CSE4020

Machine Learning
WINTER SEMESTER 2018

INSTRUCTOR: Prof. Premalatha

STUDENT: Aman Saha [15BCE1273]

Lab Assignment - II

A Study of Decision Tree Algorithm on Statlog (heart) dataset

Abstract

Decision Tree is one of the most widely used predictive models in machine learning and data mining. It is given the name because the resulting output is in a form of a tree structure. It is a rooted tree expressed as a recursive partition of the input space. Each node of the tree partitions the input space into subspaces based on some discrete attribute values. The leaf nodes of the tree are assigned to one class representing the most appropriate target value. In this experiment, we try to implement the Decision Tree algorithm, which uses the Decision Tree as it's model, on the Statlog (heart) dataset and observe the accuracy of the classifier on the test data.

Introduction

The Decision Trees can be used for the classification problems as well as the regression problems. The Decision Tree classifier classifies the input by navigating from the root of the tree to a leaf node following the test criteria of the intermediate nodes finally arriving at a particular class label. The Tree could be constructed using either Gini index or the Information Gain as the criterion. We try to use the Decision Tree classifier constructed using the Gini index criterion on our Statlog(heart) dataset.

Methodology

The first step in the construction of a Decision Tree is to decide which of the attributes of the given dataset is to be placed at the root and which of them are to be placed at the intermediate nodes. Gini index would be used as the criterion for the Tree construction. It is a metric which measures how often a random instance would be misidentified. That means an attribute with lower value of the Gini index would be preferred.

For Gini index calculations, we have to convert the continuous data into categorical form. For example, consider a dataset having 4 attributes A1, A2, A3, A4 where attribute A4 consists of class labels and the other attributes consists of continuous data. We categorize the attributes with some random values A1 (>=X), A2 (>=Y), A3 (>=Z) with X, Y, Z being some random real values.

The Gini index of an attribute is given by –

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

For example, A1 has value >=X for 10 records out of 20 records and 10 have value <X.

1. A1>=X
    1. class positive: 4/10
    2. class negative: 6/10

$$\text{Gini}(4,6) = 1 - ((4/10)^2 + (6/10)^2) = 0.48$$

2. A1<X

    1. class positive: 6/10
    2. class negative: 4/10

$$\text{Gini}(6,4) = 1 - ((6/10)^2 + (4/10)^2) = 0.48$$

Gini index for A1:

$$\text{Gini}(A1) = (10/20) * (\text{Gini}(4,6)) + (10/20) * (\text{Gini}(6,4)) = 0.48$$

Similarly, the Gini values for other attributes could be found and the attribute with the lower Gini value should be placed at the root node and other attributes at the internal node through the same way to form the Decision Tree.

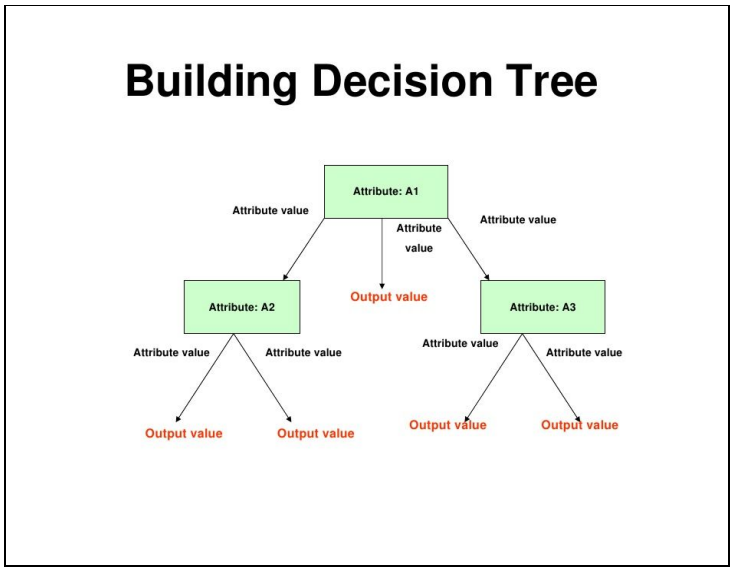

Figure 1. Decision Tree.

