Chapter 8 Exercises: Tree-Based Methods

Statistical Learning with R

Jon Geiger

March 19, 2022

Conceptual Exercise 5

Suppose we produce ten bootstrapped samples from a data set containing red and gree classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{ClassisRed}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the classification under each of these two approaches?

Solution

Because each of these values represents the probability that the class is red, there is a simple, logical cutoff value for how each of these estimates would classify Y.

The first approach is the majority vote, which involves looking at how each observation would classify the sample, and taking the majority. This is demonstrated here:

```
probs <- c(0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75)
classes_ind <- factor(ifelse(probs >= 0.5, "Red", "Green"))
classes_ind
```

```
##  [1] Green Green Green Green Red   Red   Red   Red   Red   Red
## Levels: Green Red
```

```
table(classes_ind)
```

```
## classes_ind
## Green   Red
##     4     6
```

We can see that the majority of the observations classify this X as Red, so the majority vote approach classifies it as red.

The second approach is to take the average probability and classify X according to the mean of the sampled probability. This is demonstrated here:

```r
result <- ifelse(mean(probs) >= 0.5, "Red", "Green")
cat("Mean of probabilities: ", mean(probs), "\n",
    "Classification: ", ifelse(mean(probs) >= 0.5, "Red", "Green"), sep = "")
```

```
## Mean of probabilities: 0.45
## Classification: Green
```

This approach has the opposite classification as the majority vote approach. This is because taking the average probability uses the mean, which is sensitive to outliers and skew, whereas the majority vote approach is closer to using the median on a skewed dataset. Even if the mean is below 50%, the median of the data set still falls above that value at 0.575.

Applied Exercise 10

Solution