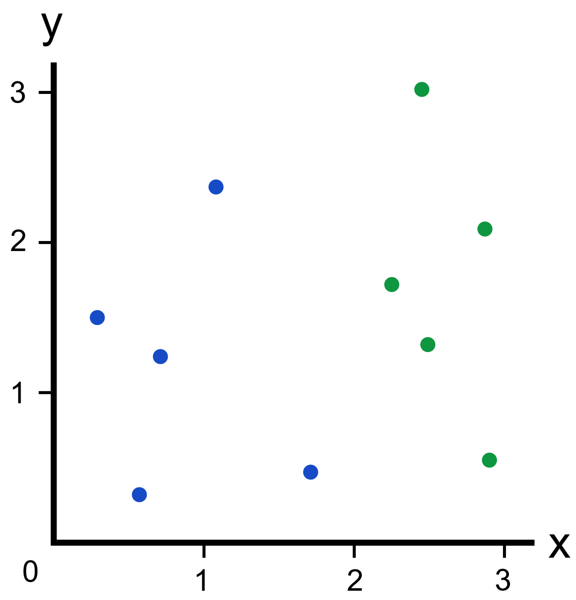
Victor Zhou

A Simple Explanation of Information Gain and Entropy

What Information Gain and Information Entropy are and how they're used to train Decision Trees.

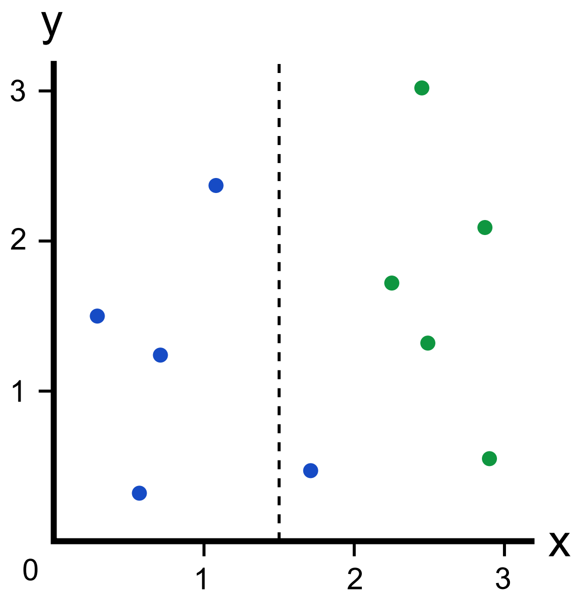
JUNE 7, 2019

Information Gain, like Gini Impurity, is a metric used to train Decision Trees. Specifically, these metrics measure the **quality of a split**. For example, say we have the following data:



The Dataset

What if we made a split at $x = 1.5$?



An Imperfect Split

This imperfect split breaks our dataset into these branches:

- Left branch, with 4 blues. ● ● ● ●
- Right branch, with 1 blue and 5 greens. ● ● ● ● ● ●

It's clear this split isn't optimal, but how good is it? **How can we *quantify* the quality of a split?**

That's where Information Gain comes in.

> *Confused? Not sure what Decision Trees are or how they're trained? Read the beginning of my [introduction to Random Forests and Decision Trees](#).*

Information Entropy

Before we get to Information Gain, we have to first talk about [Information Entropy](#). In the context of training Decision Trees, Entropy can be roughly thought of as **how much variance the data has**. For example:

- A dataset of only blues ●●●● would have very **low** (in fact, zero) entropy.

- A dataset of mixed blues, greens, and reds would have relatively **high** entropy.

Here's how we calculate Information Entropy for a dataset with C classes:

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

where p_i is the probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i).

