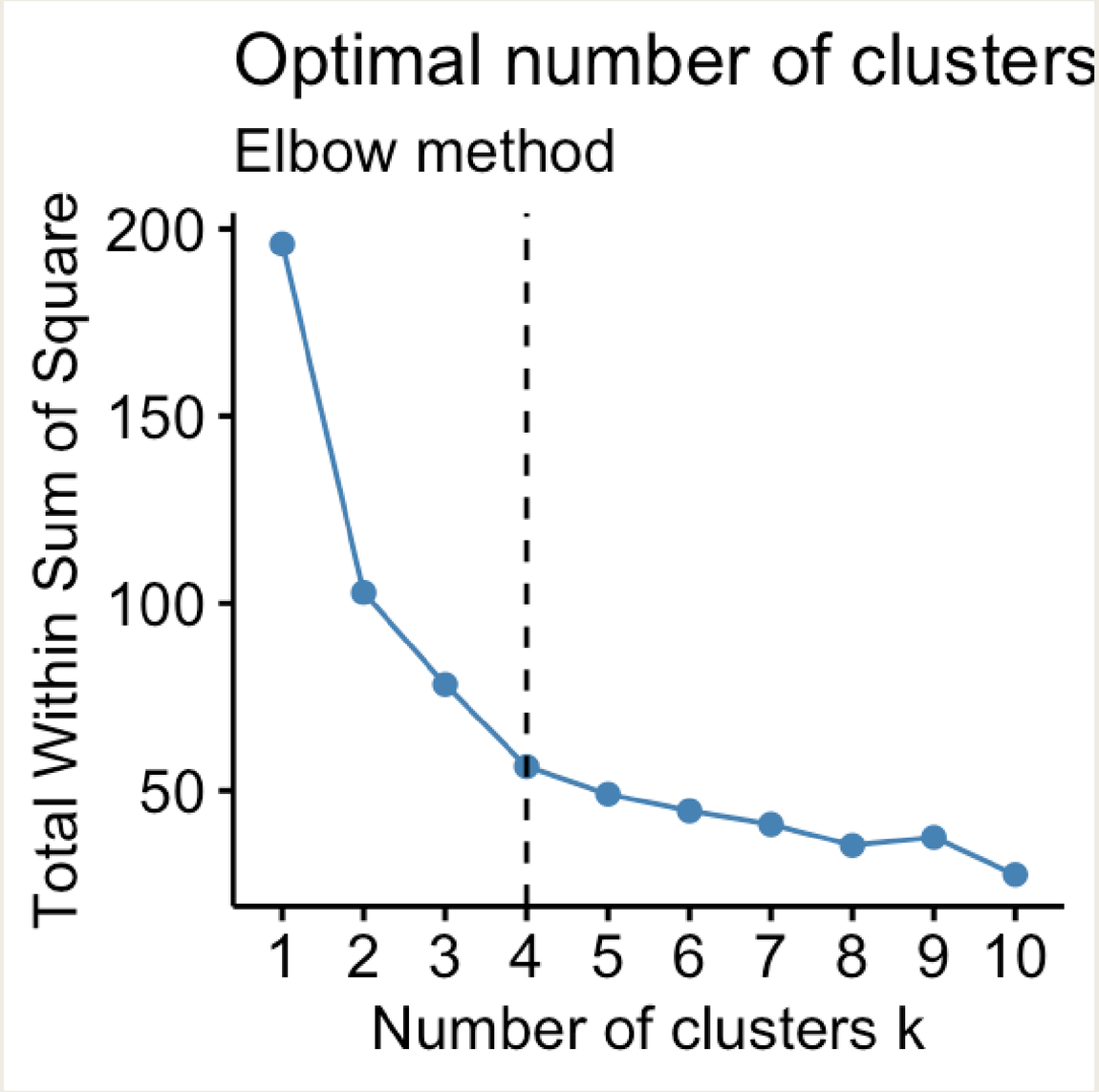
Process

- Step 1 – First, we need to specify the number of clusters, K, need to be generated by this algorithm.
- Step 2 – Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.
- Step 3 – Now it will compute the cluster centroids.
- Step 4 – Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more



Elbow Method

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.



Advantages K Means

- It is very easy to understand and implement.
- If we have large number of variables then, K-means would be faster than Hierarchical clustering.
- On re-computation of centroids, an instance can change the cluster.
- Tighter clusters are formed with K-means as compared to Hierarchical clustering.

