Logisitic Regression Report

Donal Loitam

July 16, 2022

Contents

1 Introduction

- One of the easiest algorithm in machine learning as it does not involve much mathematics.
- A Decision Tree is a supervised machine learning algorithm that can be used for both Regression and Classification problem statements
- It uses a flowchart like a tree structure wherein internal nodes represent conditions and we check at every nodes if that condition is satisfied and split accordingly until it reaches a leaf(Decision). As for instance, see the tree below

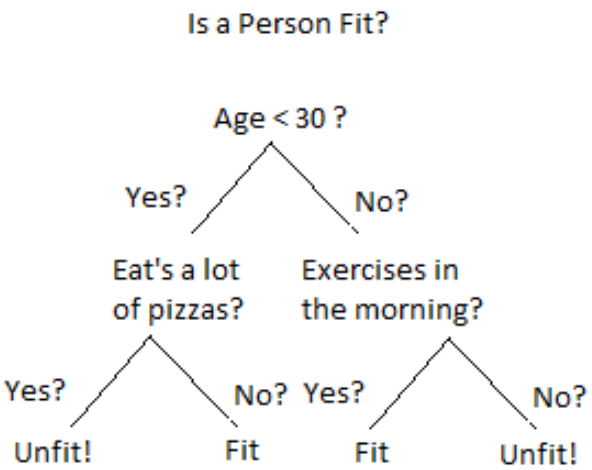

Figure 1: Example of a Decision tree: Is a person fit ?

2 Key Points of the Algorithm

So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop ?
What we generally use for decision trees is the Recursive Binary splitting

Recursive Binary Split : All the features are considered and different split points are tried using a cost function(we will discuss one later). The split with the least cost is selected. This method is recursive as we can recursively apply this to the remaining groups and continue splitting.Due to which, it is also called the **greedy algorithm** as we are choosing the least cost at every step.
Note that in our example in fig 1 the split that costed least was the **Age**

Cost function : If the dataset consists of N attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. So we have a bunch of criterias/methods that can be used as: **Entropy, Information Gain, Gini Index**.We use Gini Index
Gini Index : Gini impurity is a function that determines how well a decision tree was split. Basically, it helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5.

- Gini Index works with the categorical target variable "Success" or "Failure". It performs only Binary splits.
- Lower the Gini ,the better is that feature

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2 \qquad \in [0, 0.5] \tag{1}$$

where, p_i is the probability of an object being classified to a particular class.
Steps to Calculate Gini index for a split
(1) Calculate Gini for sub-nodes, using the above formula for success(p) and failure(q) (p^2+q^2).
(2) Calculate the Gini index for split using the weighted Gini score of each node of that split.
Calculating the Gini Index for Past Trend
Illustration:

Past Trend	Open Interest	Trading Volume	Return
+	Low	High	Up
-	High	Low	Down
+	Low	High	Up
+	high	High	Down
-	Low	High	Down
+	Low	Low	Down
-	High	High	Down
-	Low	High	Down
+	Low	Low	Down
+	High	High	Up

P(Past Trend=Positive): 6/10
P(Past Trend=Negative): 4/10

- If (Past Trend = Positive and Return = Up), probability = 4/6
- If (Past Trend = Positive and Return = Down), probability = 2/6

Gini index = $1 - ((4/6)^2 + (2/6)^2) = 0.45$

- If (Past Trend = Negative and Return = Up), probability = 0
- If (Past Trend = Negative and Return = Down), probability = 4/4

Gini index = $1 - ((0)^2 + (4/4)^2) = 0$

Weighted sum of the Gini Indices can be calculated as follows:
Gini Index for Past Trend = (6/10)0.45 + (4/10)0 = 0.27

Similarly , we calculate for **Open Interest** = 0.47 and for **Trading Volume** = 0.34. Since gini for past trend is the least we take it to br the parent node.

