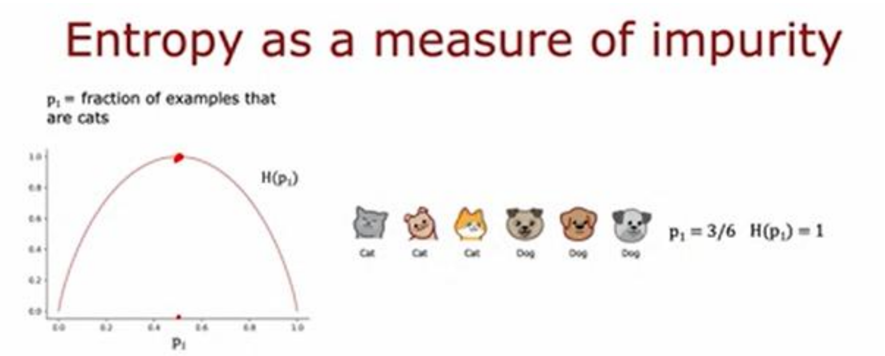
Measurement of purity

In this video, we'll look at the way of measuring the purity of a set of examples. If the examples are all cats of a single class then that's very pure, if it's all not cats that's also very pure, but if it's somewhere in between how do you quantify how pure is the set of examples? Let's take a look at the definition of entropy, which is a measure of the impurity of a set of data.

Given a set of six examples like this, we have three cats and three dogs, let's define p_1 to be the fraction of examples that are cats, that is, the fraction of examples with label one, that's what the subscript one indicates. p_1 in this example is equal to 3/6. We're going to measure the impurity of a set of examples using a function called the entropy which looks like this.

The entropy function is conventionally denoted as capital H of this number p_1 and the function looks like this curve over here where the horizontal axis is p_1, the fraction of cats in the sample, and the vertical axis is the value of the entropy. In this example where p_1 is 3/6 or 0.5, the value of the entropy of p_1 would be equal to one.



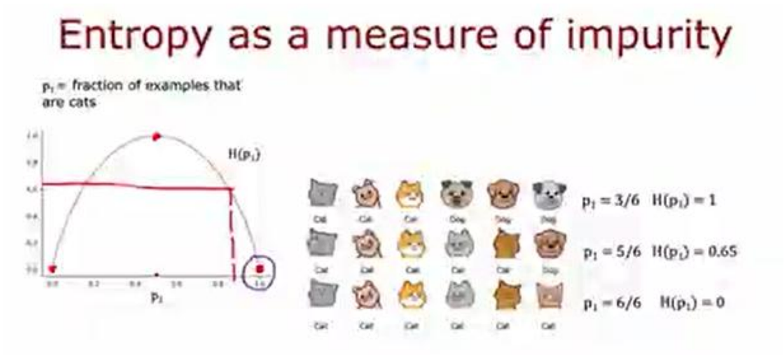
Entropy as a measure of impurity

p_1 = fraction of examples that are cats

$p_1 = 3/6 \quad H(p_1) = 1$

You notice that this curve is highest when your set of examples is 50-50, so it's most impure as an impurity of one or with an entropy of one when your set of examples is 50-50, whereas in contrast if your set of examples was either all cats or not cats then the entropy is zero. Let's just go through a few more examples to gain further intuition about entropy and how it works.

Here's a different set of examples with five cats and one dog, so p_1 the fraction of positive examples, a fraction of examples labeled one is 5/6 and so p_1 is about 0.83. If you read off that value at about 0.83 we find that the entropy of p_1 is about 0.65. And here I'm writing it only to two significant digits. Here's one more example. This sample of six images has all cats so p_1 is six out of six because all six are cats and the entropy of p_1 is this point over here which is zero.



Entropy as a measure of impurity

p_1 = fraction of examples that are cats

$p_1 = 3/6 \quad H(p_1) = 1$

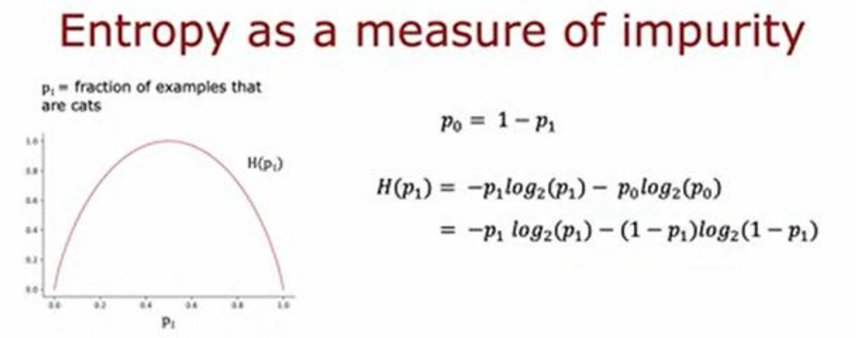
$p_1 = 5/6 \quad H(p_1) = 0.65$

$p_1 = 6/6 \quad H(p_1) = 0$

We see that as you go from 3/6 to six out of six cats, the impurity decreases from one to zero or in other words, the purity increases as you go from a 50-50 mix of cats and dogs to all cats. Let's look at a few more examples. Here's another sample with two cats and four dogs, so p_1 here is 2/6 which is 1/3, and if you read off the entropy at 0.33 it turns out to be about 0.92.

