Decision Trees

Detailed Notes

Prerequisite-

- Probability and statistics.

Objectives-

- Understand the prerequisite term such as Gini index, entropy, Information gain and pruning.
- Understand the Decision tree as a CART algorithm.
- Tuning parameter of the decision tree.
- Advantages and disadvantages of Decision Tree.

Decision Tree

A Decision Tree is one of the most popular and effective supervised learning technique for classification problem that equally works well with both categorical and quantitative variables. It is a graphical representation of all the possible solution to a decision that is based on a certain condition. In this algorithm, the training sample points are split into two or more sets based on the split condition over input variables. A simple example of a decision tree can be as a person has to take a decision for going to sleep or restaurant based on parameters like he is hungry or has 25$ in his pocket.

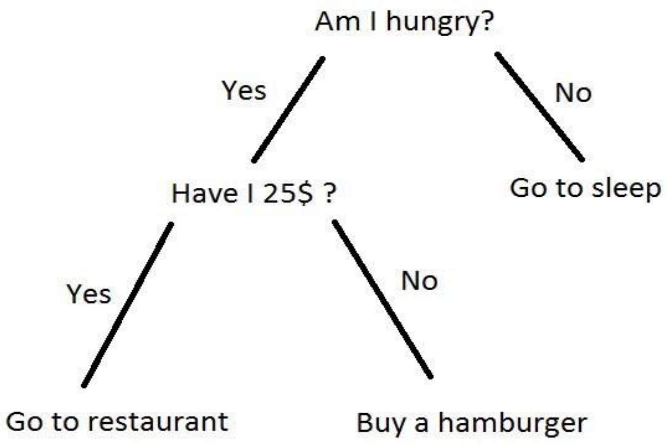


Figure 1 Example of Decision tree

Types of Decision tree

- **Categorical variable decision tree**: The type of decision tree is classified based on the response/target variable. A tree with qualitative or categorical response variable is known as Categorical variable decision tree.
- **Continuous variable decision tree:** A tree with a continuous response variable is known as a continuous variable decision tree.

Terminology

Before moving forward into the details of the decision tree and its working lets understand the meaning of some terminology associated with it.

Root node- Represent the entire set of the population which gets further divided into sets based on splitting decisions.

Decision node- These are the internal nodes of the tree, These nodes are expressed through conditional expression for input attributes.

Leaf node- Nodes which do not split further are known as leaf nodes or terminal nodes.

Splitting- The process of dividing a node into one or more sub-nodes.

Pruning- It is the reverse process of splitting where the sub-nodes are removed.

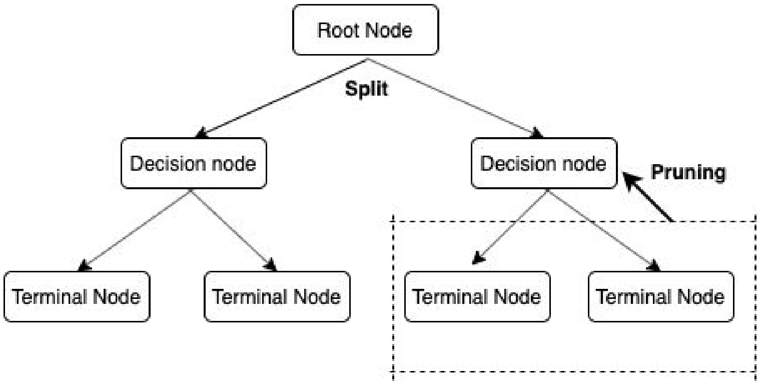


Figure 2 Decision Tree Terminology

The tree accuracy is heavily affected by the split point at a decision node. Decision trees use different criteria to decide split on decision node to get two or more sub-nodes. The resultant sub nodes must increase in the homogeneity of data points also known as the purity of nodes with respect to the target variable. The split decision is tested on all available variables and then the split with maximum purity sub-nodes is get selected.

Measures of Impurity:

Decision trees recursively split feature about to their target variable’s purity. The algorithm is designed to optimize each split such the purity will be maximized. Impurity can be measured in many ways such as Gini impurity, Entropy and information gain.

