Analytics Vidhya

GSplitting Decision Trees with Gini Impurity

BEGINNER CLASSIFICATION MACHINE LEARNING REGRESSION VIDEOS

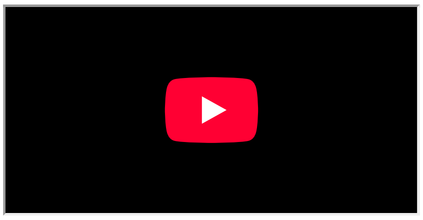
Introduction

In decision trees, making informed choices is pivotal for accurate and robust predictions. Selecting the optimal split to branch nodes significantly influences a decision tree's effectiveness. One of the powerful methods employed for this purpose is the gini impurity decision tree. This article delves into the intricacies of utilizing Gini Impurity to discern the best split in decision trees. We will explore the concepts, calculations, and real-world implications, equipping you with a comprehensive understanding of how it enhances the precision and reliability of decision tree models. Whether you're a novice or a seasoned data practitioner, uncovering the secrets behind this essential algorithm will empower you to harness the full potential of decision trees in your data analysis endeavors.

Table of contents

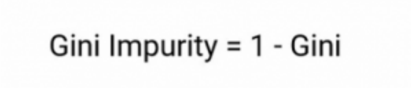
What is Gini Impurity?

Gini impurity is a measure used in decision tree algorithms to quantify a dataset's impurity level or disorder. In binary classification problems, it assesses the likelihood of an incorrect classification when a randomly selected data point is assigned a class label based on the distribution of classes in a particular node. It ranges from 0 to 0.5, where 0 indicates a perfectly pure node (all instances belong to the same class), and 0.5 signifies maximum impurity (an equal distribution of classes). In decision trees, it aids in selecting the optimal split by identifying features that result in more homogeneous subsets of data, ultimately contributing to the creation of accurate and reliable predictive models.



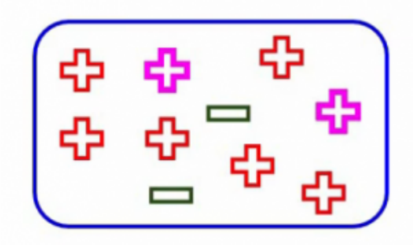
Decision Tree Algorithm for Selecting the Best Split

There are multiple algorithms that are used by the decision tree to decide the best split for the problem. Let's first look at the most common and popular out of all of them, which is **Gini Impurity**. It measures the impurity of the nodes and is calculated as:

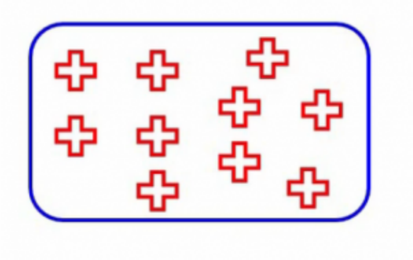


Gini Impurity = 1 - Gini

Let's first understand what Gini is and then I'll show you how you can calculate the Gini impurity for split and decide the right split. Let's say we have a node like this-



So what Gini says is that if we pick two points from a population at random, the pink ones highlighted here, then they must be from the same class. Let's say we have a completely pure node-



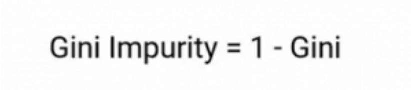
Can you guess what would be the probability that a randomly picked point will belong to the same class? Well, obviously it will be 1 since all the points here belong to the same class. So no matter which two points you picked, they will always belong to that one class and hence the probability will always be 1 if the node is pure. And that is what we want to achieve using Gini.

Gini ranges from zero to one, as it is a probability and the higher this value, the more will be the purity of the nodes. And of course, a lesser value means lesser pure nodes.

Properties of Gini Impurity

Let's look at its properties before we actually calculate the gini impurity decision tree to decide the best split.

We decide the best split based on the gini impurity in decision tree and as we discussed before, Gini impurity is:



Gini Impurity = 1 - Gini

Here Gini denotes the purity and hence Gini impurity tells us about the impurity of nodes. Lower the Gini impurity we can safely infer the purity will be more and hence a higher chance of the homogeneity of the nodes.

> Gini works only in those scenarios where we have **categorical** targets. It does not work with continuous targets.

A very important point to note to keep in mind. For example, if you want to predict the house price or the number of bikes that have been rented, Gini is not the right algorithm. It only performs binary splits either yes or no, success or failure, and so on. So it will only split a node into two sub-nodes. These are the properties of Gini impurity.

Also Read: How to Split a Decision Tree – The Pursuit to Achieve Pure Nodes

Steps to Calculate Gini Impurity for a Split

Let's now look at the steps to calculate the Gini split.