This is actually quite impure and in particular this set is more impure than this set because it's closer to a 50-50 mix, which is why the impurity here is 0.92 as opposed to 0.65. Finally, one last example, if we have a set of all six dogs then p_1 is equal to 0 and the entropy of p_1 is just this number down here which is equal to 0 so there's zero impurity or this would be a completely pure set of all not cats or all dogs. Now, let's look at the actual equation for the entropy function H(p_1).

Recall that p_1 is the fraction of examples that are equal to cats so if you have a sample that is 2/3 cats then that sample must have 1/3 not cats.



Entropy as a measure of impurity

p_1 = fraction of examples that are cats

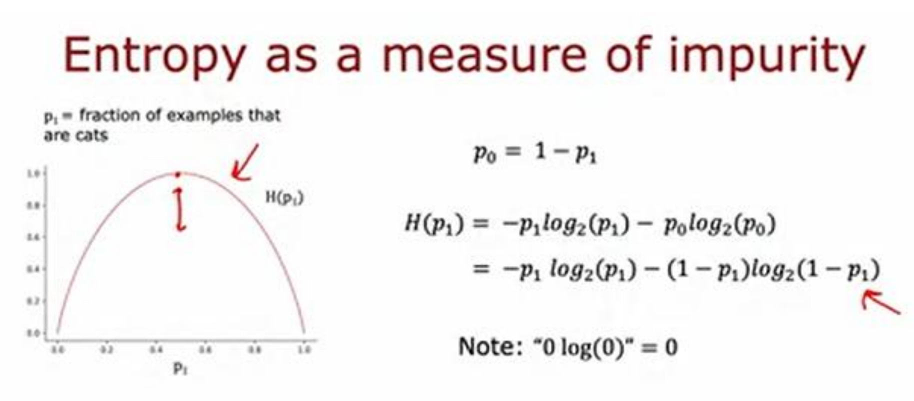
$$p_0 = 1 - p_1$$

$$H(p_1) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$
$$= -p_1\, log_2(p_1) - (1-p_1) log_2(1-p_1)$$

Let me define p_0 to be equal to the fraction of examples that are not cats to be just equal to 1 minus p_1. The entropy function is then defined as negative p_1log_2 (p_1), and by convention when computing entropy we take logs to base two rather than to base e, and then minus p_0log_2(p_0). Alternatively, this is also equal to negative p_1log_2(p_1) minus 1 minus p_1 log_2(1 minus p_1).

If you were to plot this function in a computer you will find that it will be exactly this function on the left. We take log_2 just to make the peak of this curve equal to one, if we were to take log_e or the base of natural logarithms, then that just vertically scales this function, and it will still work but the numbers become a bit hard to interpret because the peak of the function isn't a nice round number like one anymore.

One note on computing this function, if p_1 or p_0 is equal to 0 then an expression like this will look like 0log(0), and log(0) is technically undefined, it's actually negative infinity. But by convention for the purposes of computing entropy, we'll take 0log(0) to be equal to 0 and that will correctly compute the entropy as zero or as one to be equal to zero.



Entropy as a measure of impurity

p_1 = fraction of examples that are cats

$$p_0 = 1 - p_1$$

$$H(p_1) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$
$$= -p_1\, log_2(p_1) - (1-p_1) log_2(1-p_1)$$

Note: "0 log(0)" = 0

If you're thinking that this definition of entropy looks a little bit like the definition of the logistic loss that we learned about in the last course, there is actually a mathematical rationale for why these two formulas look so similar. But you don't have to worry about it and we won't get into it in this class. But applying this formula for entropy should work just fine when you're building a decision tree.

To summarize, the entropy function is a measure of the impurity of a set of data. It starts from zero, goes up to one, and then comes back down to zero as a function of the fraction of positive examples in your sample. There are other functions that look like this, they go from zero up to one and then back down. For example, if you look in open source packages you may also hear about something called the Gini criteria, which is another function that looks a lot like the entropy function, and that will work well as well for building decision trees.

But for the sake of simplicity, in these videos I'm going to focus on using the entropy criteria which will usually work just fine for most applications. Now that we have this definition of entropy, in the next video let's take a look at how you can actually use it to make decisions as to what feature to split on in the nodes of a decision tree.