Gini Impurity-

Gini index is the measure of how often a randomly chosen element from the set would be incorrectly labelled. Mathematically the **impurity** of a set can be expressed as:

$$gini\ impurity = 1 - \sum_i p_i^2$$

Where the p_i represents the probability of random selection of class i observation.

Let consider a simple example of a bag contains some balls (4 red balls and 0 blue balls). The Gini index would be:

$$gin\ impurity = 1 - \left\{P\ (Red)^2 + \ P\ (Blue)^2\right\} = 1 - \left\{\ 1^2 + 0^2\right\} = 0$$

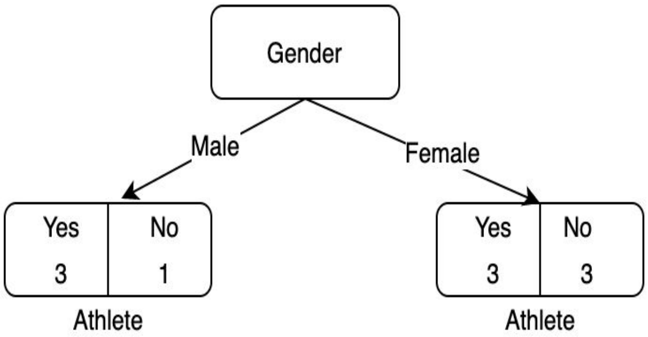
The impurity measurement is zero because no matter what ball we take out we would never incorrectly label it. Similarly, if we consider 2 red and 2 blue balls, the Gini index would be **0.5.** Gini primarily works with categorical variables and perform the binary split. Lower the Gini index shows higher homogeneity. Split selection can be done in two simple steps.

1. Calculate the Gini impurity using the above formula for each sub-nodes.
2. Calculate the Gini for the split using the weighted approach of Gini scores of each node of the split.

Example- Let we have been given with a data of some students with a target variable that whether a student is athlete or not. The input attributes are gender (Male/Female) and education (UG/PG).

Person	Gender	Education	Athlete
Student1	Male	UG	Yes
Student2	Female	PG	No
Student3	Male	UG	Yes
Student4	Male	PG	No
Student5	Female	UG	Yes
Student6	Female	UG	Yes
Student7	Female	PG	No
Student8	Male	UG	Yes
Student9	Female	PG	Yes
Student10	Female	PG	No

Case 1- Let first make a split with respect to **gender** variable: Figure is showing the split-



Step 1- calculate the Gini impurity for each sub-node

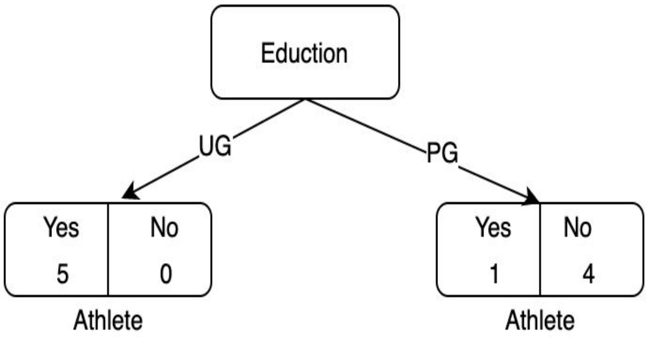
$$gini_{male} = 1 - \left\{P\ (yes)^2 + P\ (No)^2\right\} = 1 - \left\{\ \left(\tfrac{3}{4}\right)^2 + \left(\tfrac{1}{4}\right)^2\right\} = 0.375$$

$$gini_{female} = 1 - \left\{P\ (yes)^2 + P\ (No)^2\right\} = 1 - \left\{\ \left(\tfrac{3}{6}\right)^2 + \left(\tfrac{3}{6}\right)^2\right\} = 0.5$$

Step 2- Calculate the Gini for split using weights of each Gini score.

$$gini_{gender} = \tfrac{4}{10}\,(gini_{male}) + \tfrac{6}{10}\left(gini_{female}\right) = 0.4\ \ (0.375) + 0.6\,(0.5) = 0.45$$

