Lecture 15: Classification Trees
STAT 452, Spring 2021

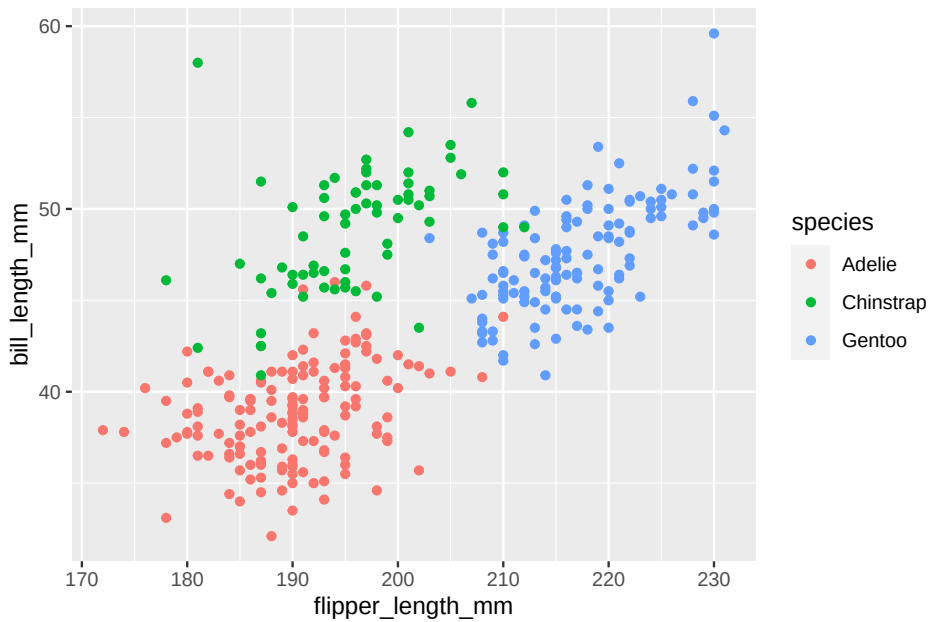
- A **classification tree** is very similar to a regression tree, except that it is used to predict a qualitative response rather than a quantitative response.

- Recall for regression trees, the prediction for a new test set observation is given by the mean response of the training observations that belong to the same terminal note.

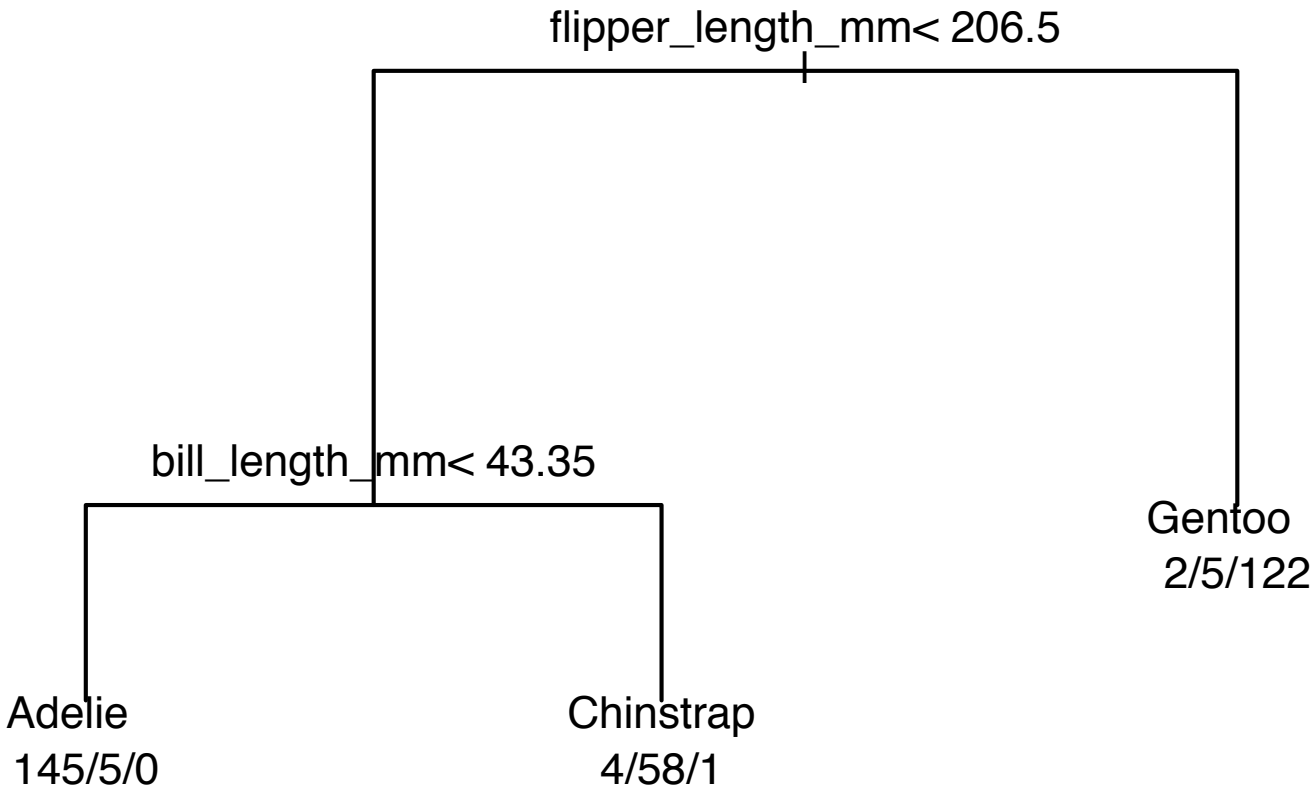
- In contrast, for a classification tree, the class prediction for a new test set observation is given by *most commonly occurring class* of training observations that belong to the same terminal node.

