WS #13 - CART

Monday, October 28, 2024

Your Name: _____

Names of people you worked with: _____

What is your favorite type of bread? And, more importantly, have you ever made it yourself?

Task:

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

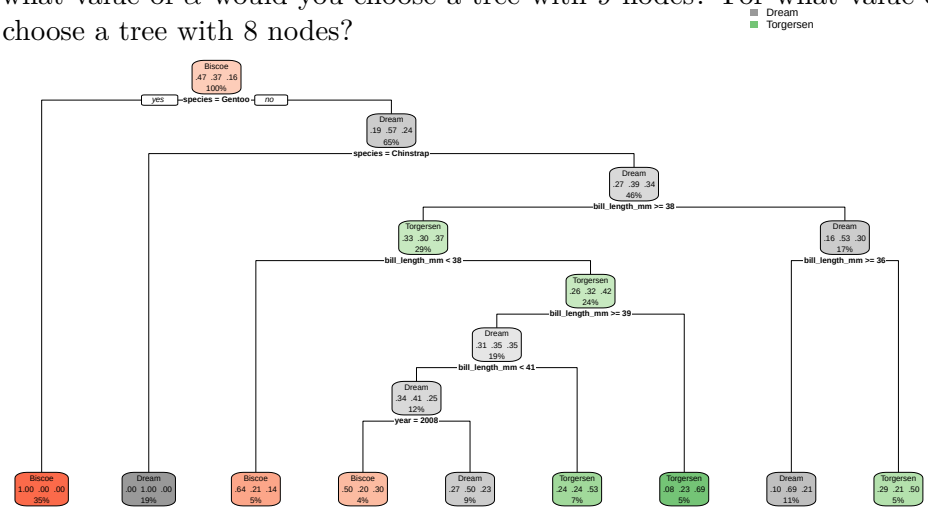
Let $|T|$ be the number of nodes in a given tree.

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?



```
== 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 258 136 Biscoe (0.47286822 0.36821705 0.15891473)
    2) species=Gentoo 90   0 Biscoe (1.00000000 0.00000000 0.00000000) *
    3) species=Adelie,Chinstrap 168  73 Dream (0.19047619 0.56547619 0.24404762)
      6) species=Chinstrap 49   0 Dream (0.00000000 1.00000000 0.00000000) *
      7) species=Adelie 119  73 Dream (0.26890756 0.38655462 0.34453782)
       14) bill_length_mm>=37.55 76  48 Torgersen (0.32894737 0.30263158 0.36842105)
         28) bill_length_mm< 38.3 14   5 Biscoe (0.64285714 0.21428571 0.14285714) *
         29) bill_length_mm>=38.3 62  36 Torgersen (0.25806452 0.32258065 0.41935484)
           58) bill_length_mm>=39.4 49  32 Dream (0.30612245 0.34693878 0.34693878)
            116) bill_length_mm< 41.35 32  19 Dream (0.34375000 0.40625000 0.25000000)
              232) year=2008 10   5 Biscoe (0.50000000 0.20000000 0.30000000) *
              233) year=2007,2009 22  11 Dream (0.27272727 0.50000000 0.22727273) *
            117) bill_length_mm>=41.35 17   8 Torgersen (0.23529412 0.23529412 0.52941176) *
           59) bill_length_mm< 39.4 13   4 Torgersen (0.07692308 0.23076923 0.69230769) *
       15) bill_length_mm< 37.55 43  20 Dream (0.16279070 0.53488372 0.30232558)
         30) bill_length_mm>=35.55 29   9 Dream (0.10344828 0.68965517 0.20689655) *
         31) bill_length_mm< 35.55 14   7 Torgersen (0.28571429 0.21428571 0.50000000) *
```

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`.

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

$$C_\alpha(T = 8) = 52 + \alpha \cdot 8$$

2.

$$\begin{aligned} C_\alpha(T = 9) &< C_\alpha(T = 8) \\ 49 + \alpha \cdot 9 &< 52 + \alpha \cdot 8 \\ \alpha &< 3 \end{aligned}$$

If $\alpha < 3$, keep the tree with 9 terminal nodes. If $\alpha > 3$, keep the tree with 8 terminal nodes.