Random Forest

Abstract

This report consists of my basic understanding of one of the popular Ml methods ”Random Forest”.

1 Random Forest

Random Forest is a supervised machine learning algorithm used to solve classification and regression problems but more preffered for classification problems.It has quite an applications in bank sector,stock market and so on.It comes under non-parametric machine learning as the model will not assume any relationship between dependent and independent variables.As the name suggests, random forest simply generates decision trees on different training data subsets and takes their majority vote for predicting the outcome. This type of learning is known as ”**Ensemble learning**”.

2 Ensemble Learning

Ensemble learning is the process of combining multiple classifiers to solve a complex problem and to improve the performance of model.So in this case the final output predicted is based on the predictions of many individual classifiers. In general, ensemble learning uses two kinds of methods,

1) Bagging
2) Boosting

Bagging: In this method, the classifiers are trained with training samples which are subsets of training data which are randomly taken with replacement. So the same data points can be present in two or more classifiers.The output is based on majority voting.
Boosting : In Boosting the weak learners are combined squentially to form a strong learner.By the word sequential we mean each model tries to compensate the weakness of the predecessor.
In this case, random forest follows bagging method to do ensemble learning.

3 Understanding Random Forest

As we discussed earlier,Random forest generates the decision trees and predicts based on the outcomes of individual trees.So increasing the number of decision trees will decrease the possibilty of overfitting and increase the performance of model.The figure 3.1 shows how the outcome is predicted in random forest. The process of dividing the training data randomly into different subsets with replacements is known as row

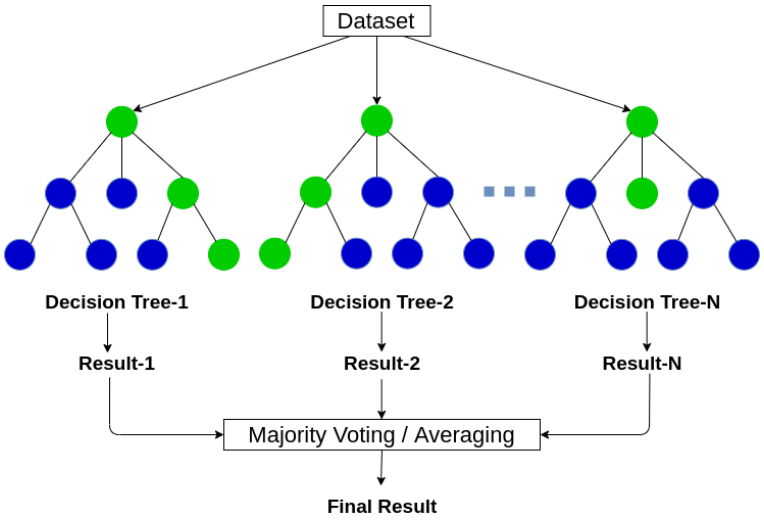

Fig. 3.1: Predicting the outcome by random forest

sampling.The step that involves row sampling is known as ”Bootstraping”.After each model trained independently, they all predict the output for a commom data and majority of them is taken as final output. This step is known as ”Aggregation”.

4 Advantages of Random Forest

1) Decision trees are tend to have risk of overfitting as they are more dependent on training data.So,by combining the decision trees we are able to decrease the risk of overfitting as averaging the outcomes of individual trees decreases the prediction error.
2) It runs efficiently on large data bases.
3) Robust against the outliers since they are averaged out in the aggregation step.

5 Disadvantages of Random Forest

1) It is hard to interpret the final outcome since we don’t know exactly what features are responsible for the output as in decision trees.
2) Sometimes it becomes hard to build a random forest due to limited computational power.

6 Miscellaneous

- Random forest expect it’s decision trees to have low correlation values among themselves.
- The features to be taken for decision trees should represent the target attribute since we are trying produce accurate results not to guess the result.

7 Quesions

1) What is ensemble learning?
2) Which method does random forest follow while doing ensmeble learning?
3) How random forest predicts the outcome of a data?
4) What will happen if the decision trees have high correlation values?
5) Give some advantages of random forest

8 Solutions

1) The ensemble learning is the process of combining multiple classifiers which helps in improving the performance of model.
2) The random forest follows bagging method in ensemble learning, it parallely trains the decision trees with data samples taken from data set with replacement.
3) Random forest averages/takes the majority of outcomes predicted by the individual decision trees.
4) If the decision trees with high correlation values predicting the outcome then there will be more variance and prediction error.As it will be no different from a single decision tree.
5) Refer the section 4