Q34

Rahul Atre

2023-11-16

Q34 [Ensemble Learning]

Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of P(Class is Red|X): 0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

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

Let us first load the probability estimates data into a list:

```r
probability_estimates = c(0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75)
```

If we consider 0.5 as the decision threshold probability, then for any estimate that has a probability of 0.5 or higher would be in class Red, whereas less than 0.5 would be in Class green.

Calculating the majority vote for both classes:

```r
sum(probability_estimates >= 0.5) #Number of estimates in Class Red
```

```
## [1] 6
```

```r
sum(probability_estimates < 0.5) #Number of estimates in Class Green
```

```
## [1] 4
```

Therefore, for the majority vote approach, the final classification is **Class Red** since there are more estimates with a probability of 0.5 or higher.

For the average probability approach, we can calculate the average probability of the 10 estimates and compare it to the decision threshold probability, which is 0.5. Similar to majority vote, if the average probability is greater or equal to 0.5, it classifies in Class Red, and Green otherwise.

```r
prob_estimate_mean = mean(probability_estimates)
prob_estimate_mean
```

```
## [1] 0.45
```

Since the average probability of the estimates is 0.45, which is less than 0.5, the final classification would be **Class Green**.