Disadvantages K Means

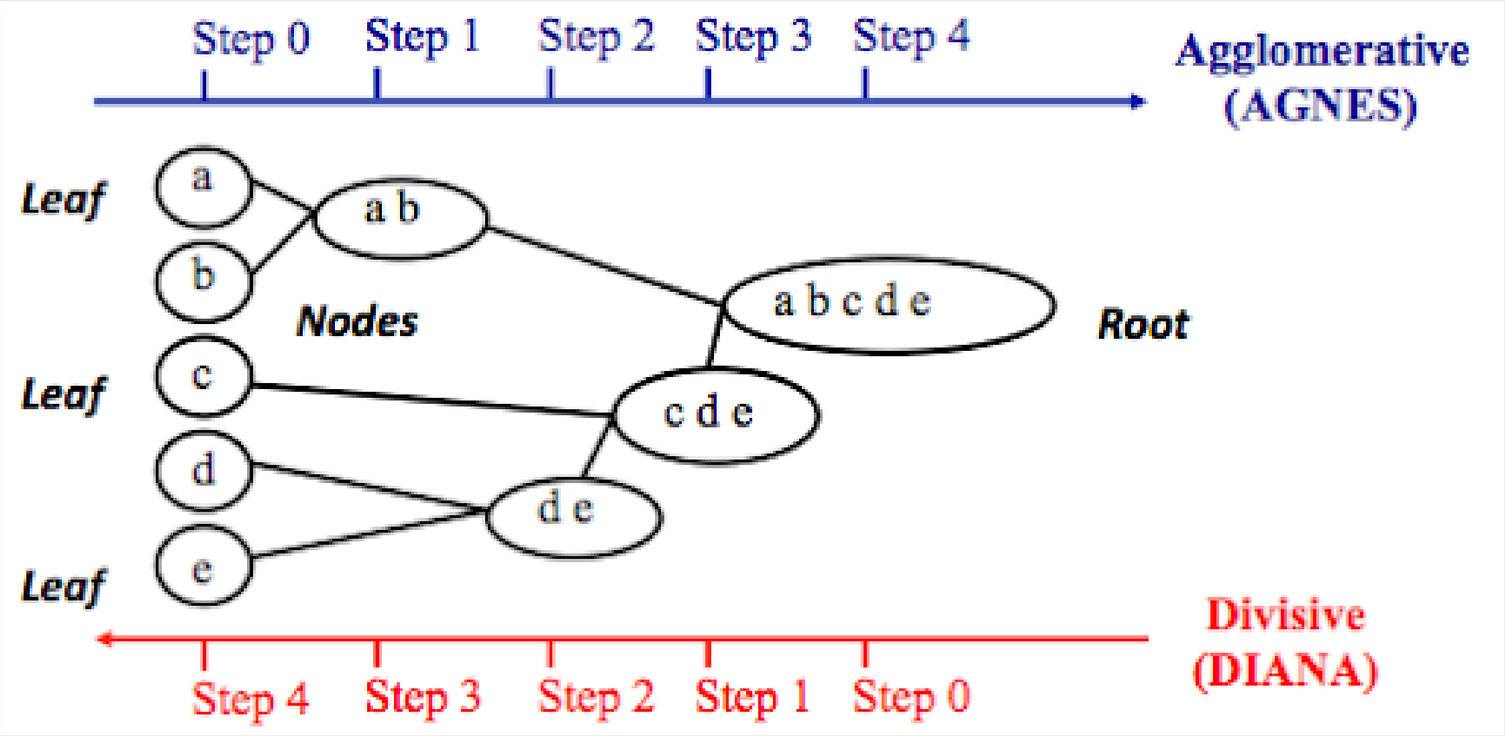
- It is a bit difficult to predict the number of clusters i.e. the value of k.
- Output is strongly impacted by initial inputs like number of clusters (value of k)
- Order of data will have strong impact on the final output.
- It is very sensitive to rescaling. If we will rescale our data by
- means of normalization or standardization, then the output will
- completely change.

Hierarchial Clustering

Hierarchical cluster analysis or HCA is an unsupervised clustering algorithm which involves creating clusters that have predominant ordering from top to bottom.

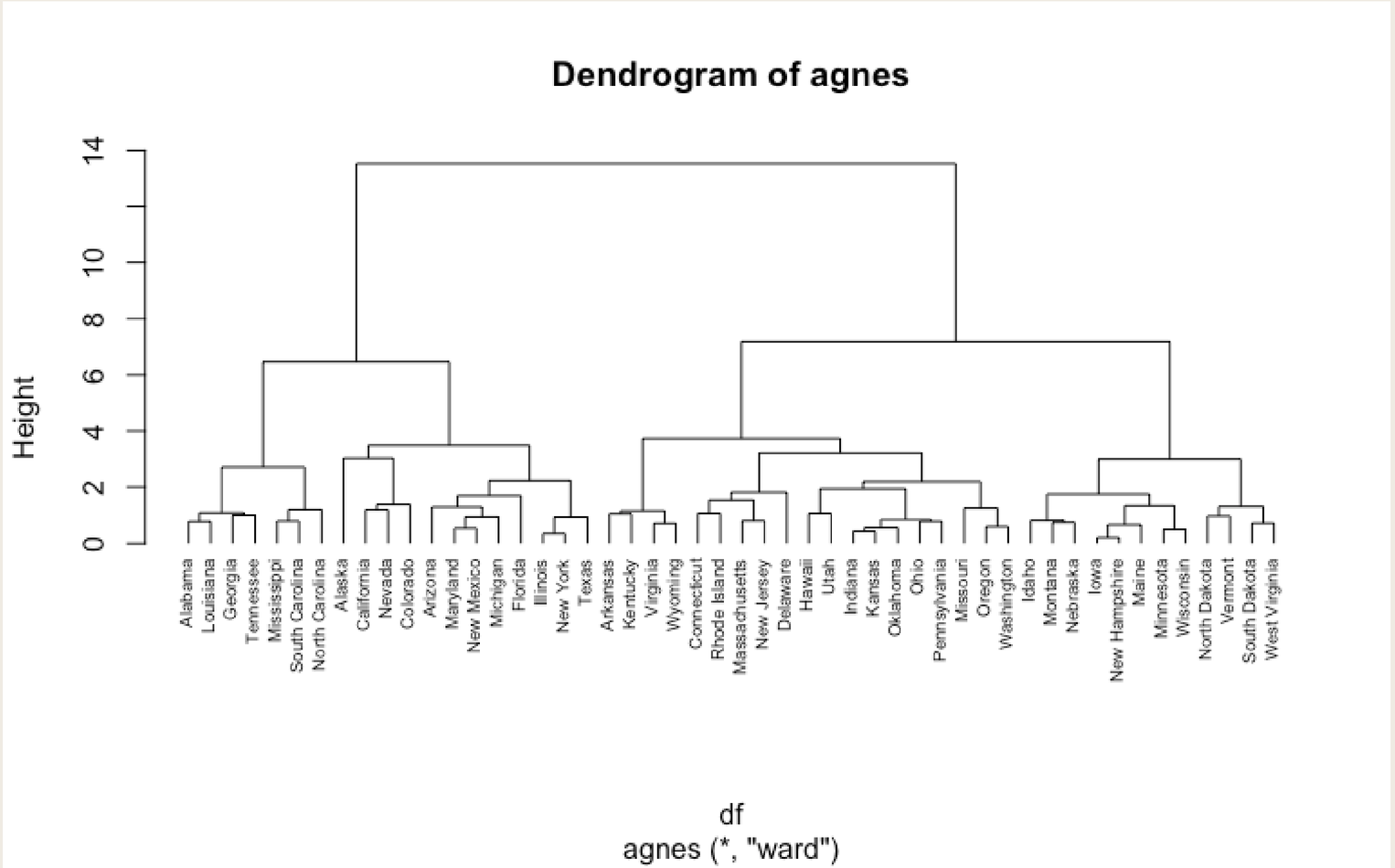
This clustering technique is divided into two types:

- Agglomerative Hierarchical Clustering
- Divisive Hierarchical Clustering



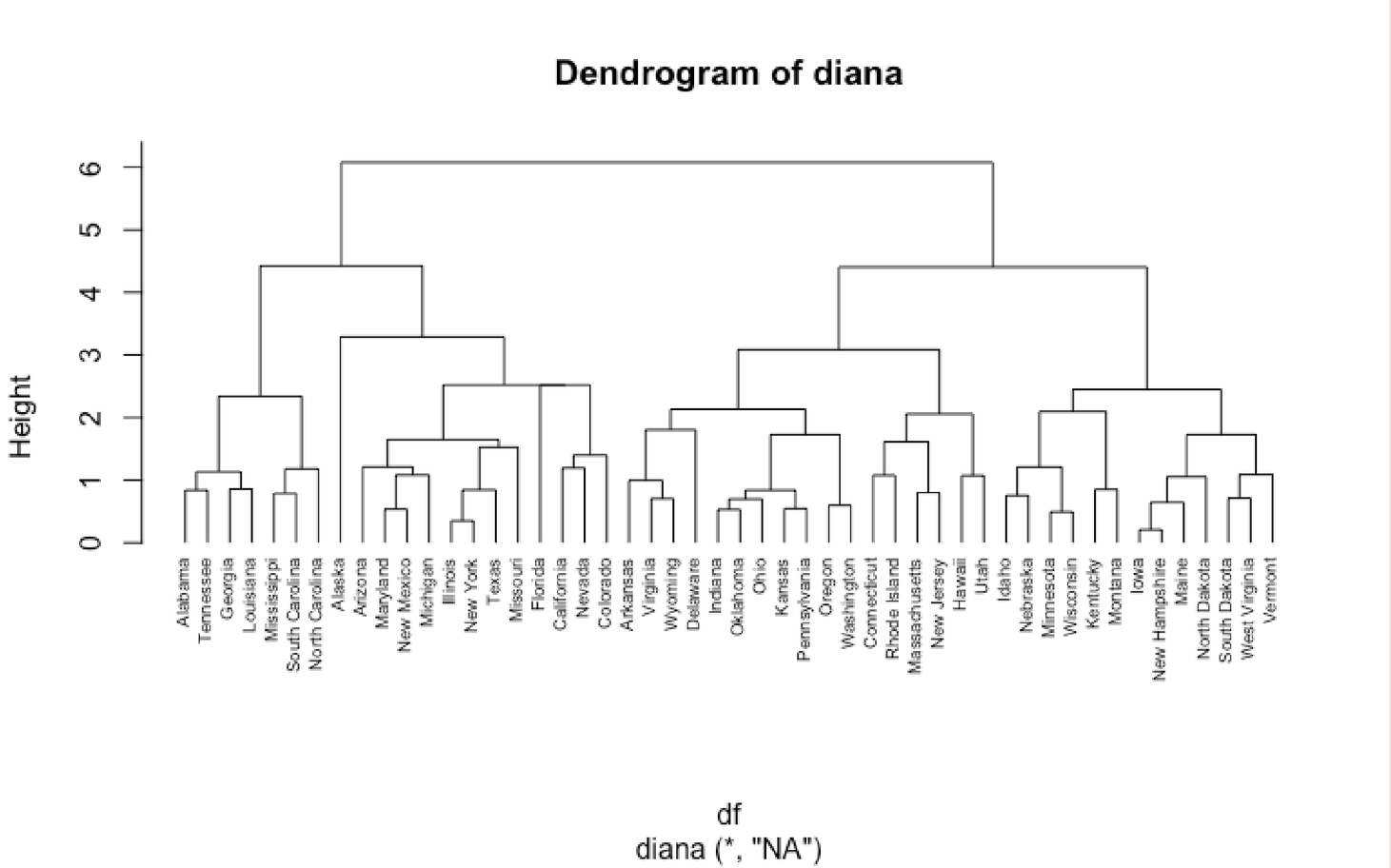
Agglomerative Hierarchical Clustering

In agglomerative hierarchical algorithms, each data point is treated as a single cluster and then successively merge or agglomerate (bottom-up approach) the pairs of clusters. The hierarchy of the clusters is represented as a dendrogram or tree structure.



Divisive Hierarchical Clustering

In divisive hierarchical algorithms, all the data points are treated as one big cluster and the process of clustering involves dividing (Top-down approach) the one big cluster into various small clusters.



