BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI – HYDERABAD CAMPUS
BITS F464 : MACHINE LEARNING
FIRST SEMESTER 2023-2024

A worked-out example on Random Forests

Here's a step-by-step solved numerical problem on the random forest classifier.

Let's assume the following dataset that involves classifying different types of flowers based on their petal length and width :

Petal Length	Petal Width	Flower Type
1.4	0.2	Setosa
1.3	0.3	Setosa
3.1	1.5	Versicolor
3.9	1.1	Versicolor
5.7	2.0	Verginica
5.9	1.8	Verginica

We want to have the final prediction for a test data point using the majority vote of the multi-class classification outputs by multiple decision trees.

Step 1: Building Random Forest.

Let's say we decide to build a Random Forest with two decision trees. The algorithm will randomly select subsets of the data and features to build each tree. For simplicity, let's assume the first decision tree uses petal length and the second decision tree uses petal width.

Step 2: Building the Decision Trees.

1. Let's consider the first decision tree that uses petal length as the deciding feature.

Calculate Gini Impurity for the root node:

- Total instances: 6
- Setosa instances: 2
- Versicolor instances: 2
- Virginica instances: 2

$$\text{Gini Impurity} = 1 - (P(Setosa)^2 + P(Versicolor)^2 + P(Virginica)^2$$
$$= 1 - ((^2/_6)^2 + (^2/_6)^2 + (^2/_6)^2)$$
$$= {}^2/_3$$

Split based on petal length <= 3.1:

- Instances on the left: 3
- Instances on the right: 3

Calculate Gini Impurity for left node (petal length <= 3.1):

- Setosa instances: 2
- Versicolor instances: 1
- Virginica instances: 0

$$\text{Gini Impurity} = 1 - ((2/3)^2 + (1/3)^2 + (0/3)^2)$$
$$= 4/9$$

Calculate Gini Impurity for right node (petal length > 3.1):

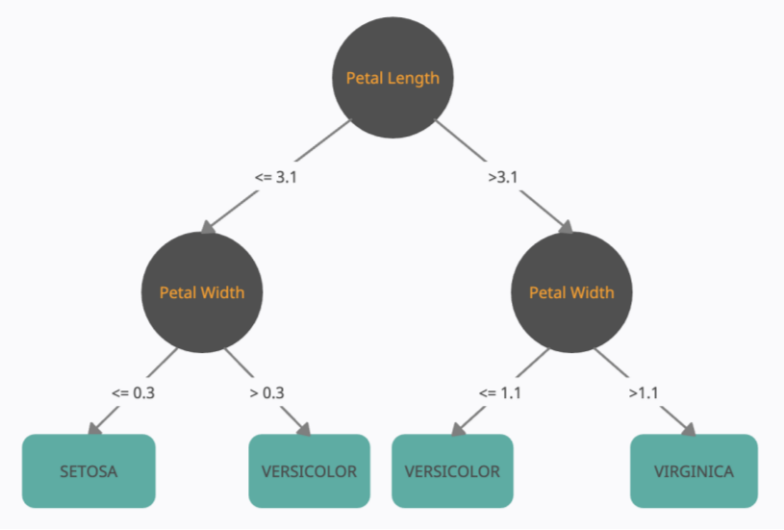
- Setosa instances: 0
- Versicolor instances: 1
- Virginica instances: 2

$$\text{Gini Impurity} = 1 - ((0/3)^2 + (1/3)^2 + (2/3)^2)$$
$$= 4/9$$

Calculate Information Gain:

$$\text{Information Gain} = \text{Parent Gini Impurity} - (\text{Weighted Average Gini Impurity of Child Nodes})$$
$$= 2/3 - ((3/6) \cdot (4/9) + (3/6) \cdot (4/9))$$
$$= 2/9$$

The decision tree continues to split based on other features, and we calculate Gini Impurity and Information Gain at each node. In the end, we'll have the first decision tree as follows:



2. Now consider the second decision tree that uses petal width as the deciding feature.

Split based on petal width <= 1.5:

- Instances on the left: 4
- Instances on the right: 2

Calculate Gini Impurity for left node (petal width <= 1.5):