The easiest way to understand this is with an example. Consider a dataset with 1 blue, 2 greens, and 3 reds: . Then

$$E = -(p_b \log_2 p_b + p_g \log_2 p_g + p_r \log_2 p_r)$$

We know $p_b = \frac{1}{6}$ because $\frac{1}{6}$ of the dataset is blue. Similarly, $p_g = \frac{2}{6}$ (greens) and $p_r = \frac{3}{6}$ (reds). Thus,

$$\begin{aligned} E &= -(\frac{1}{6}\log_2(\frac{1}{6}) + \frac{2}{6}\log_2(\frac{2}{6}) + \frac{3}{6}\log_2(\frac{3}{6})) \\ &= \boxed{1.46} \end{aligned}$$

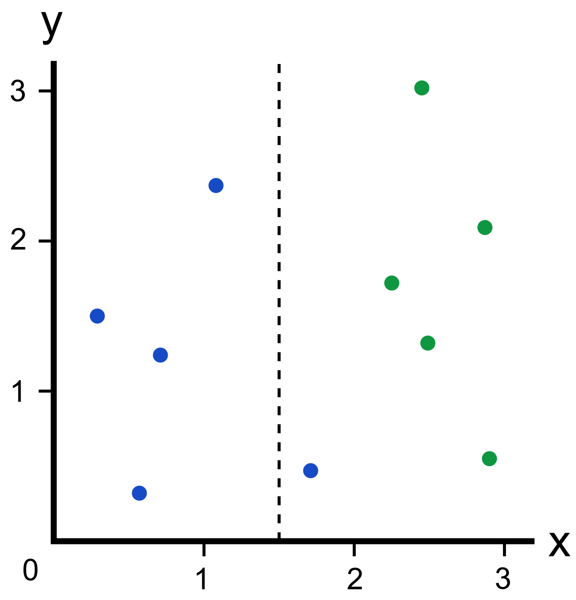
What about a dataset of all one color? Consider 3 blues as an example: . The entropy would be

$$E = -(1 \log_2 1) = \boxed{0}$$

Information Gain

It's finally time to answer the question we posed earlier: **how can we quantify the quality of a split?**

Let's consider this split again:



An Imperfect Split

Before the split, we had 5 blues and 5 greens, so the entropy was

$$\begin{aligned} E_{before} &= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) \\ &= \boxed{1} \end{aligned}$$

After the split, we have two branches.

Left Branch has 4 blues, so $E_{left} = \boxed{0}$ because it's a dataset of all one color.

Right Branch has 1 blue and 5 greens, so

$$\begin{aligned} E_{right} &= -(\frac{1}{6} \log_2 (\frac{1}{6}) + \frac{5}{6} \log_2 (\frac{5}{6})) \\ &= \boxed{0.65} \end{aligned}$$

Now that we have the entropies for both branches, we can determine the quality of the split by **weighting the entropy of each branch by how many elements it has**. Since Left Branch has 4 elements and Right Branch has 6, we weight them by 0.4 and 0.6, respectively:

$$\begin{aligned} E_{split} &= 0.4 * 0 + 0.6 * 0.65 \\ &= \boxed{0.39} \end{aligned}$$

We started with $E_{before} = 1$ entropy before the split and now are down to 0.39! **Information Gain = how much Entropy we removed**, so

$$\text{Gain} = 1 - 0.39 = \boxed{0.61}$$

This makes sense: **higher Information Gain = more Entropy removed**, which is what we want. In the perfect case, each branch would contain only one color after the split, which would be zero entropy!

Recap

Information Entropy can be thought of as how unpredictable a dataset is.

- A set of only one class (say, blue ●●●) is extremely predictable: anything in it is blue. This would have **low** entropy.
- A set of many mixed classes ●●● is unpredictable: a given element could be any color! This would have **high** entropy.

The actual formula for calculating Information Entropy is:

$$E = -\sum_i^C p_i \log_2 p_i$$

Information Gain is calculated for a split by subtracting the weighted entropies of each branch from the original entropy. When training a Decision Tree using these metrics, the best split is chosen by maximizing Information Gain.

Want to learn more? Check out my explanation of [Gini Impurity](), a similar metric, or my in-depth guide [Random Forests for Complete Beginners]().

This blog is open-source on Github.

