WU #14 - CART

Math 154 - Jo Hardin

Thursday, October 28, 2021

Name: _____

Consider the decision tree and resulting fit from running a model to classify the penguin home island.

Let $|T|$ be the number of nodes in a given tree.

1. Find (as a function of α)
$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{i \in R_m} I(y_i \neq k(m)) + \alpha \cdot |T|$$
for the final tree as well as two different trees with **one** fewer terminal nodes.
2. For what value of α would you choose a tree with 9 nodes? For what value of α would you choose a tree with 8 nodes?

Solution:

1. For the full tree, there are 49 misclassificiations (5+5+11+8+4+9+7 = 49).

If we prune back `year`, we go from 16 (5+11) missclassifications (in those two nodes) to 19 misclassifications (3 additional missclassificaitons by pruning).

If we prune back `bill_length_mm`, we go from 16 (9+7) missclassifications (in those two nodes) to 20 misclassifications (4 additional missclassificaitons by pruning).

We will prune back `year`.

$$C_\alpha(T = 9) = 49 + \alpha \cdot 9$$

$$C_\alpha(T = 8) = 53 + \alpha \cdot 8$$

2.
$$\begin{aligned} C_\alpha(T = 9) &< C_\alpha(T = 8) \\ 49 + \alpha \cdot 9 &< 53 + \alpha \cdot 8 \\ \alpha &< 4 \end{aligned}$$

If $\alpha < 4$, keep the tree with 9 terminal nodes. If $\alpha > 4$, keep the tree with 8 terminal nodes.

```
## == Workflow [trained] ==========================================
## Preprocessor: Recipe
## Model: decision_tree()
##
## -- Preprocessor ----------------------------------------------
## 1 Recipe Step
##
## * step_mutate()
##
## -- Model -----------------------------------------------------
## n= 258
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 258 140 Biscoe (0.473 0.368 0.159)
##     2) species=Gentoo 90   0 Biscoe (1.000 0.000 0.000) *
##     3) species=Adelie,Chinstrap 168  73 Dream (0.190 0.565 0.244)
##       6) species=Chinstrap 49   0 Dream (0.000 1.000 0.000) *
##       7) species=Adelie 119  73 Dream (0.269 0.387 0.345)
##        14) bill_length_mm>=38 76  48 Torgersen (0.329 0.303 0.368)
##          28) bill_length_mm< 38 14   5 Biscoe (0.643 0.214 0.143) *
##          29) bill_length_mm>=38 62  36 Torgersen (0.258 0.323 0.419)
##            58) bill_length_mm>=39 49  32 Dream (0.306 0.347 0.347)
##             116) bill_length_mm< 41 32  19 Dream (0.344 0.406 0.250)
##               232) year=2008 10   5 Biscoe (0.500 0.200 0.300) *
##               233) year=2007,2009 22  11 Dream (0.273 0.500 0.227) *
##             117) bill_length_mm>=41 17   8 Torgersen (0.235 0.235 0.529) *
##            59) bill_length_mm< 39 13   4 Torgersen (0.077 0.231 0.692) *
##        15) bill_length_mm< 38 43  20 Dream (0.163 0.535 0.302)
##          30) bill_length_mm>=36 29   9 Dream (0.103 0.690 0.207) *
##          31) bill_length_mm< 36 14   7 Torgersen (0.286 0.214 0.500) *
```