Example

- To illustrate a classification tree we'll use the `penguins` data.
- The response variable is the `species` of penguin: Adelie, Chinstrap, and Gentoo.
- The predictor variables we'll consider are `flipper_length_mm` and `bill_length_mm`.

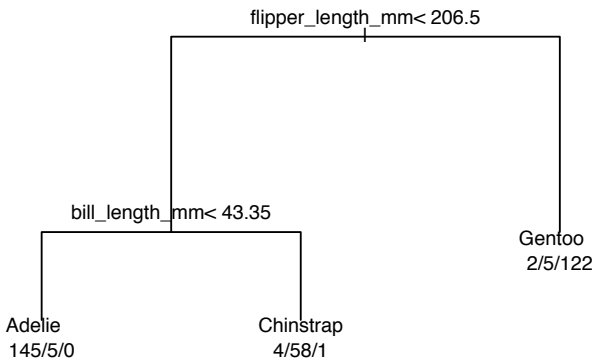


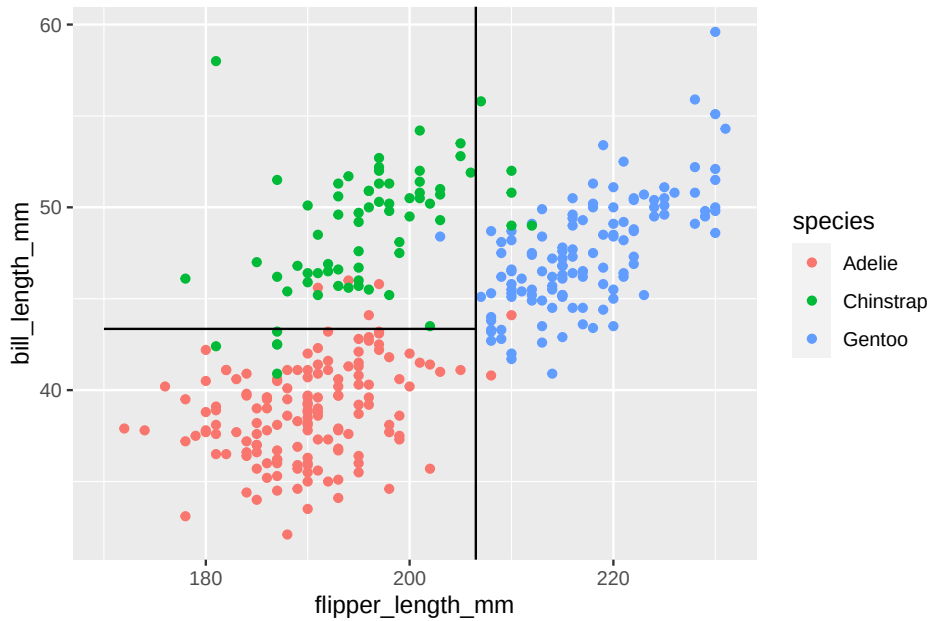
```
# load packages
> library(palmerpenguins)
> library(tidyverse)
> library(rpart)

# fit classification tree
> t1 <- rpart(species ~ flipper_length_mm + bill_length_mm,
              data = penguins, method = "class")

# plot classification tree
> par(cex=0.7, xpd=NA)
> plot(t1)
> text(t1, use.n = TRUE)
```


