By Paul Naidoo, EOLAS

Random Forests

Random Forests are an ensemble learning method primarily used for classification and regression tasks.

Decision Trees

The Decision Tree (DT) is the building block of the Random Forest method. A DT is a means of modelling data that splits data into subsets based on feature values, creating a tree-like structure of decisions.

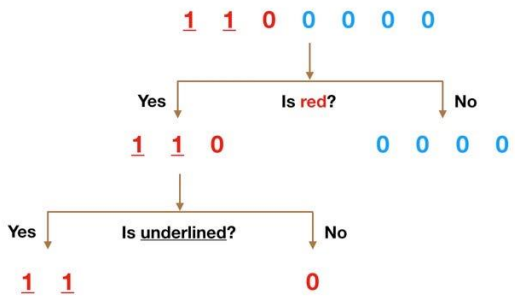

Figure 1: a simple Decision Tree (source)

Figure 1 illustrates a simple Decision Tree. Our data, at the top, is sorted by a series of 'nodes': conditions by which the data can be split into groups. Based on the answer to the (binary) question at each node, data is sorted along one or the other 'branch'. In the example above, data is split into various categories, each one referred to as a 'leaf'.

How Random Forests Work

1. **Decision Trees**: Random Forests operate by constructing multiple Decision Trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.
2. **Bootstrap Aggregation (Bagging)**: Random Forests use a technique called bagging to create multiple decision trees. Bagging involves generating multiple subsets of the original dataset by sampling with replacement. Each subset is used to train a separate decision tree.
3. **Random Feature Selection**: When constructing each tree, Random Forests introduce randomness by selecting a random subset of features at each split. This helps in creating diverse trees and reduces the risk of overfitting.

4. **Voting/Averaging**: Once all the trees are constructed, the Random Forest makes predictions by aggregating the results of the individual trees. For classification tasks, it uses majority voting, where the class that gets the most votes is the final prediction. For regression tasks, it averages the predictions of all the trees.

Advantages of Random Forests

- **Robustness**: By averaging multiple trees, Random Forests reduce the risk of overfitting and improve generalization.
- **Feature Importance**: They can provide estimates of feature importance, helping in understanding which features (variables) are most influential in making predictions.
- **Versatility**: Random Forests can handle both classification and regression tasks and are effective with large datasets and high-dimensional spaces.

Limitations

- **Complexity**: The model can become complex and computationally intensive with a large number of trees.
- **Interpretability**: While individual decision trees are easy to interpret, the ensemble of many trees can be less transparent.