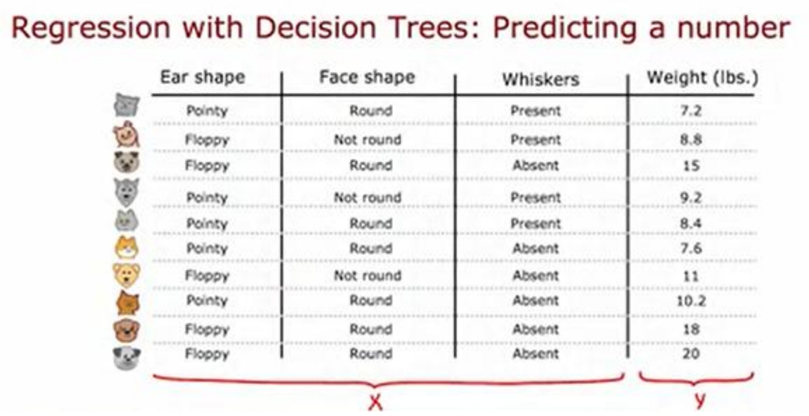
Regression trees (optional)

So far we've only been talking about decision trees as classification algorithms. In this optional video, we'll generalize decision trees to be regression algorithms so that we can predict a number. Let's take a look. The example I'm going to use for this video will be to use these three valued features that we had previously, that is, these features X, In order to predict the weight of the animal, Y.

So just to be clear, the weight here, unlike the previous video is no longer an input feature. Instead, this is the target output, Y, that we want to predict rather than trying to predict whether or not an animal is or is not a cat. This is a regression problem because we want to predict a number, Y.



Regression with Decision Trees: Predicting a number

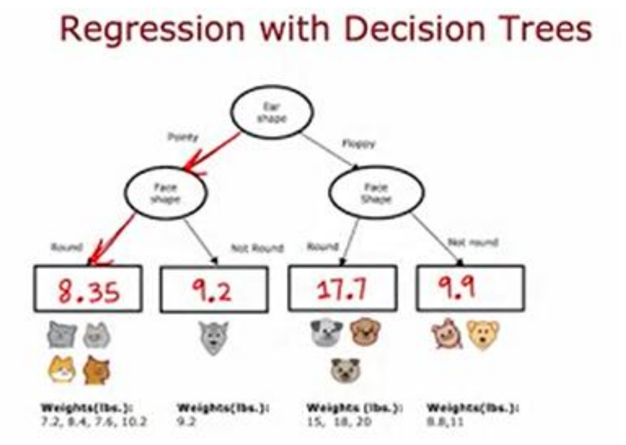
Ear shape	Face shape	Whiskers	Weight (lbs.)
Pointy	Round	Present	7.2
Floppy	Not round	Present	8.8
Floppy	Round	Absent	15
Pointy	Not round	Present	9.2
Pointy	Round	Present	8.4
Pointy	Round	Absent	7.6
Floppy	Not round	Absent	11
Pointy	Round	Absent	10.2
Floppy	Round	Absent	18
Floppy	Round	Absent	20

Let's look at what a regression tree will look like. Here I've already constructed a tree for this regression problem where the root node splits on ear shape and then the left and right sub tree split on face shape and also face shape here on the right. And there's nothing wrong with a decision tree that chooses to split on the same feature in both the left and right side branches. It's perfectly fine if the splitting algorithm chooses to do that.

If during training, you had decided on these splits, then this node down here would have these four animals with weights 7.2, 7.6 and 10.2. This node would have this one animal with weight 9.2 and so on for these remaining two nodes. So, the last thing we need to fill in for this decision tree is if there's a test example that comes down to this node, what is there weights that we should predict for an animal with pointy ears and a round face shape?

The decision tree is going to make a prediction based on taking the average of the weights in the training examples down here. And by averaging these four numbers, it turns out you get 8.35. If on the other hand, an animal has pointy ears and a not round face shape, then it will predict 9.2 or 9.2 pounds because that's the weight of this one animal down here. And similarly, this will be 17.70 and 9.90.

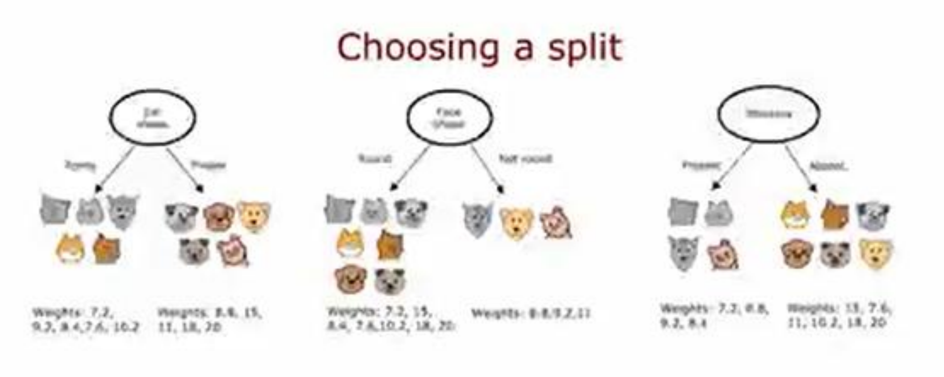
So, what this model will do is given a new test example, follow the decision nodes down as usual until it gets to a leaf node and then predict that value at the leaf node which I had just computed by taking an average of the weights of the animals that during training had gotten down to that same leaf node. So, if you were constructing a decision tree from scratch using this data set in order to predict the weight. The key decision as you've seen earlier this week will be, how do you choose which feature to split on?