Case 2- Let first make the second split with respect to the **education** variable: Figure is showing the split-



Step 1- calculate the Gini impurity for each sub-node

$$gini_{UG} = 1 - \left\{P\ (yes)^2 + P\ (No)^2\right\} = 1 - \left\{\ \left(\tfrac{5}{5}\right)^2 + \left(\tfrac{0}{5}\right)^2\right\} = 0.0$$

$$gini_{PG} = 1 - \left\{P\ (yes)^2 + P\ (No)^2\right\} = 1 - \left\{\ \left(\tfrac{1}{5}\right)^2 + \left(\tfrac{4}{5}\right)^2\right\} = 0.32$$

Step 2- Calculate the Gini for split using weights of each Gini score.

$$gini_{education} = \tfrac{5}{10}\,(gini_{UG}) + \tfrac{3}{10}\,(gini_{PG}) = 0.5\ \ (0.0) + 0.3\,(0.32) = 0.096$$

As it is clear from the calculations that the Gini score for education is lower than for gender variable so the best variable for making a split is **education.** (gives more pure sub-nodes)

Entropy and Information gain

Entropy- In Layman term, Entropy is nothing but the measure of disorder. We can also think it as a measure of purity. The mathematical formula of Entropy is as follows-

$$E = \sum_{i=1}^{c} -\ p_i log_2 p_i$$

Where p_i is the probability of class i.

For the given example we have two labels for athlete class (Yes/No). Therefore the athlete could be either Yes or No. So, the entropy for the given set is defined as:

$$P\ (Yes) = \tfrac{6}{10} \quad and\ P\ (No) = \tfrac{4}{10}$$

$$Entropy = \ -\tfrac{6}{10}\left(\tfrac{6}{10}\right) - \tfrac{4}{10}\left(\tfrac{4}{10}\right) \quad = 0.442 + 0.529 = 0.971$$

The entropy comes out to be 0.971 which is considered a high entropy, Which means a low level of Purity. Entropy is measured between 0 and 1.

Information Gain

Information Gain measures the reduction in Entropy and decides which attribute would be selected as a decision node. In general, information gain has again calculated the subtraction of decision node entropy to the weighted average of the entropies for the children of the decision node. That is for ”m” points in the first child node and n points in the second child node, the information gain is:

$$IG = Entropy\,(Decisionnode) - \frac{m}{m+n}Entropy\,(FirstChild) - \frac{n}{m+n}Entropy(SecondChild)$$

A Simple Graph between a class and Entropy would be like:

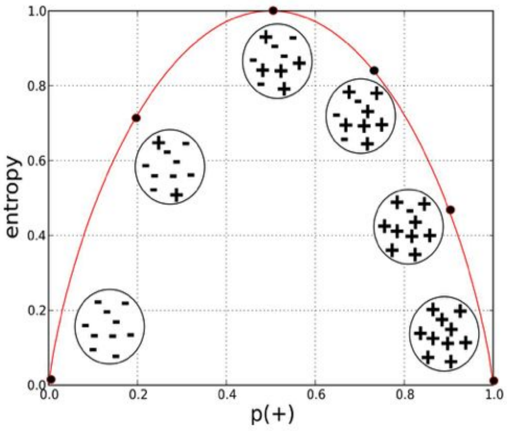


Figure 3 Entropy graph w.r.t impurity

Gini vs Entropy-

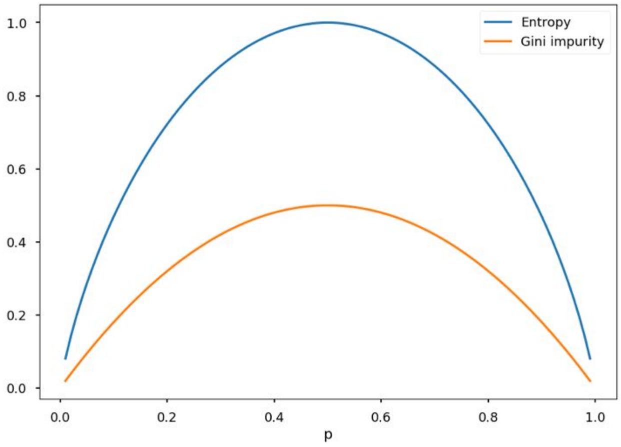


Figure 4 Gini vs Entropy Graph

Pruning-

One of the problems with the decision tree is it gets easily overfit with training sample and becomes too large and complex. A complex and large tree poorly generalizes the new samples data whereas a small tree fails to capture the information of training sample data.

