Lecture 15: Classification Trees

STAT 452, Spring 2021

Here we go over an example of using cross-validation (hold-out method) with the `penguins` data.

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

```
# remove missing data
penguins2 <- penguins %>%
  select(species, flipper_length_mm, bill_length_mm) %>%
  na.omit()
```

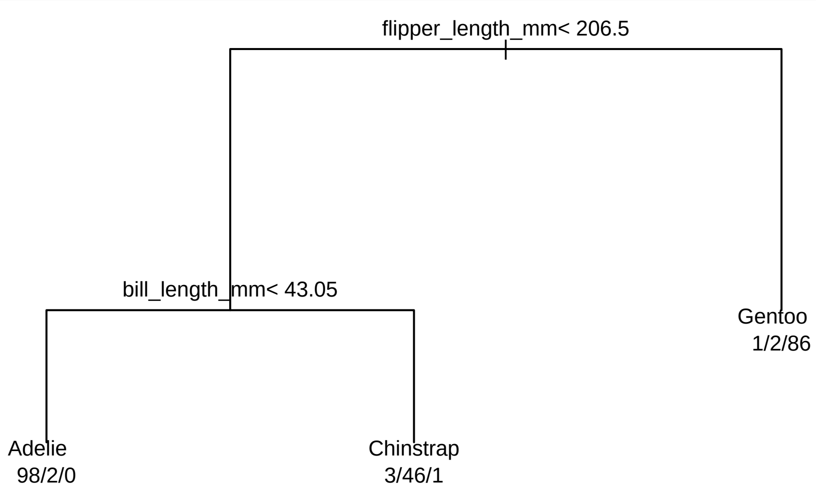
Randomly split the data into a 70% training and 30% test set.

```
set.seed(123)
n <- nrow(penguins2)
train_index <- sample(1:n, round(0.7*n))
penguins_train <- penguins2[train_index, ]
penguins_test <- penguins2[-train_index, ]
```

Next we fit a classification tree on the training set.

```
t1 <- rpart(species ~ flipper_length_mm + bill_length_mm,
            data = penguins_train, method = "class")
```

```
# plot of tree fit to training data
par(cex=0.7, xpd=NA)
plot(t1)
text(t1, use.n = TRUE)
```



Next we make predictions for the penguin species on the test set, and then compute the confusion matrix and accuracy.

```r
# make predictions on test set
preds1 <- predict(t1, newdata = penguins_test, type = "class")
```

```r
# make confusion matrix
tb <- table(prediction = preds1, actual = penguins_test$species)
addmargins(tb)
```

```
##            actual
## prediction  Adelie Chinstrap Gentoo Sum
##   Adelie        44         2      0  46
##   Chinstrap      4        13      0  17
##   Gentoo         1         3     36  40
##   Sum           49        18     36 103
```

```r
# Accuracy (percent correctly classified)
(44 + 13 + 36) / 103
```

```
## [1] 0.9029126
```

The accuracy is about 90%, which is comparable to what we got using kNN (leture 11). Although, the advantage of the classification tree, is that we have a tree model that is easy to interpret. kNN, on the other hand, is more of a "black-box", that's only useful for making predictions and does not describe the relationships between the predictors and response variable.