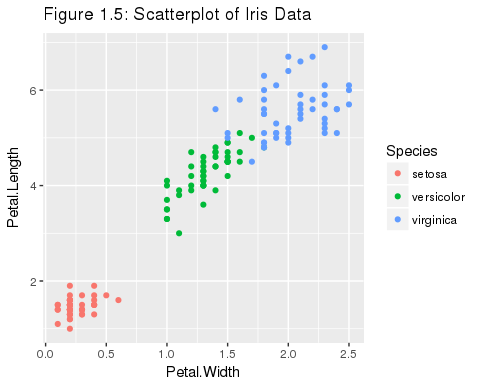
CART lab

Liam Niehus-Staab

11/19/2018

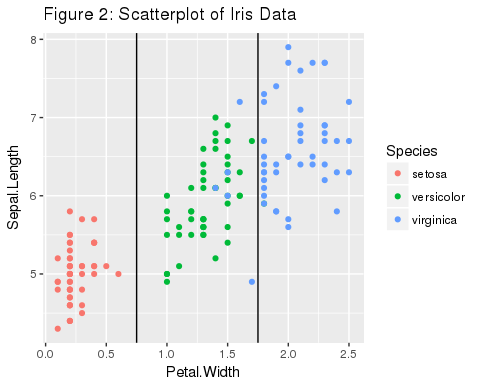
### 1) There are four explanatory variables in this model. Modify Figure 1 by choosing the two explanatory variables that creates the most distinct species groups. In other words, we want each species of iris in its own cluster.

ggplot(data = iris, aes(x = Petal.Width, y = Petal.Length)) +   
 geom\_point() + aes(colour = Species) + labs(title = "Figure 1.5: Scatterplot of Iris Data")



### 2) Create a new scatterplot with Petal.Width on the x-axis and Sepal.Width on the y-axis. Draw two verticle lines on this scatterplot. These lines will break up the predictor space into three regions. Ideally each region will contain only one species of iris.

ggplot(data = iris, aes(x = Petal.Width, y = Sepal.Length)) +   
 geom\_point() + aes(colour = Species) + labs(title = "Figure 2: Scatterplot of Iris Data") + geom\_vline(xintercept = 0.75) + geom\_vline(xintercept = 1.75)

 ### 3) Based upon the vertical lines created in Question 2, write mathematical rules to identify the three regions. For example, when x is between 0 and 4, we are in setosa region, when x is between 4 and 5 we are in versicolor region. These are called spliting rules and will be used to identify regions corresponding to specific species within our decision tree. 0 < x < .75 setosa .75 < x < 1.75 versicolor 1.75 < x < 2.6 verginica

### 4) Using the treeall model, if petal width = 1.5 and petal length = 5. what would you classify the species to be?

virginica

### 5) View the tree1table data frame. How many misclassifications were there? Using tree1, which species was the most difficult to accurately predict?

There were 6 miscalculations: 1 versicolor was predicted to be a virginica and 5 virginica were predicted to be versicolor.

### 6) Based upon the summary table, what is the misclassification rate?

8%

### 7) How many rows are in the iris[-train] dataset?

150

### 8) Why does changing the seed create different misclassification rates?

The seed determines the random number generation order that is used in training the tree. As such, the tree will be trained (and thus constructed) differently for each different random number seed. Obviously different trees will not always have the same misclassification rates.