- Setosa instances: 2
- Versicolor instances: 2
- Virginica instances: 0

$$\text{Gini Impurity} = 1 - (({}^{2}/_{4})^{2} + ({}^{2}/_{4})^{2} + ({}^{0}/_{4})^{2})$$
$$= {}^{1}/_{2}$$

Calculate Gini Impurity for right node (petal width > 1.5):

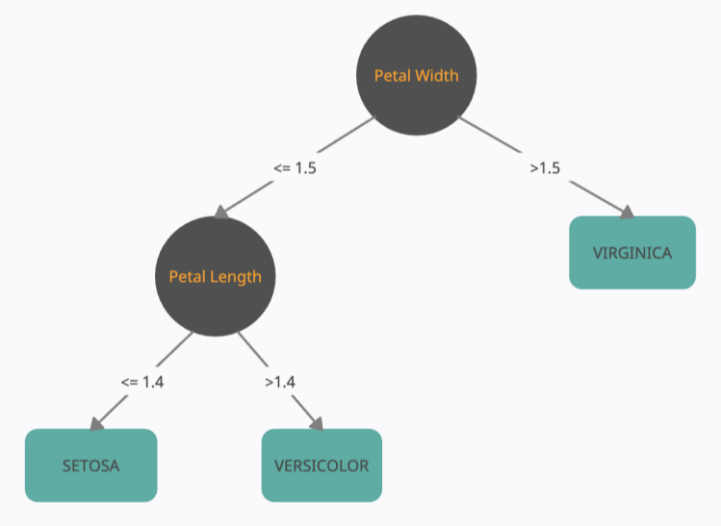
- Setosa instances: 0
- Versicolor instances: 0
- Virginica instances: 2

$$\text{Gini Impurity} = 1 - (({}^{0}/_{4})^{2} + ({}^{0}/_{4})^{2} + ({}^{2}/_{4})^{2})$$
$$= {}^{3}/_{4}$$

Calculate Information Gain:

$$\text{Information Gain} = {}^{2}/_{3} - (({}^{4}/_{6}) \cdot ({}^{1}/_{2}) + ({}^{2}/_{6}) \cdot ({}^{3}/_{4}))$$
$$= {}^{1}/_{12}$$

Finally, we'll have the second decision tree as follows:



Step 3: Majority Voting.

Let's perform majority voting for a test data point using the two decision trees we've built.

Suppose we have a test data point with the following characteristics:

- Petal Length: 4.2
- Petal Width: 1.6

Decision Tree 1: Based on Decision Tree 1, we follow the splits as described earlier:

- Petal Length > 3.1: Goes to the right node
- Petal Width > 1.1: Goes to the right node

⇒ Virginica

Decision Tree 2: Based on Decision Tree 2, we follow the splits as described earlier:

- Petal Width > 1.5: Goes to the right node

⇒ Virginica

Majority Voting: Both Decision Tree 1 and Decision Tree 2 classify the test data point as "Virginica" in their respective trees. Since both trees have made the same classification, the majority voting result is "Virginica".

Step 4: Conclusion.

Based on the majority voting result, the Random Forest ensemble classifies the test flower as of type "Virginica". This is the final prediction based on the combination of both decision trees' classifications.

This numerical problem illustrates Bagging (Bootstrap Aggregating), a powerful ensemble learning technique in machine learning. It aims to improve the predictive performance and reduce overfitting of models by generating multiple subsets of the training data through bootstrapping (random sampling with replacement). Each subset is used to train a base model, such as a decision tree or any other learner. Bagging then combines the predictions from these base models, often through majority voting for classification problems or averaging for regression tasks. This ensemble approach helps to smooth out the variance in individual models and enhance overall model robustness, making it a fundamental technique for building more accurate and stable machine learning models.

Please note that in a real-world application, there might be more complex decision trees with multiple splits and features considered at each step. Also, the Random Forest algorithm would use a larger ensemble of trees to make more accurate predictions. The example provided here is a simplified illustration to help you understand the steps involved in building a decision tree within a Random Forest.