Advantages Hierarchial Clustering

- No assumption of a particular number of clusters(i.e. k-means)
- May correspond to meaningful taxonomies

Disadvantages Hierarchial Clustering

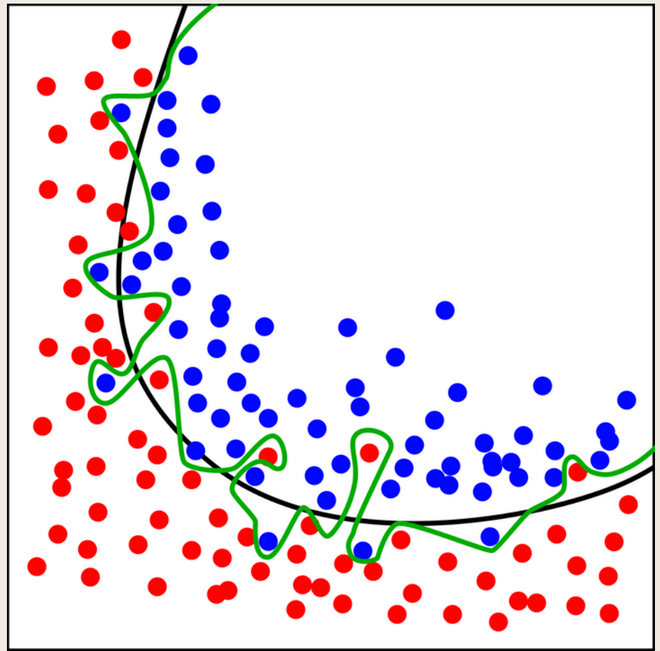
- HC is computationally expensive $O(N^2Log(N))$ hence is not recommended on massive datasets whereas k-means using linear timeIt is sensitive to noise and outliers.
- We have to define the number of clusters like k-means clustering algorithm

Cross Validation

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

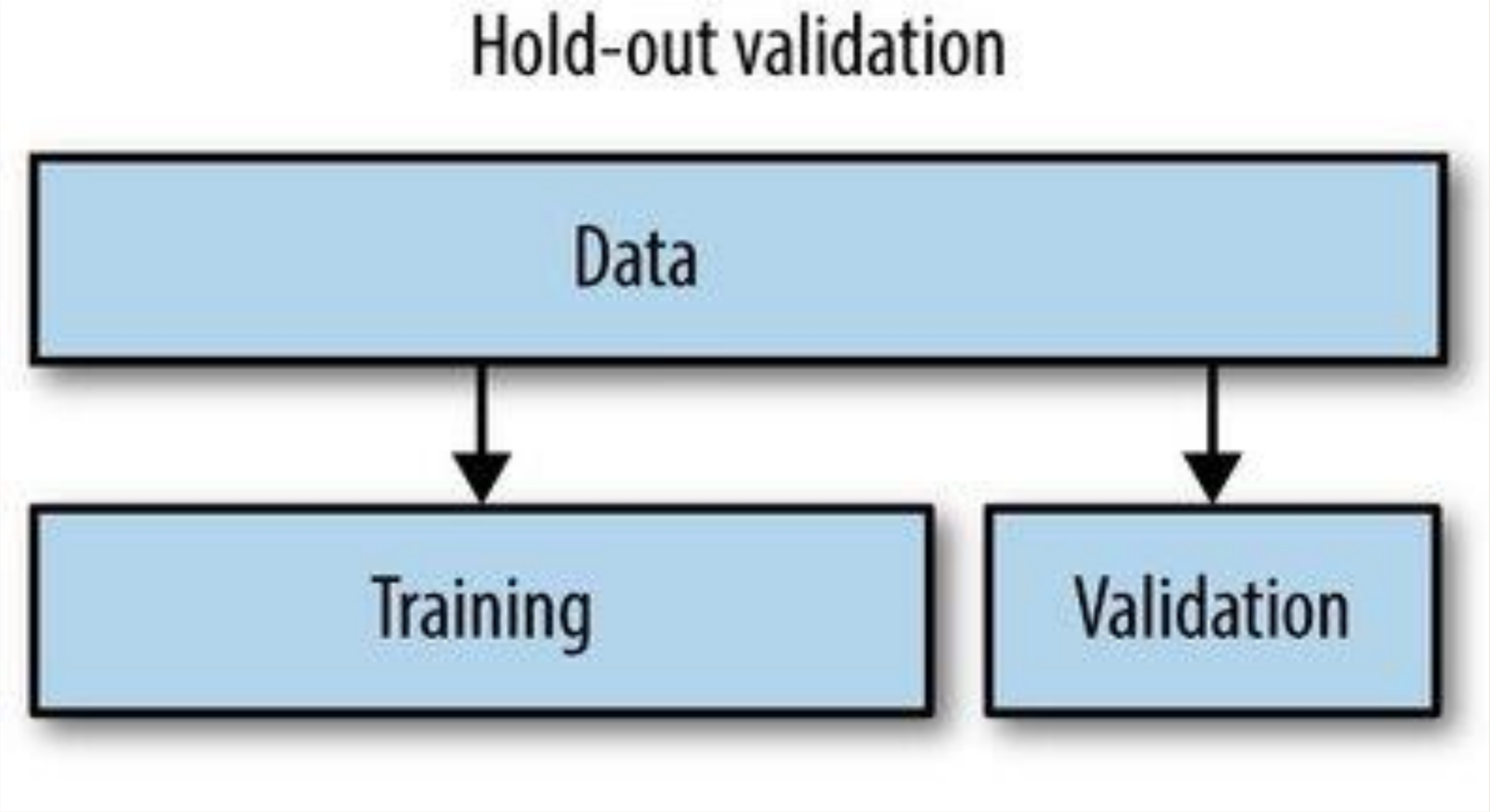
The three steps involved in cross-validation are as follows :