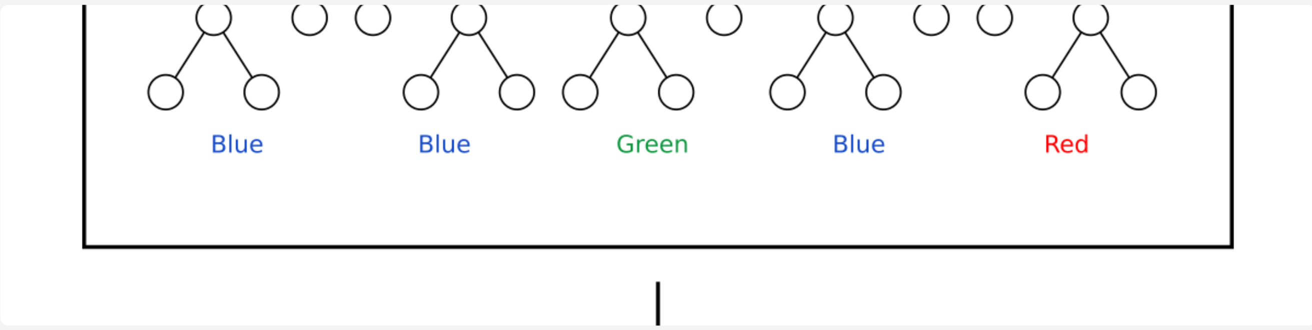
Tags: Machine Learning | For Beginners | Decision Trees | Random Forests

YOU MIGHT ALSO LIKE

A Simple Explanation of Gini Impurity

March 29, 2019

What Gini Impurity is (with examples) and how it's used to train Decision Trees.



Random Forests for Complete Beginners

April 10, 2019

The definitive guide to Random Forests and Decision Trees.

Victor Zhou @victorczhou

Software Engineer. I blog about web development, machine learning, and more topics.

SHARE THIS POST

Facebook | Twitter | LinkedIn | Reddit

Login

Add a comment

M↓ MARKDOWN

COMMENT ANONYMOUSLY

ADD COMMENT

Upvotes Newest Oldest

Anonymous
1 point · 15 months ago

A great explanation! I have a question here: What is the difference between Information gain and Gini impurity? They both seem to help split the data by giving how much Impurity will be removed by one split, so we know which split to use.

Victor Zhou MODERATOR
0 points · 15 months ago

You're right - in most cases they're very similar and serve the same purpose.

Michael Yang
1 point · 14 months ago

Solid explanation, better and clearer than my professor's example!

Anonymous
1 point · 16 months ago

One of the best example I have come across explaining decision trees

Anonymous
0 points · 15 months ago

It's short, simple, but extremely useful. Muchas Gracias

Anonymous
0 points · 3 months ago

Awesome article, it really helped me out a lot! Keep up the great work.

Anonymous
0 points · 14 months ago

Thanks Victor, very clear and easy to understand :)

Anonymous
0 points · 12 months ago

Its very simply explained and very useful information given.

Anonymous

0 points · 2 months ago

very clear! thanks

Anonymous
0 points · 14 months ago

Thanks so much, the course I did made this seem overly complicated, but I get it now.

Anonymous
0 points · 15 months ago

What if I look at the minimum of the Gini impurity instead of the maximum of the information gain? Both of them can be used to find the best split?

Anonymous
0 points · 15 months ago

I love this. Very clear and easy to understand. Thanks for the explanation.

Anonymous
0 points · 12 months ago

Very good explanations. You would make an excellent online-instructor.

Anonymous
0 points · 18 months ago

Crystal clear. Thanks a lot for your work. Best.

Minh Le
0 points · 18 months ago

This is such a great article! Thanks for sharing!

> **Anonymous**
> **0 points** · 11 days ago
>
> hi

Anonymous
0 points · 8 months ago

Great explanation. Thanks!

Delali christinme
0 points · 10 months ago

this is an amazing post. in 5-8 minutes i have understood something which have been bugging me for hours. Xie Xie nin.

Anonymous

0 points · 6 months ago

Oh god you just save me from my quiz!!! Your explanation is super clear and easy to understand.

phanikiran s

0 points · 4 months ago

A very simple..no jargon and heavy math based explanation...loved it...believe me...I have gone through many articles on youtube,medium,etc...nobody has been able to come close to making it this easy to understand... Just a question..is there a max score for entropy and IG?

Eddie Otudor

0 points · 4 months ago

Very well Explained! Thanks so much Sir!