Step 1: Calculate GI for Sub-nodes

First, we calculate the Gini impurity for sub-nodes, as you've already discussed gini impurity in decision tree is, and I'm sure you know this by now:

Gini impurity = 1 – Gini

Here is the sum of squares of success probabilities of each class and is given as:

Considering that there are n classes.

Step 2: Calculate GI of the Split

Once we've calculated the Gini impurity for sub-nodes, we calculate the gini impurity decision tree of the split using the weighted impurity of both sub-nodes of that split. Here the weight is decided by the number of observations of samples in both the nodes. Let's look at these calculations using an example, which will help you understand this even better.

For the split on the performance in the class, remember this is how the split was?

We have two categories, one is “above average” and the other one was “below average”.

Above Average

When we focus on the above average, we have 14 students out of which 8 play cricket and 6 do not. The probability of playing cricket would be 8 divided by 14, which is around **0.57**, and similarly, for not playing cricket, the probability will be 6 divided by 14, which will be around **0.43**. Here for simplicity, I’ve rounded up the calculations rather than taking the exact number.

Below Average

Similarly, when we look at below average, we calculated all the numbers and here they are- the probability of playing is **0.33** and of not playing is **0.67-**

Let's now calculate the GI of the sub-nodes for above average and here's the calculation-

It will be, one minus the square of the probability of success for each category, which is 0.57 for playing cricket and 0.43 for not playing cricket. So after this calculation Gini comes out to be around **0.49**. The Below average node will do the same calculation as Gini. For below average:

It comes up to be around 0.44. Just take a pause and analyze these numbers.

Now to calculate the gini impurity in decision tree of the split, we will take the weighted Gini impurities of both nodes, above average and below average. In this case, *the weight of a node is the number of samples in that node divided by the total number of samples in the parent node*. So for the above-average node here, the weight will be **14/20**, as there are 14 students who performed above the average of the total 20 students that we had.

And the weight for below average is **6/20**. *So, the weighted Gini impurity will be the weight of that node multiplied by the Gini impurity of that node*. The weighted gini impurity decision tree for **performance in class split** comes out to be:

Step 3: Calculate GI for Split on Class

Similarly, here we have captured the gini index decision tree for **the split on class**, which comes out to be around **0.32**–

Now, if we compare the two Gini impurities for each split-

We see that the Gini impurity for the split on **Class** is less. And hence class will be the first split of this decision tree.

Similarly, for each split, we will calculate the Gini impurities and the split producing minimum Gini impurity will be selected as the split. And you know, that the minimum value of gini index decision tree means that the node will be purer and more homogeneous.

Conclusion

In conclusion, understanding the Gini Impurity measure is pivotal for enhancing the accuracy and reliability of decision tree models. By quantifying the impurity level of data nodes, Gini Impurity aids in identifying optimal splits, leading to more homogeneous subsets and ultimately more accurate predictions. With a range from 0 to 0.5, where lower values signify purer nodes, Gini Impurity serves as a crucial tool in decision tree algorithms. Mastery of this concept equips data practitioners, whether beginners or experts, to unlock the full potential of decision trees in their data analysis endeavors.

If you are looking to kick start your Data Science Journey and want every topic under one roof, your search stops here. Check out Analytics Vidhya's Certified AI & ML BlackBelt Plus Program!

If you have any questions, let me know in the comments section!

Q1. What is Gini Impurity formula?

A. The Gini Impurity formula is: $1 – (p_1)^2 – (p_2)^2$, where p_1 and p_2 represent the probabilities of the two classes in a binary classification problem.

Q2. What is the Gini Impurity of 0?

A. A Gini Impurity of 0 indicates a perfectly pure node, where all instances belong to the same class.

Q3. What is the Gini coefficient and Gini Impurity?

A. The Gini coefficient measures income inequality, while Gini Impurity assesses impurity in decision tree nodes, helping to determine optimal splits for classifying data.

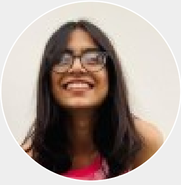
Q4. What is the difference between entropy and Gini impurity?

A. Entropy measures a set's disorder level, while Gini impurity quantifies the probability of misclassifying instances. Both are used in decision trees to determine node splits, but Gini favors larger partitions.

Q5. What is the highest Gini impurity?

The highest Gini impurity is 0.5. This indicates that a node has an equal distribution of classes, meaning that it is completely impure.

Article Url - https://www.analyticsvidhya.com/blog/2021/03/how-to-select-best-split-in-decision-trees-gini-impurity/



Himanshi Singh

I'm a data lover who enjoys finding hidden patterns and turning them into useful insights. As the Manager – Content and Growth at Analytics Vidhya, I help data enthusiasts learn, share, and grow together.

Thanks for stopping by my profile – hope you found something you liked □