- Reserve some portion of sample data-set.
- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.



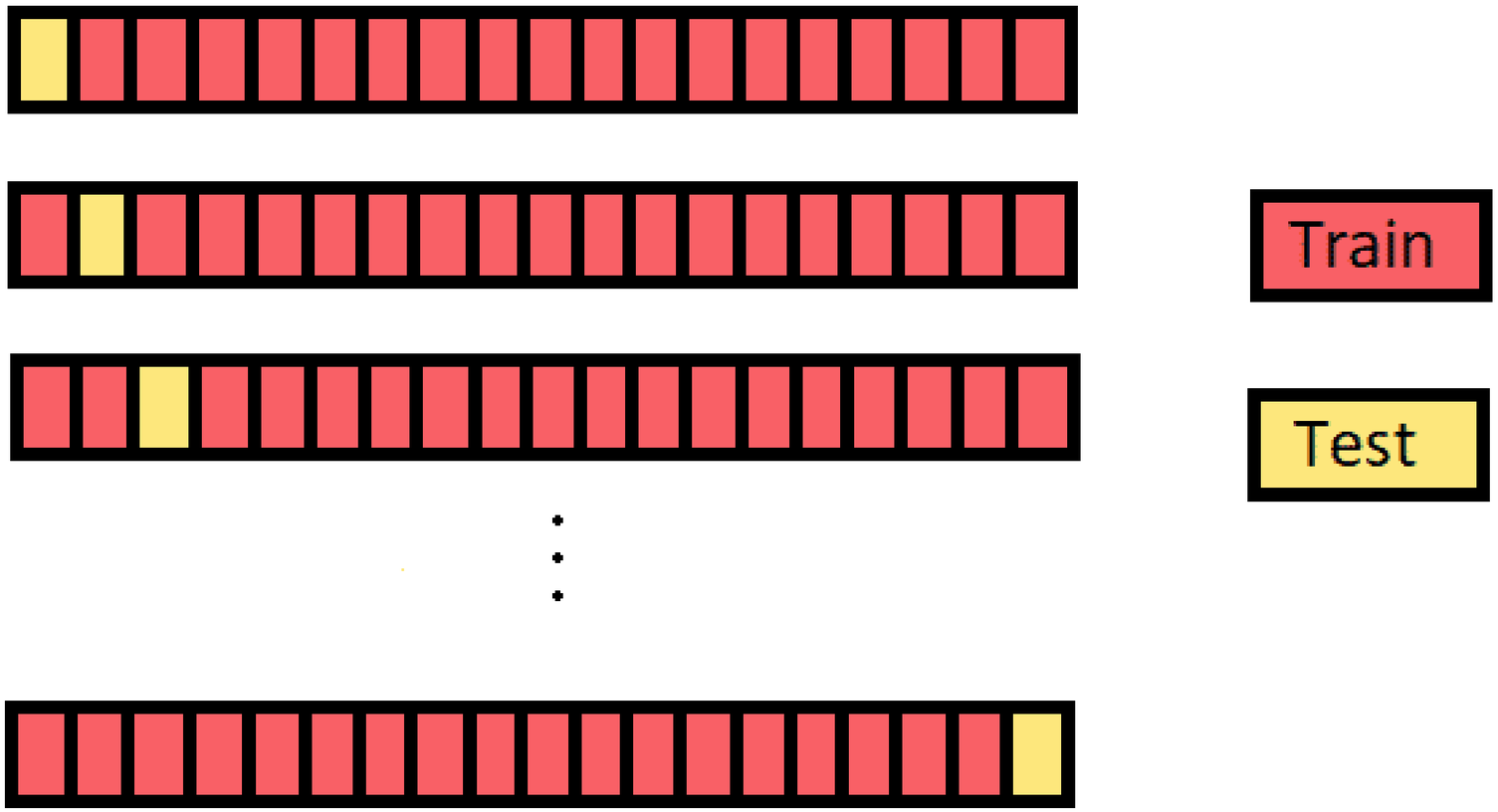
Hold Out Cross Validation

it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with four different combinations of predicted and actual values.