Pruning may be defined as shortening the branches of the tree. The process of reducing the size of the tree by turning some branch node into a leaf node and removing the leaf node under the original branch.

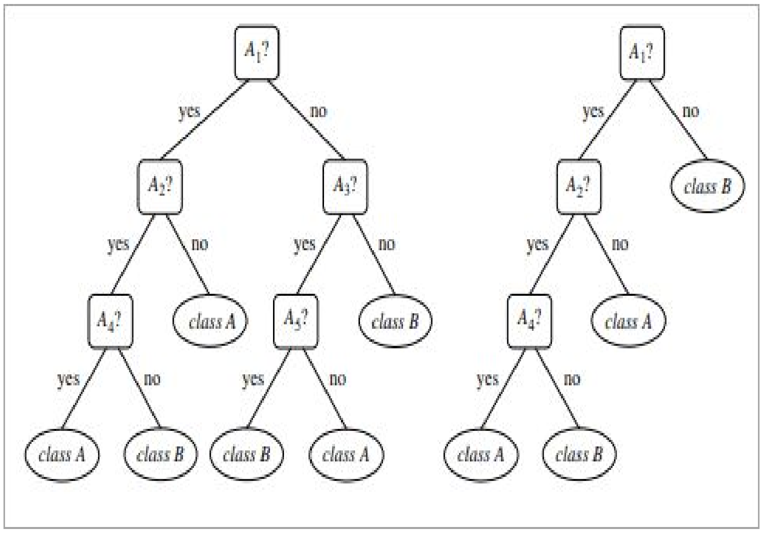


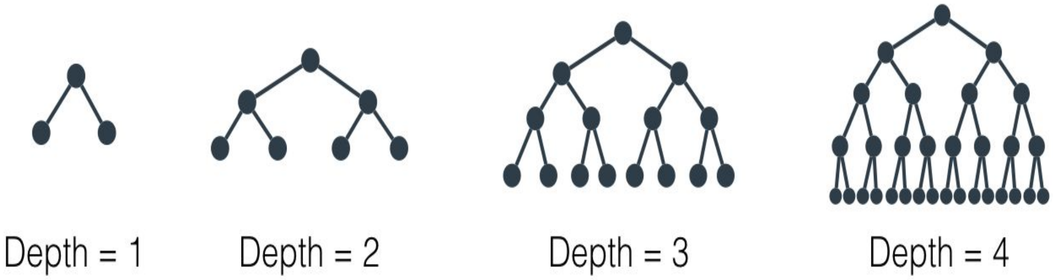
Figure 5 Pruning in Decision Tree

Pruning is very useful in the decision tree because sometimes what happens is that the decision tree may fit the training data very well but performs very poorly in testing or new data. So, by removing branches we can reduce the complexity of tree which help in reducing the overfitting of the tree.

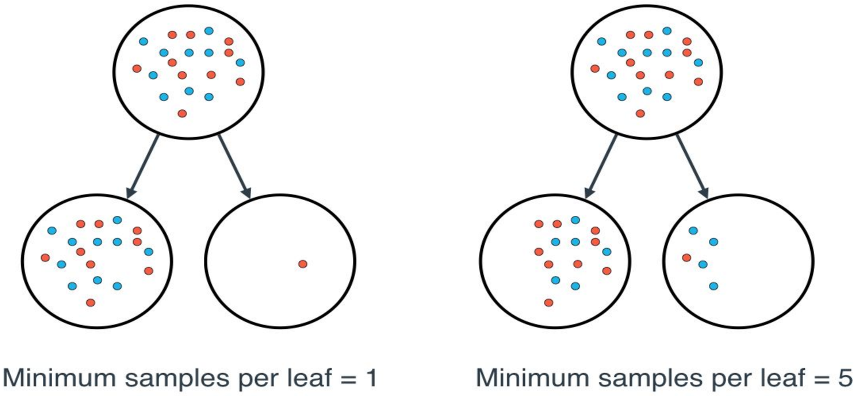
Hyperparameters for Decision Trees

To generate decision trees that will generalize to new problems well, we can tune different aspect of trees. We call these aspects of decision tree “hyperparameters”. Some of the Important Hyperparameters used in decision trees are as follows:

Maximum Depth- The maximum depth of the decision tree is simply the largest length between the root to leaf. A tree of maximum length **k** can have at most **2**k** leaves.



The minimum number of samples per leaf- While splitting a node, one could run into the problem of having 990 samples in one of them, and 10 on the other. This will not take us too far in our process and would be a waste of resources and time. If we want to avoid this, we can set a minimum for the number of samples we allow on each leaf.



The maximum number of the feature- We can have too many features to build a decision tree. While splitting, in every split, we have to check the entire data-set on each of the features. This can be very expensive. A solution for this is to limit the number of features that one looks for in each split. If this number is large enough, we're very likely to find a good feature among the ones we look for (although maybe not the perfect one). However, if it's not as large as the number of features, it will speed up our calculations significantly.

Example-

The decision tree for the example discussed in Gini calculation will be-

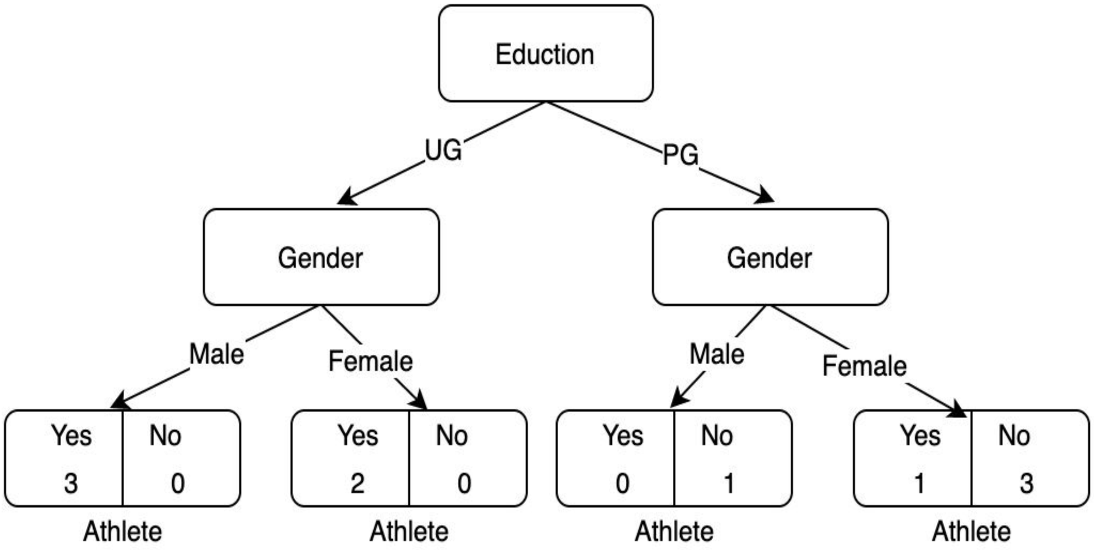


Figure 6 Decision Tree

Advantages of Decision tree

- Decision tree does not require normalization and scaling of data.

- Missing value in the data-set also does not affect the process of making the decision tree of the particular dataset.
- It is very easy to explain the decision tree to anyone. It does not require much knowledge of statistics and visualization also helps very much.
- A decision tree requires less time and effort during data pre-processing than other algorithms.

Disadvantages of Decision Tree

- A small change in the data-set can result in a large change in the structure of the decision tree causing instability in the model.
- It requires more time to train a model in decision tree than any other algorithm presents out there.
- In a Decision tree calculation can be far more expensive than the other algorithm.
- It is not advised to apply decision tree for regression or predicting continuous values.

****