Random Forest
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

19 Jul 2022

1. Introduction

Random forest is a supervised machine learning algorithm that is used for both classification and regression. As the name suggests, a random forest model constructs multiple decision trees and aggregates all of their outputs for assigning labels to a data point. Even though random forest sacrifices the interpretability of decision trees, their accuracy is much higher.

2. Bootstrap Aggregating

Suppose we are given m labelled samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$. An algorithm called bootstrap aggregation, also known as bagging, is used to generate a random forest. The algorithm is as follows.

i) We create a bootstrapped dataset of size m from these samples. A bootstrapped dataset is obtained by random sampling from a given dataset with replacement, i.e., repetitions are allowed.
ii) This bootstrapped database is now used to create a decision tree. However, the decision tree is created by considering a random subset of a fixed size k of the features at each decision node. In other words, instead of comparing the Gini impurity indices of all the features for choosing the best classifying feature, compare the indices for only some k random features chosen at every step and use one of these for the decision node. For a classification tree, k is usually chosen to be $\sqrt{n}$ and $\frac{n}{3}$ for regression trees.
iii) Repeat the above steps several times choosing a different bootstrapped dataset each time. This will give us multiple uncorrelated decision trees, each built using a different set of samples and using different features for each decision node. This causes the trees to have a lot of variety and helps improve the accuracy of the model compared to a singlar decision tree.

Now that we have our random forest ready, we can assign labels to a query data point by aggregating the outputs of all the decision trees in the forest. In case of classification, we can simply take a majority vote and in case of regression, we can consider the average of all outputs.

3. Cross-Validation

For assessing the performance of a random forest model, the out-of bag error is commonly used. For each tree in the forest, we consider the out-of-bag dataset, the dataset containing the samples that were not selected in the bootstrapped database. We run all such out-of-bag samples through every tree that didn't use those samples and count the number of times these samples were misclassified. The proportion of out-of-bag samples that were misclassified is the out-of-bag error.

This out-of-bag error can also be helpful in determining the value of k - the size of the random subset of features chosen at every decision node. We iterate over some values of k starting from an initial guess (usually $\sqrt{n}$) and choose the one that minimizes the out-of-bag error.

4. Questions

i) Define a forest.
ii) Why does random forest perform better than decision tree?
iii) Why is it important to have a large number of trees?
iv) What is the purpose of creating bootstrapped datasets?
v) Can random forest be used for unsupervised learning?

5. Answers

i) A forest is a disjoint union of one or more trees, i.e. an acyclic undirected graph. In other words, a forest is an undirected graph in which any two vertices are connected by at most one path.
ii) Random forest performs better than decision tree because it reduces the variance of the model while keeping the bias roughly the same. Noisy data can affect the decision rule of a tree drastically, but when averaged over multiple trees, its effect is negligible.
iii) Because the random forest algorithm is based on bagging, if there are a small number of trees, some samples may be insufficiently represented in the final model as the samples are randomly chosen for forming trees.
iv) Bootstrapping is done to ensure that the trees we obtain are not correlated as a different subset of samples is being chosen everytime. Having several correlated trees would increase the variance of the model.
v) Random forest can be used for unsupervised learning provided we are given a dissimilarity measure that can quantify how similar or dissimilar two input data points are. The random forest can then use decision rules that will separate dissimilar data and classify similar data into the same category.