MITOCW | MIT15_071S17_Session_4.2.05_300k

In the previous video, we generated a CART tree with three splits, but why not two, or four, or even five?

There are different ways to control how many splits are generated.

One way is by setting a lower bound for the number of data points in each subset.

In R, this is called the minbucket parameter, for the minimum number of observations in each bucket or subset.

The smaller minbucket is, the more splits will be generated.

But if it's too small, overfitting will occur.

This means that CART will fit the training set almost perfectly.

But this is bad because then the model will probably not perform well on test set data or new data.

On the other hand, if the minbucket parameter is too large, the model will be too simple and the accuracy will be poor.

Later in the lecture, we will learn about a nice method for selecting the stopping parameter.

In each subset of a CART tree, we have a bucket of observations, which may contain both possible outcomes.

In the small example we showed in the previous video, we have classified each subset as either red or gray depending on the majority in that subset.

In the Supreme Court case, we'll be classifying observations as either affirm or reverse.

Instead of just taking the majority outcome to be the prediction, we can compute the percentage of data in a subset of each type of outcome.

As an example, if we have a subset with 10 affirms and two reverses, then 87% of the data is affirm.

Then, just like in logistic regression, we can use a threshold value to obtain our prediction.

For this example, we would predict affirm with a threshold of 0.5 since the majority is affirm.

But if we increase that threshold to 0.9, we would predict reverse for this example.

Then by varying the threshold value, we can compute an ROC curve and compute an AUC value to evaluate our model.

In the next video, we'll build a CART tree in R to predict the decisions of Justice Stevens and evaluate our model using a ROC curve.