Random Forest
EPOCH IIT HYDERABAD

Rishitha Pallala

19 August 2023

1 Introduction

Random forests are machine learning models which heavily use decision trees. They are better than decision trees as the later has high inaccuracy while working on new data. They are good at handling continuous and categorical data, so they can be used for both classification and regression problems

2 Bootstrap Aggregation

The process is also known as bagging .The working of the random forest requires creation of decision trees and data sets. the process is outlined below:

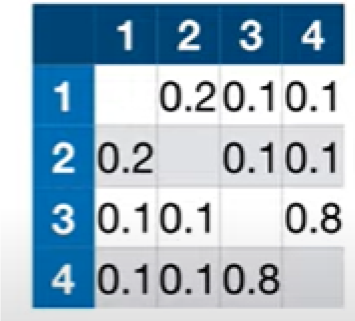
1. Creating bootstrapped data sets of the same size as original data , but duplicates are allowed . Usually, 33% of the original data doesn't make it to the bootstrapped data set.

2. Building decision trees, we build decision trees for the bootstrapped data, by considering a random subset of variables at each step.

3. Calculating the accuracy: we test the decision trees with "out of the bag " samples , and consider the result given by many decision trees as the final classification. The accuracy is directly proportional to the prediction of the out of the bag samples.

3 Proximity Matrix And Weighted Frequencies

Missing data is common in original(used for training and testing) and new data sets .



	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

1. To estimate the missing data in original data set , we initially assume the missing data as the majority in classification and average in regression.
2. Later , we run the data through all the decision trees and create a proximity matrix with probabilities that both the values end at the same leaf node.
3. we calculate the weighted frequencies(WF) given by equation (1) for classification , and equation (2) for regression;Where f is the frequency of the value, and w is the weight. these substitute the missing data.
4. we iterate the process a few times , so get the most accurate value,similar iterative methods are used for new data also.

$$WF_c = f_o w_o + f_1 w_1 \tag{1}$$

$$WF_R = \sum_{i=1}^{n} f_i w_i \tag{2}$$

Thus, random forest can be used when we have a lot of features and it's stability due to majority voting is it's selling point.