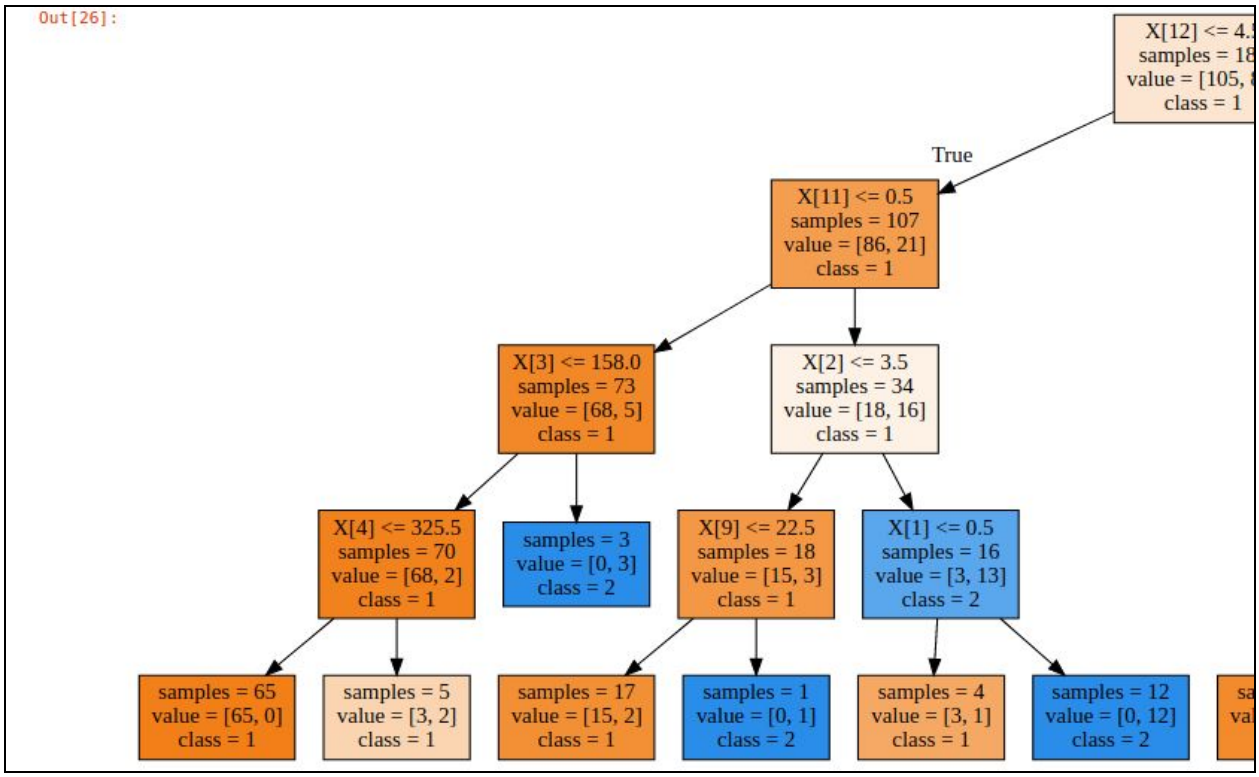

Figure 2. Left Subtree of the Decision Tree.