Regression with Decision Trees

Let me illustrate how to make that decision with an example. At the root node, one thing you could do is split on the ear shape and if you do that, you end up with left and right branches of the tree with five animals on the left and right with the following weights. If you were to choose the split on the face shape, you end up with these animals on the left and right with the corresponding weights that are written below.

And if you were to choose to split on whiskers being present or absent, you end up with this. So, the question is, given these three possible features to split on at the root node, which one do you want to pick that gives the best predictions for the weight of the animal?

When building a regression tree, rather than trying to reduce entropy, which was that measure of impurity that we had for a classification problem, we instead try to reduce the variance of the weight of the values Y at each of these subsets of the data. So, if you've seen the notion of variants in other contexts, that's great.


Choosing a split
Ear shape
Pointy
Floppy
Weights: 7.2, 9.2, 8.4,7.6, 10.2
Weights: 8.8, 15, 11, 18, 20
Face Shape
Round
Not round
Weights: 7.2, 15, 8.4, 7.6,10.2, 18, 20
Weights: 8.8,9.2,11
Whiskers
Present
Absent
Weights: 7.2, 8.8, 9.2, 8.4
Weights: 15, 7.6, 11, 10.2, 18, 20

This is the statistical mathematical notion of variants that we'll used in a minute. But if you've not seen how to compute the variance of a set of numbers before, don't worry about it. All you need to know for this slide is that variants informally computes how widely a set of numbers varies. So for this set of numbers 7.2, 9.2 and so on, up to 10.2, it turns out the variance is 1.47, so it doesn't vary that much.

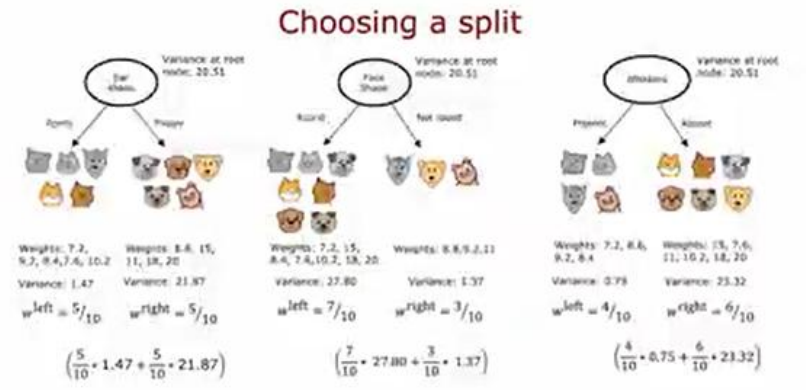
Meanwhile, here 8.8, 15, 11, 18 and 20, these numbers go all the way from 8.8 all the way up to 20. And so the variance is much larger, turns out to the variance of 21.87. And so the way we'll evaluate the quality of the split is, we'll compute same as before, W left and W right as the fraction of examples that went to the left and right branches.

And the average variance after the split is going to be 5/10, which is W left times 1.47, which is the variance on the left and then plus 5/10 times the variance on the right, which is 21.87. So, this weighted average variance plays a very similar role to the weighted average entropy that we had used when deciding what split to use for a classification problem.

And we can then repeat this calculation for the other possible choices of features to split on. Here in the tree in the middle, the variance of these numbers here turns out to be 27.80. The variance here is 1.37. And so with W left equals seven-tenths and W right as three-tenths, and so with these values, you can compute the weighted variance as follows. Finally, for the last example, if you were to split on the whiskers feature, this is the variance on the left and right, there's W left and W right.

And so the weight of variance is this. A good way to choose a split would be to just choose the value of the weighted variance that is lowest. Similar to when we're computing information gain, I'm going to make just one more modification to this equation. Just as for the classification problem, we didn't just measure the average weighted entropy, we measured the reduction in entropy and that was information gain.

For a regression tree we'll also similarly measure the reduction in variance. Turns out, if you look at all of the examples in the training set, all ten examples and compute the variance of all of them, the variance of all the examples turns out to be 20.51. And that's the same value for the roots node in all of these, of course, because it's the same ten examples at the roots node.