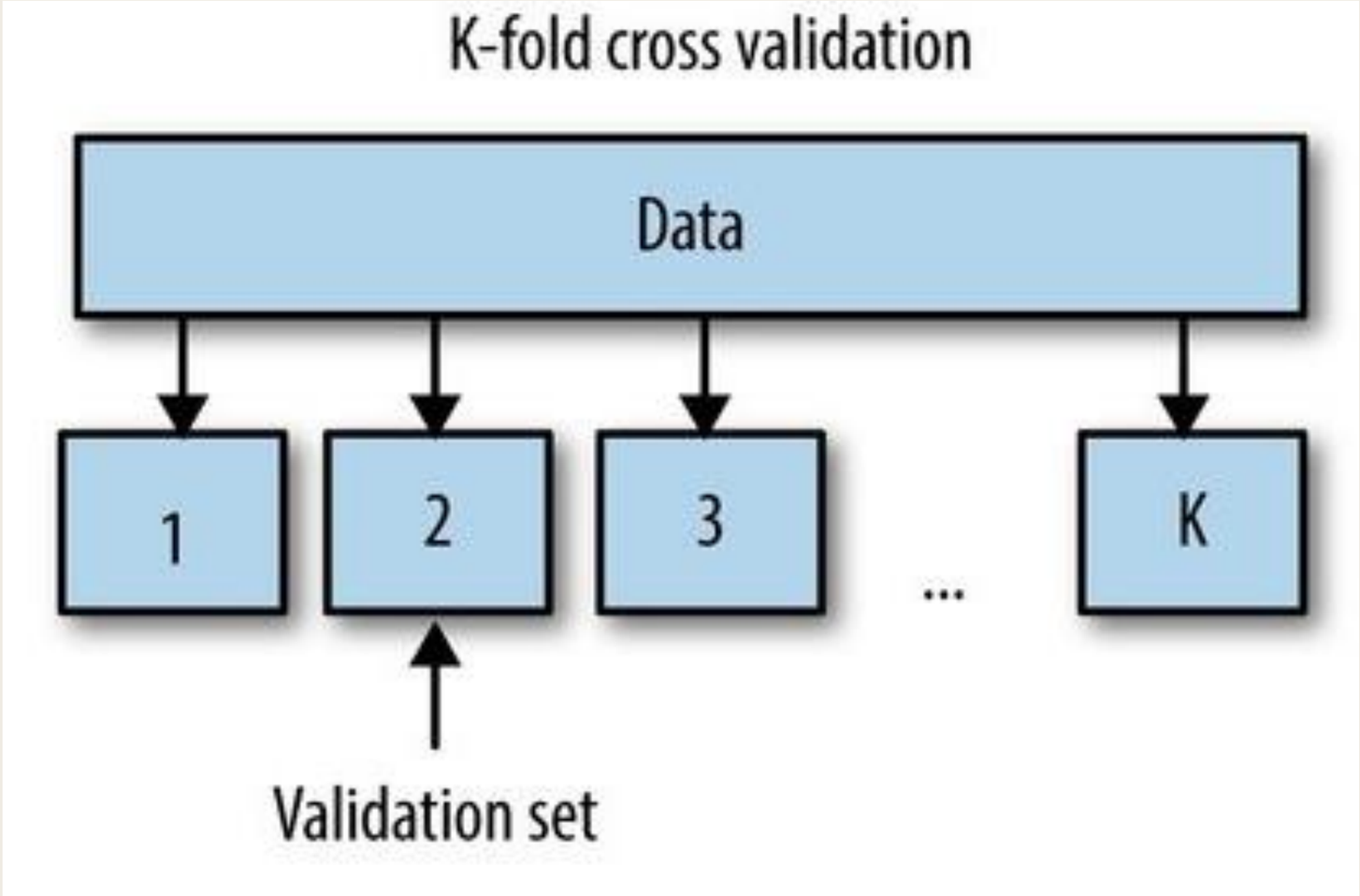
Leave One Out Cross Validation

it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with four different combinations of predicted and actual values.



