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,

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

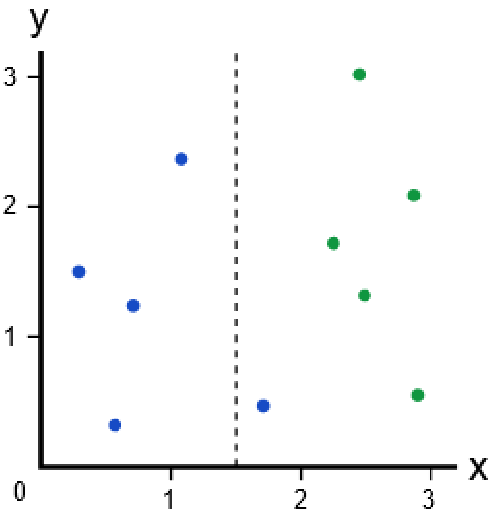
What about a dataset of all one color? Consider 3 blues as an example: 🔵🔵🔵. The entropy would be

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

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.