3 How to avoid/counter Overfitting in Decision Trees?

You must be asking this question to yourself that when do we stop growing our tree? Usually, real-world datasets have a large number of features, Such trees take time to build and can lead to overfitting.If there is no limit set on a decision tree, it will give you 100 % accuracy on the training data set because in the worse case it will end up making 1 leaf for each observation One way to remove overfitting is :

Pruning Decision Trees : It helps in improving the performance of the tree on unseen **test datasets** by cutting the nodes or sub-nodes which are not significant
Post pruning - (Minimum error). The tree is pruned back to the point where the cross-validated error is a minimum. Cross-validation is the process of building a tree with most of the data and then using the remaining part of the data to test the accuracy of the decision tree.
Pre pruning - At each stage of splitting the tree, we check the cross-validation error. If the error does not decrease significantly enough then we stop.

4 Some Questions

1. Explain the CART Algorithm for Decision Trees.
Ans. The CART stands for Classification and Regression Trees is a greedy algorithm that searches for a best split at the top level, then repeats the same process at each of the remaining levels. The solution provided by the greedy algorithm is not guaranteed to be optimal, it often produces a solution that's reasonably good since finding the optimal Tree is an NP-Complete problem that requires exponential time complexity.
2. Briefly explain the properties of Gini Impurity.
Ans. Let X (discrete random variable) takes values y_+ and y_- (two classes). Now, consider :

- **Case 1** When 100% observations belong to y_+ / y_-

$$Gini = 1 - (1^2 + 0^2) = 0 \tag{2}$$

- **Case 2** When 50% observations belong to y_+

$$Gini = 1 - ((0.5)^2 + (0.5)^2) = 0.5 \tag{3}$$

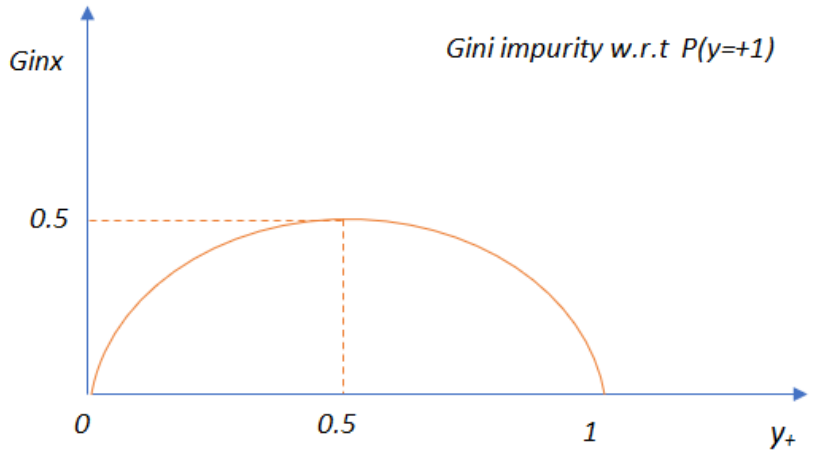

Figure 2: Gini

3. Do we require Feature Scaling for Decision Trees? Explain.
Ans Decision tree based models take decisions on the basis of parameters.

Say you have a variable "Age"(take multiple values). Now in a decision tree model, a tree is formed on the basis of micro decisions that the algorithm makes to form a tree. And example of this micro decision could be,

If $20 > Age > 30$. (either Yes/No)
Now let those feature values were to be scaled down let's say by 100

$0.2 > Age/100 > 0.3$ even then the decision wouldn't have changed. Keep in mind the value of Age too has been scaled down by 100.

Hence, decision tree models don't require scaling of feature values.
4. List down the problem domains in which Decision Trees are most suitable.
Ans.

- Decision Trees are most suitable for tabular data.
- The outputs are discrete.

5. List down some popular algorithms used for deriving Decision Trees along with their attribute selection measures.
Ans ID3 (Iterative Dichotomiser): Uses Information Gain as attribute selection measure.
CART (Classification and Regression Trees) – Uses Gini Index as attribute selection measure.