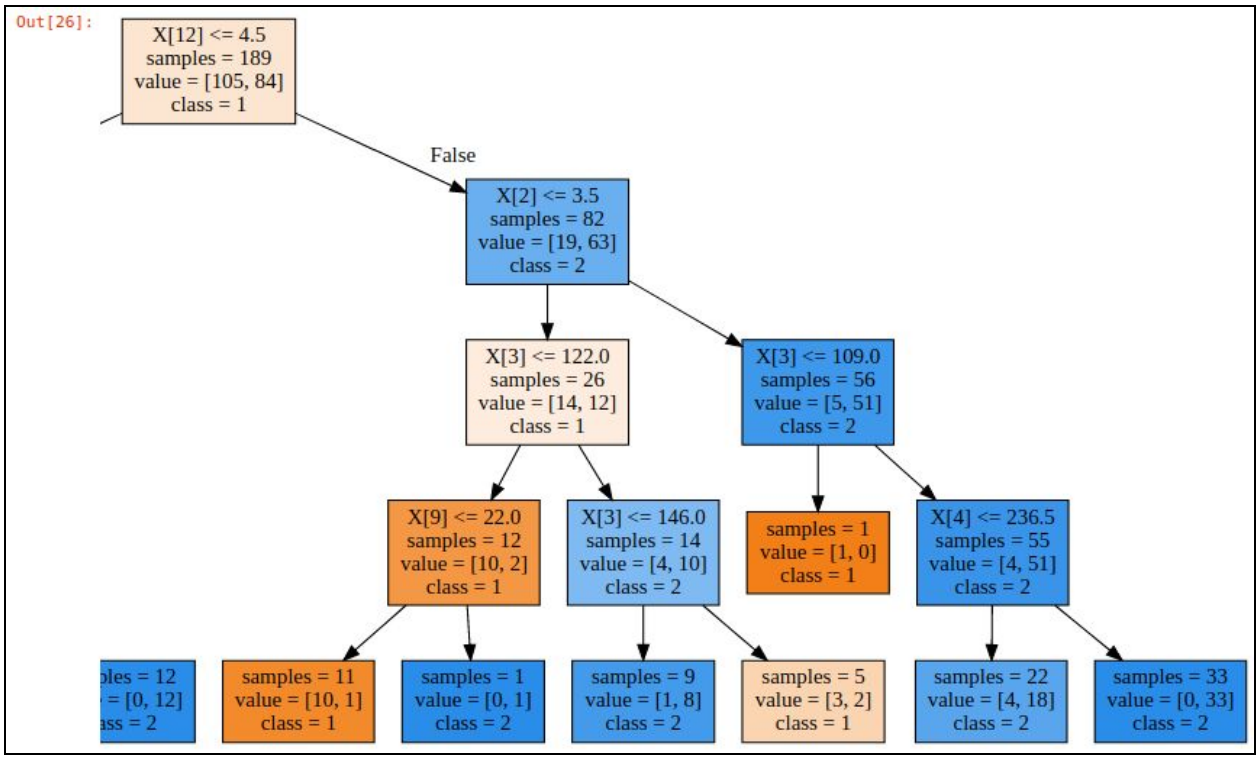

Figure 3. Right Subtree of the Decision Tree.

Dataset

We work on the Statlog (Heart) dataset in this experiment. The task is to classify the (1) absence or (2) presence of a heart disease.

The dataset consists of the attributes Age ([29,77] years), Sex[0-F,1-M], ChestPainType(4types), RestBloodPressure(rest bp), SerumCholestoral (inmg/dl), FastingBlood- Sugar (120>mg/dl) , ResElectrocardiographic (values 0,1,2), MaxHeartRate, Exercise- Induced, Oldpeak(depression induced by exercise), Slope (peak exercise ST segment), MajorVessels([0-3] colored by fluoroscopy), Thal (3 = normal; 6 = fixed defect; 7 = reversible defect), Class([1,2]). The Class represents the absence (1) or presence (2) of a heart disease in a person.

Experiment

Firstly, we split the dataset into training data and test data Xtrain, ytrain, Xtest, ytest where Xtrain, Xtest are the predictor attribute values for the training data and test data respectively and ytrain, ytest are the target attribute values for the training data and test data respectively.

The Decision tree in the Figure 2. and Figure 3. is constructed for the given training data using the Gini index as the criterion. The accuracy of the system is then measured based on the training data as well as the test data as shown in Figure 5.

```
Accuracy on training set: 0.931
Accuracy on test set: 0.667
```

Figure 5. Accuracy measurements.

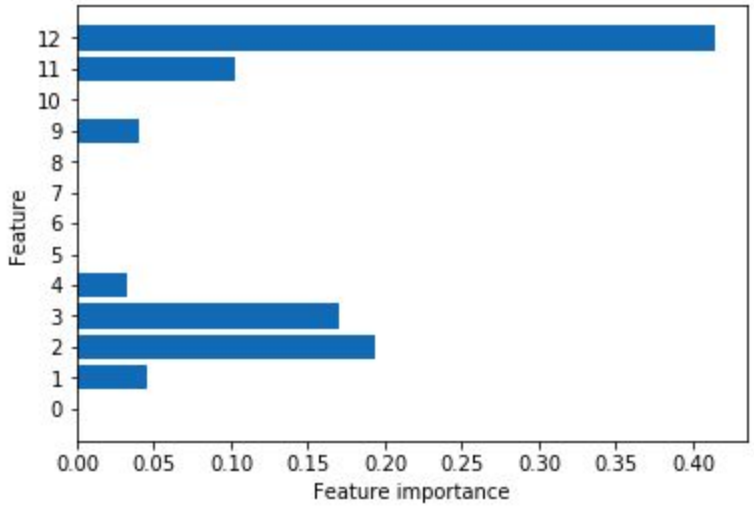

Figure 6. Feature Importance.

The Figure 6. is a plot between the features / attributes and their corresponding importances which clearly shows that the *Thal* attribute gives a major contribution in the determination of the presence or absence of the heart disease and hence is placed on the root of the Decision Tree.

Conclusion

The system built using the Decision Tree displayed 93.1% accuracy on the training data and 66.7 % accuracy on the test dataset. The Tree also highlighted that *Thal* had the most contribution in the determination of the presence or the absence of the heart disease.

References

1. Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004.
2. http://sci2s.ugr.es/keel/dataset.php?cod=99#sub1