random forest akram assi

Machine Learning

Random forest is one of those algorithms which comes to the mind of every data scientist to apply on a given problem. It has been around for a long time and has successfully been used for such a wide number of tasks that it has become common to think of it as a basic need. It is a versatile algorithm and can be used for both regression and classification.

• Definition

The random forest is a supervised learning algorithm that generates and fuses several decision trees randomly into one "forest." The goal is not to focus on a single learning model, but rather on a set of decision models to improve accuracy. The key difference between this approach and the traditional decision tree algorithms is the random generation of the root nodes function splitting nodes

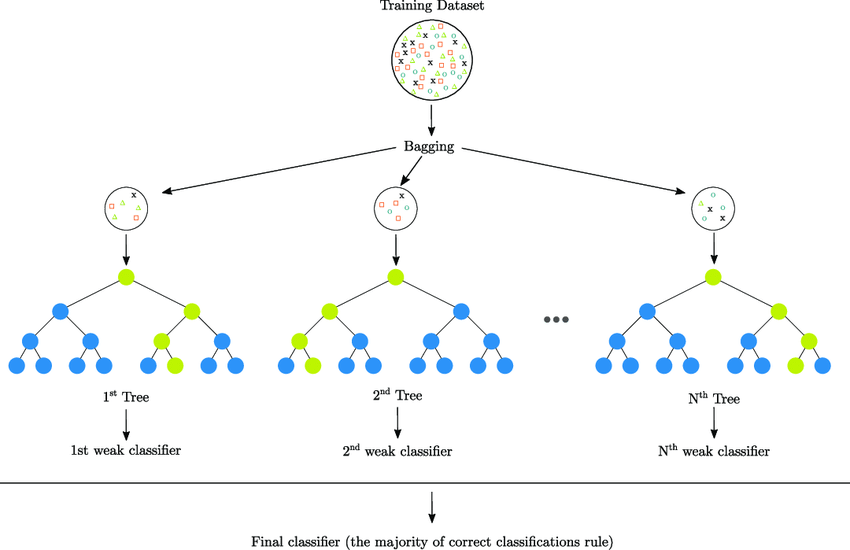
Why are you using the Random Forest ?

* It can be used for both classification and regression tasks.
* Classifiers may process missing values.
* Classifiers can also be modeled to represent categorical values.
* Overfitting is less likely to occur as more decision-making trees are added to the forest.

How does a random forest work ?

1. creating a random forest, follow the steps as follows:

* Randomly select “X” features from total “T” features where X < T
* Among the “X” features, calculate the node “S” using the best split point
* Split the node into daughter nodes using the best split method
* Repeat the previous steps until “F” number of nodes has been reached
* Build forest by repeating all steps for “n” number times to create “n” number of trees
* Model production process is fast

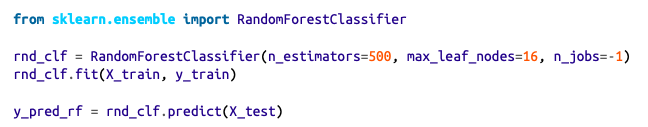


1. Experiment with random forest on test samples

Calculates each result of each sample:

* Calculate each predicted for all decision trees
* Choose the final prediction .determined by the purpose of the model (ex. mode, median, mean .....)





n\_estimators : is number of decision trees

max\_leaf\_node: depth of tree

n\_jops=-1 to use multicore

random forest vs decision trees and visualize tree

example Breast Cancer Dataset in jupyter notebook file

Advantages

* The predictive performance can compete with the best supervised learning algorithms
* They provide a reliable feature importance estimate
* They offer efficient estimates of the test error without incurring the cost of repeated model training associated with cross-validation

Disadvantages

* An ensemble model is inherently less interpretable than an individual decision tree
* Training a large number of deep trees can have high computational costs and use a lot of memory
* Predictions are slower