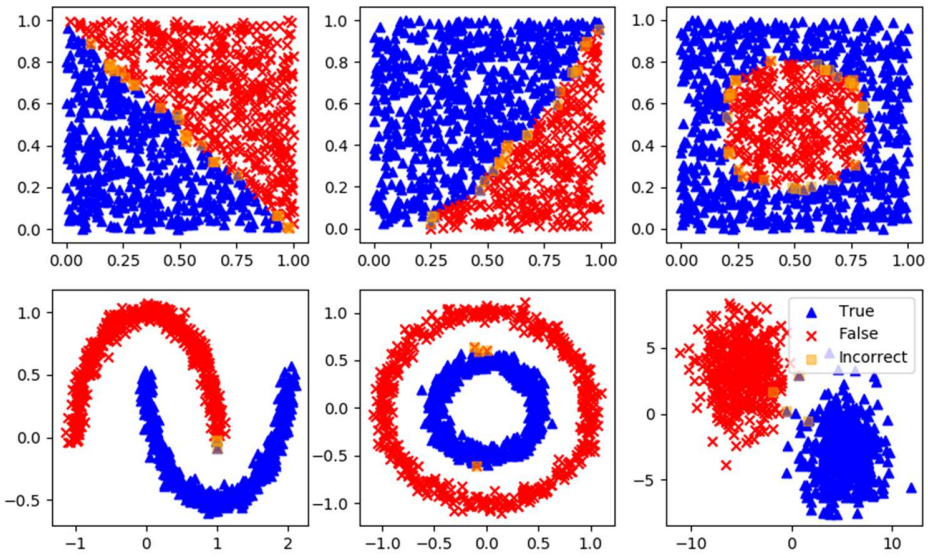


Random Forest Classifier

FROM CONCEPT TO ALGORITHM

Aryan Chauhan | ML Part - 2 | 29-07-2024

Table of Contents

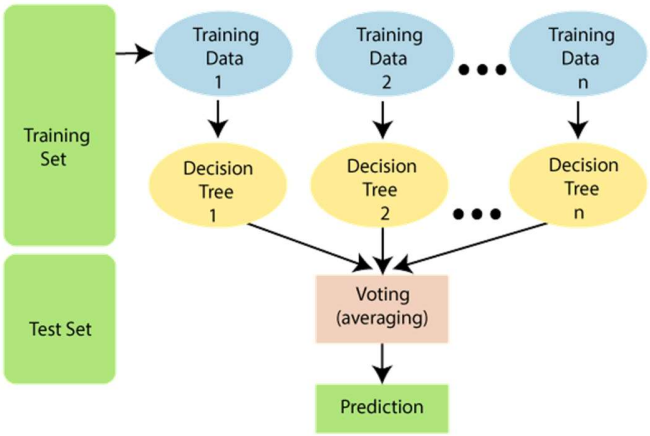
1. Introduction

Random Forest Classifier is a popular supervised machine learning algorithm. It is based on ensemble learning that involves the combination of multiple classifiers to solve complex problems while improving the accuracy of a model.

A Random Forest Classifier is a classification algorithm that contains n number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of the entire dataset.

Instead of relying on one decision tree, a random forest classifier considers the outputs of multiple decision trees and then calculates the majority vote of the decision tree to dictate its final decision.

No of trees are directly proportional to the accuracy of the model and inversely proportional to overfitting of data to the model.



2. Assumptions made in RFC

In a random forest classifier, obviously altogether as an entire unit, the decision trees will output the correct output. However, it is quite possible that internally every decision tree doesn't predict the correct output.

This means that certain trees are correct for the RFC model while others are not.

Hence, we need to make some assumptions and considerations to make sure that the decision tree is appropriate for the RFC

The considerations I prefer you make are as follows,

1. There should be some actual value in the feature variable of the dataset so that the classifier can predict actual values and not the guessed result.
2. The predictions from the trees should have less correlation with each other.

3. Why use RFC?

Training a Random Forest Classifier requires less time and is comparatively less complex.

Predictions made by the RFC are highly accurate even for large data sets.

I have observed RFC to also maintain a very good accuracy over datasets where a lot of data is missing.

These are the three main reasons why one should use Random Forest Classifier.

4. How does RFC work?

Random Forest Classifier works in two phases. First is to create the random forest by combining N decision trees and second is to make predictions for each tree created in the first phase. The working process of a RFC can be drawn into five steps :

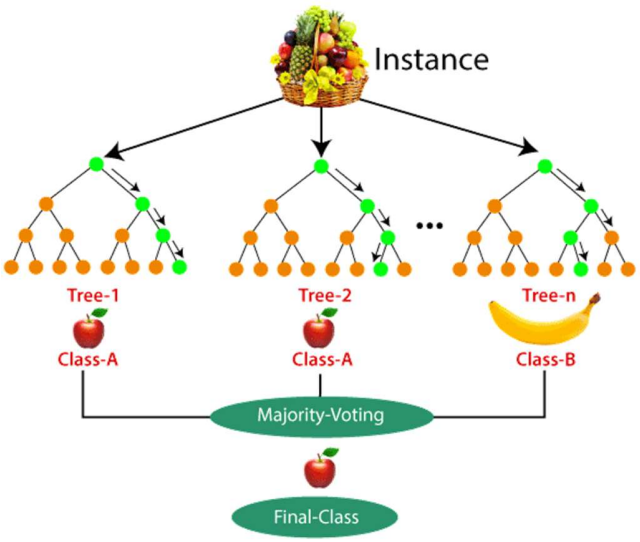
1. Select random K data points from the dataset.

2. Build a decision tree based on the subset that contains K data points.
3. Build N decision trees that will eventually create the architecture of the RFC
4. For new data points, find the predictions of each decision tree and assign the new data points to the category that wins the majority votes.

EXAMPLE

Consider that a fruit dataset is given to the random forest classifier. The data is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result and when a new data point occurs, then based on the majority of the result, the random forest classifier will assign the new data point to the fruit.

Check the below image to understand it better



5. Applications of RFC

1. Banking – Used for loan risk identification.
2. Medicine – Used to predict risk of disease and disease trends.
3. Land Use – Area of similar land use can be predicted.
4. Marketing – Marketing trends.

6. Advantages

1. RFC is capable of handling large datasets with a lot of dimensions.
2. It enhances the accuracy of the model while preventing overfitting.
3. RFC is used for both classification as well as Regression problems.
4. Feature importance is visible in RFC which allows tuning of features based on choice.
5. Parallelization is achieved in RFC leading to faster computation of data and less time complexity.

7. Disadvantages

1. Computational complexity to train multiple decision trees is high hence hardware core memory required is more.
2. While individual trees are easy to interpret, the ensemble of multiple trees can be seen as a black box model making it harder to interpret the overall decision process.
3. While the variance of the algorithm on the data reduces, the biasness can increase if each individual tree isn't deep enough.
4. RFCs are not recommended for extrapolation beyond the range of the training datasets i.e. they can perform well for the data in the range of training dataset but might not beyond the training dataset

8. Summary

In summary, a Random Forest Classifier is a powerful and widely used machine learning algorithm for classification tasks. It operates by combining multiple decision trees, each constructed from a subset of the training data. The final classification decision is made based on the majority voting from these individual decision trees. This ensemble approach enhances the model's accuracy and robustness by reducing overfitting and improving generalization to unseen data. Random Forest Classifiers are particularly valued for their ability to handle large datasets with higher

dimensionality and for providing insights into feature importance, making them a versatile and reliable choice in various classification applications.

End of Document