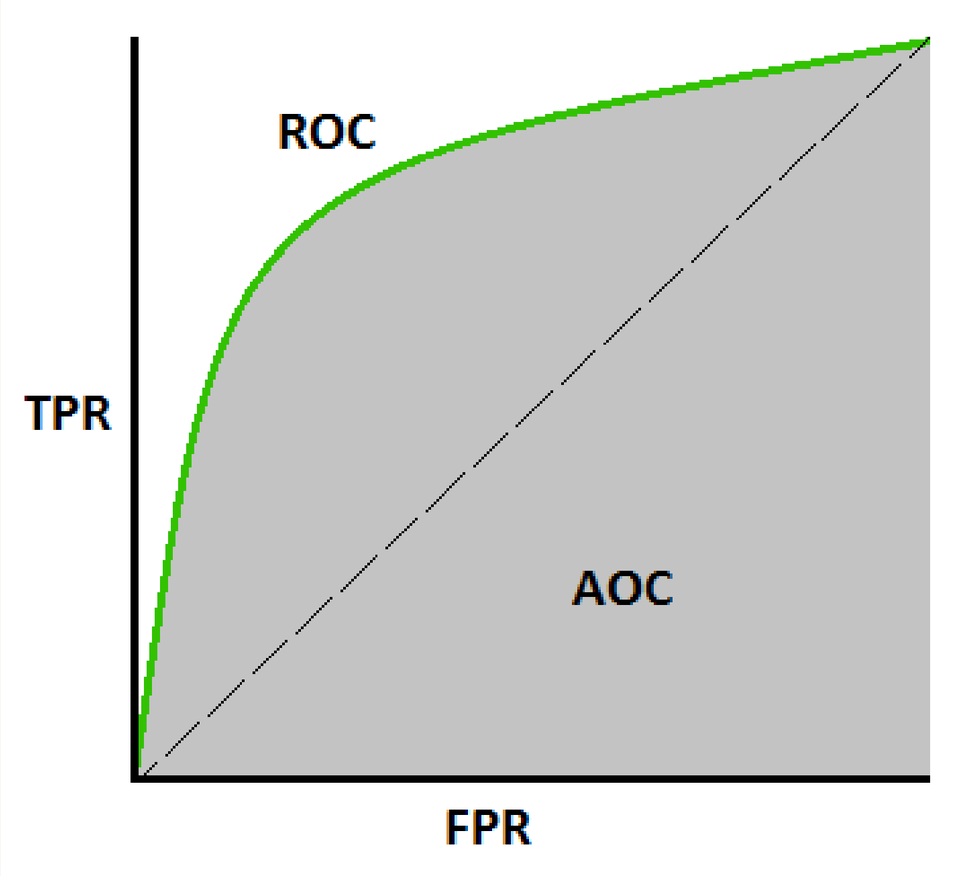
K-Fold Cross Validation

it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with four different combinations of predicted and actual values.