flipper_length_mm< 206.5
bill_length_mm< 43.35
Gentoo
2/5/122
Adelie
145/5/0
Chinstrap
4/58/1

- The tree has two internal notes and three terminal nodes (leaves).
- The tree partitions the predictor space into three regions.
- If `flipper_length_mm` ≥ 206.5, the predicted species is Gentoo.
- If `flipper_length_mm` < 206.5 and `bill_length_mm` ≥ 43.35, then the predicted species is Chinstrap.
- If `flipper_length_mm` < 206.5 and `bill_length_mm` < 43.35, then the predicted species is Adelie.

Illustration of the partitioning of the predictor space that corresponds with classification the tree. We see that the observations falling in each region are predominately of the same species.





Details of Classification Trees

How does the tree-building algorithm select the splitting rules?

- In the example, it appears that the splits were selected in such a way so that each partition predominately contained observations from a single class (species).
- If the observations that fall into a region are predominately a single class, that node is considered pure.
- There are various measures of **node-purity** that one can use as a splitting criteria.

- The most intuitive measure of node purity is the **misclassification error rate**. This is the fraction of the training observations in each region that do not belong to the most common class.

$$E_m = 1 - \max_k \hat{p}_{mk}$$

Here $\hat{p}_{mk}$ represents the proportion of training observations in the mth region that are from the kth class.

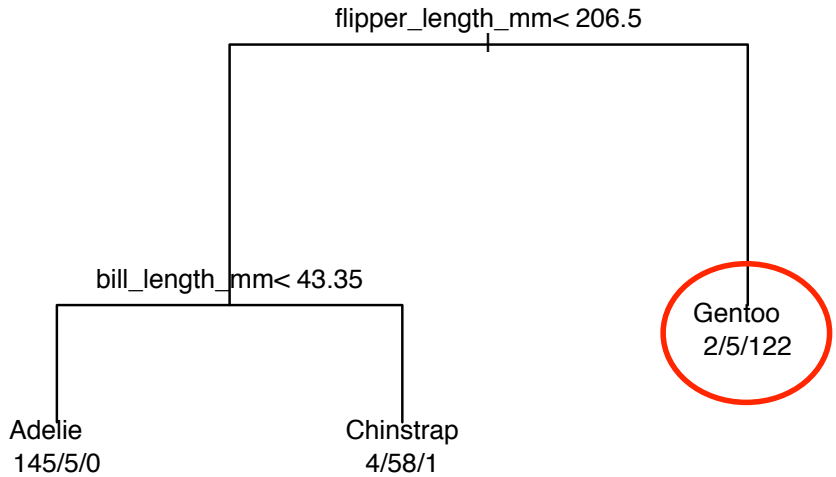
- Another measure of node purity is the **Gini index** defined by

$$G_m = \sum_{i=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

which is a measure of total variance across the K classes. It takes on a small value when all of the $\hat{p}_{mk}$'s are close to 0 or 1, which imples that the data in a region are predominately of a single class.

- By default `rpart()` uses the Gini index, which turns out to be a preferable metric for tree-building.

Your Turn



(a) Suppose we decide to go on a field trip to the Palmer, Archipelago in Antarctica. While there we find a penguin, and measure the flipper length to be 185 mm and the bill length to be 38 mm. Use the classification tree to predict the species?

(b) Calculate the missclassification error rate for the node circled in red.

Your Turn

Suppose we build a classification tree, where the response variable has 2 categories labeled A and B.

(a) Suppose that the data that fall into a certain terminal node are 90% A and 10% B. Calculate the Gini index for this node.

(b) Suppose that the data that fall into a certain terminal node are 60% A and 40% B. Calculate the Gini index for this node.

(c) For which node is the Gini index lower?