Decision Trees

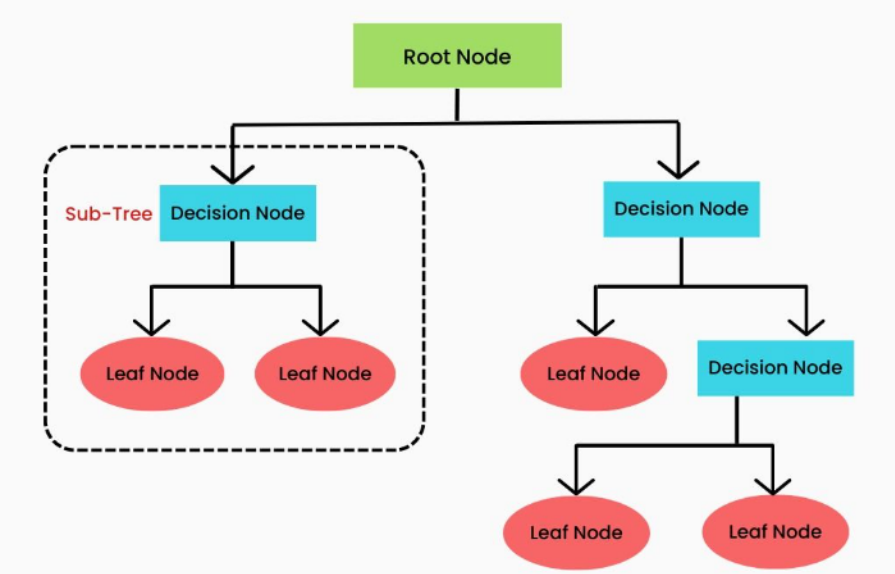
Epoch IIT Hyderabad

Arin Aggarwal
MA22BTECH11006

Contents

1 Introduction

A decision tree is a predictive model that uses a flowchart-like structure to make decisions based on input data. It divides data into branches and assigns outcomes to leaf nodes. Decision trees are used for classification tasks. The tree consists of a root node, decision/internal nodes and leaf nodes.



2 Working

At every level, we need to decide which attribute to use for that node. This can be done using many ways, like Information Gain (using Entropy) and Gini impurity to name a few. Our aim is to reduce the 'impurity' at every step, i.e. to increase the homogeneity/reduce the difference in number of possible outcomes at a node.

When we achieve purity, there is a chance that we overfitted the data. A process called Pruning is done to delete the unnecessary nodes/stop the tree at an appropriate time in order to get the optimal decision tree.

There are 2 kinds of Prunings:

(i **Pre-Pruning**: It involves tuning the parameters prior to the training. This includes methods such as setting max depth or setting minimum number of observations in a leaf in order to node out. Here we need a value that will not overfit as well as underfit our data and for this, we can use GridSearch Cross Validation. Cross-validation is a technique for validating the model efficiency by training it on the subset of input data and testing on previously unseen subset of the input data.

(ii **Post-Pruning**: This is done after the tree has grown to its full depth. This is done using Cost Complexity Pruning. For this, we need to to compute ccp_alpha values and train our decision tree with these different values and plot testing and training scores against gini index for them. The ccp_alpha value that gives highest testing accuracy is the best choice as it generalizes the model better.

3 Mathematical Formulations

3.1 Gini Indices

$$Gini\ Index = \sum_i 1 - P_i^2$$

Weighted Average of Gini Indices is taken to determine effective index of the node.

The lower its value, the better

3.2 Information Gain and Entropy

$$Information\ Gain = E(X) - E(X|Y)$$

where X and Y are features

and

$$Entropy(E(X)) = -\sum_i p_i log_2(i)$$

and

$$E(X|Y) = Weighted\ Average = p(X)E(X) + p(Y)E(Y)$$

The higher its value, the better.