WS #15 - Bagging

Monday, November 3, 2025

Math 154 - Jo Hardin

Your Name: _____

Names of people you worked with: _____

What spring class are you most excited about?

Task:

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$|X$):

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

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 (for this particular observation) under each of the two approaches?

Also: to discuss with your classmate: how were the numbers 0.1, 0.15, 0.2, etc. calculated?

Solution:

Majority vote:

In 6 out of the 10 trees, the value of X was predicted to be Red. The predicted value is **RED**.

Average Probability:

```
mean(c(0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75))
```

```
[1] 0.45
```

The average predicted probability is 0.45, so the predicted value of X is **GREEN**.

Majority vote is used in R packages like **ranger** by default, partly because the default prediction is `type = "response"` (which means the prediction output will be either red or green). If the prediction is changed to `type = "prob"` then the RF classification will be done using average probability.

How were the numbers 0.1, 0.15, 0.2, etc. calculated?

Answer: the observation X is sent down the trained tree until it gets to a terminal node. The number 0.1 indicates that 10% of the training observations in that terminal node are red.