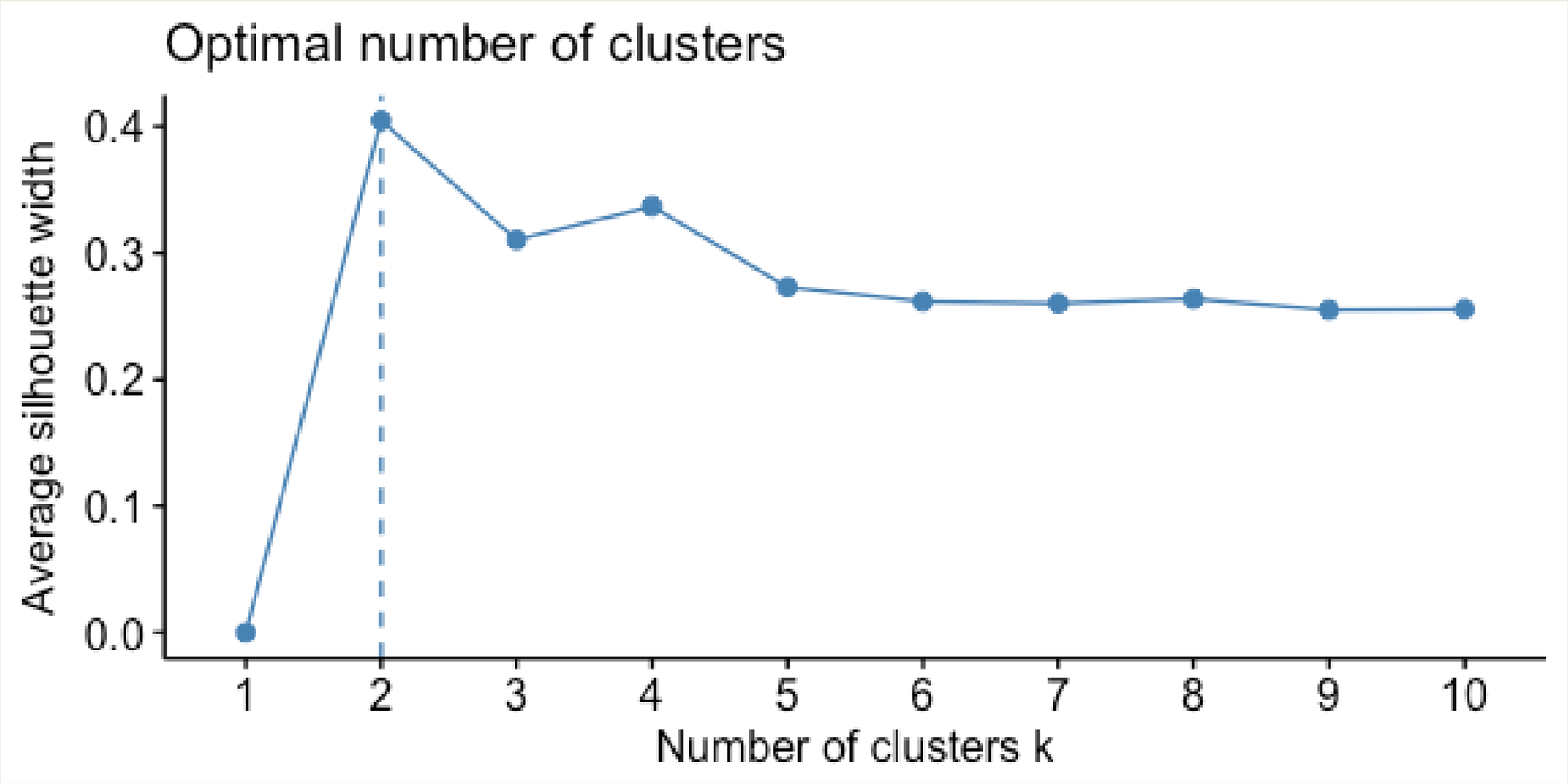
AUC & ROC

- An ROC curve (receiver operating characteristic curve) is a graphshowing the performance of a classification model at all classification thresholds
- AUC stands for "Area under the ROC Curve." That is, AUC measures theentire two-dimensional area underneath theentire ROC curve (think integral calculus) from (0,0) to (1,1).

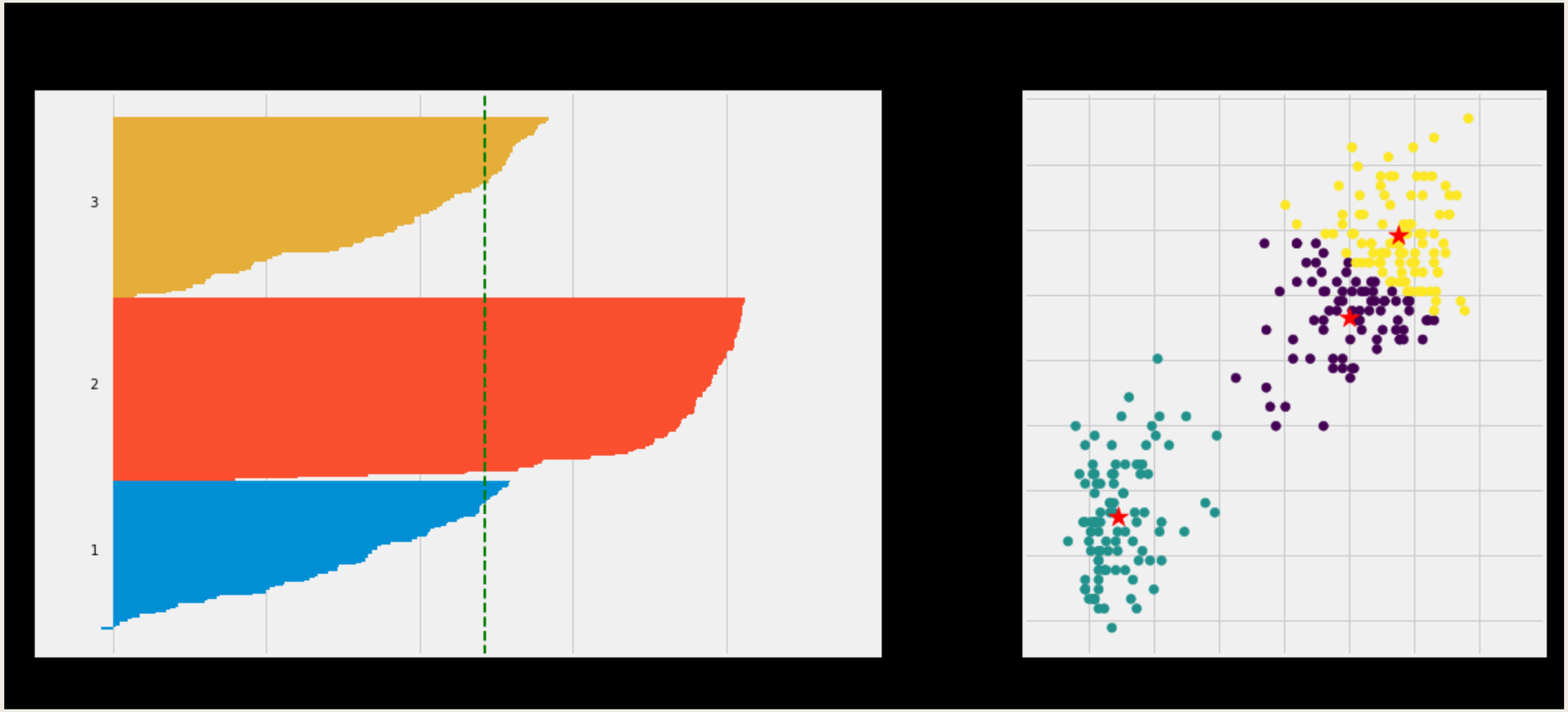


Shilouette Method

it is a performance measurement for machine learning clustering to determine number of cluster using elbow method too.



Shilouette Method



References

- www.tutorialspoint.com
- www.iykra.com
- Hands-OnMachine Learningwith Scikit-Learn& TensorFlow
- https://www.datavedas.com/linear-regression/
- www.towardsdatascience.com
- www.medium.com
- https://ml-cheatsheet.readthedocs.io
- https://www.geeksforgeeks.org
- https://www.analyticsvidhya.com
- https://uc-r.github.io/hc_clustering

Thank You!

KEEP LEARNING, STAY HUNGRY!