Choosing a split
Variance at root node: 20.51
Ear shape
Pointy
Floppy
Weights: 7.2, 9.2, 8.4, 7.6, 10.2
Weights: 8.8, 15, 11, 18, 20
Variance: 1.47
Variance: 21.87
$w^{left} = 5/10$
$w^{right} = 5/10$
$\left(\frac{5}{10} \cdot 1.47 + \frac{5}{10} \cdot 21.87\right)$
Variance at root node: 20.51
Face Shape
Round
Not round
Weights: 7.2, 15, 8.4, 7.6, 10.2, 18, 20
Weights: 8.8, 9.2, 11
Variance: 27.80
Variance: 1.37
$w^{left} = 7/10$
$w^{right} = 3/10$
$\left(\frac{7}{10} \cdot 27.80 + \frac{3}{10} \cdot 1.37\right)$
Variance at root node: 20.51
Whiskers
Present
Absent
Weights: 7.2, 8.8, 9.2, 8.4
Weights: 15, 7.6, 11, 10.2, 18, 20
Variance: 0.75
Variance: 23.32
$w^{left} = 4/10$
$w^{right} = 6/10$
$\left(\frac{4}{10} \cdot 0.75 + \frac{6}{10} \cdot 23.32\right)$

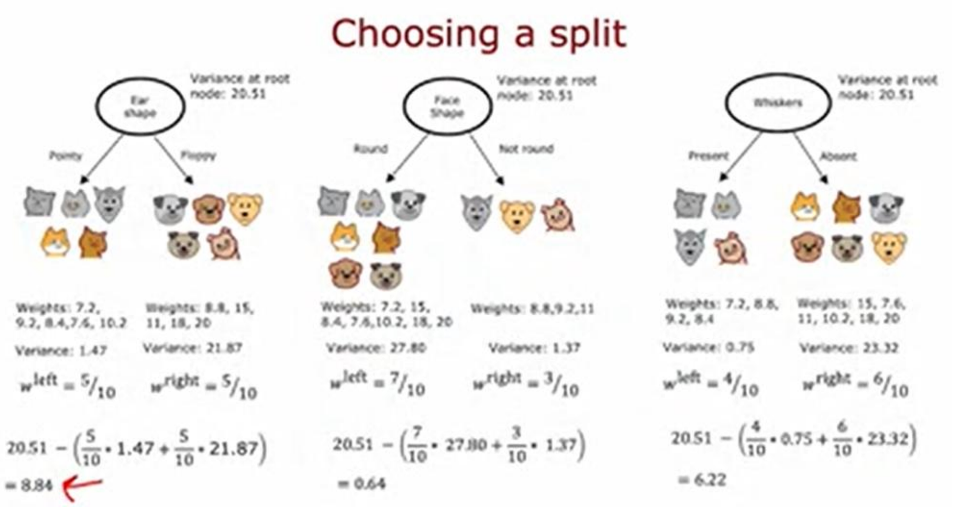
And so what we'll actually compute is the variance of the roots node, which is 20.51 minus this expression down here, which turns out to be equal to 8.84. And so at the roots node, the variance was 20.51 and after splitting on ear shape, the average weighted variance at these two nodes is 8.84 lower. So, the reduction in variance is 8.84.

And similarly, if you compute the expression for reduction in variance for this example in the middle, it's 20.51 minus this expression that we had before, which turns out to be equal to 0.64. So, this is a very small reduction in variance. And for the whiskers feature you end up with this which is 6.22. So, between all three of these examples, 8.84 gives you the largest reduction in variance.

So, just as previously we would choose the feature that gives you the largest information gain for a regression tree, you will choose the feature that gives you the largest reduction in variance, which is why you choose ear shape as the feature to split on.

Having chosen the ear shape feature to split on, you now have two subsets of five examples in the left and right side branches and you would then, again, we say recursively, where you take these five examples and do a new decision tree focusing on just these five examples, again, evaluating different options of features to split on and picking the one that gives you the biggest variance reduction. And similarly on the right. And you keep on splitting until you meet the criteria for not splitting any further. And so that's it.

With this technique, you can get your decision treat to not just carry out classification problems, but also regression problems. So far, we've talked about how to train a single decision tree.



Choosing a split

Ear shape
Variance at root node: 20.51
Pointy
Floppy
Face Shape
Variance at root node: 20.51
Round
Not round
Whiskers
Variance at root node: 20.51
Present
Absent

Weights: 7.2, 9.2, 8.4,7.6, 10.2

Weights: 8.8, 15, 11, 18, 20

Variance: 1.47

Variance: 21.87

$w^{\text{left}} = {}^{5}/_{10}$ $w^{\text{right}} = {}^{5}/_{10}$

$$20.51 - \left(\frac{5}{10} \cdot 1.47 + \frac{5}{10} \cdot 21.87\right)$$

$= 8.84$

Weights: 7.2, 15, 8.4, 7.6,10.2, 18, 20

Weights: 8.8,9.2,11

Variance: 27.80

Variance: 1.37

$w^{\text{left}} = {}^{7}/_{10}$ $w^{\text{right}} = {}^{3}/_{10}$

$$20.51 - \left(\frac{7}{10} \cdot 27.80 + \frac{3}{10} \cdot 1.37\right)$$

$= 0.64$

Weights: 7.2, 8.8, 9.2, 8.4

Weights: 15, 7.6, 11, 10.2, 18, 20

Variance: 0.75

Variance: 23.32

$w^{\text{left}} = {}^{4}/_{10}$ $w^{\text{right}} = {}^{6}/_{10}$

$$20.51 - \left(\frac{4}{10} \cdot 0.75 + \frac{6}{10} \cdot 23.32\right)$$

$= 6.22$