Classification2

Prof. Eric A. Suess

March 11, 2020

Today

Today we will work with the C5.0 decision trees algorithm and the credit dataset from the book.

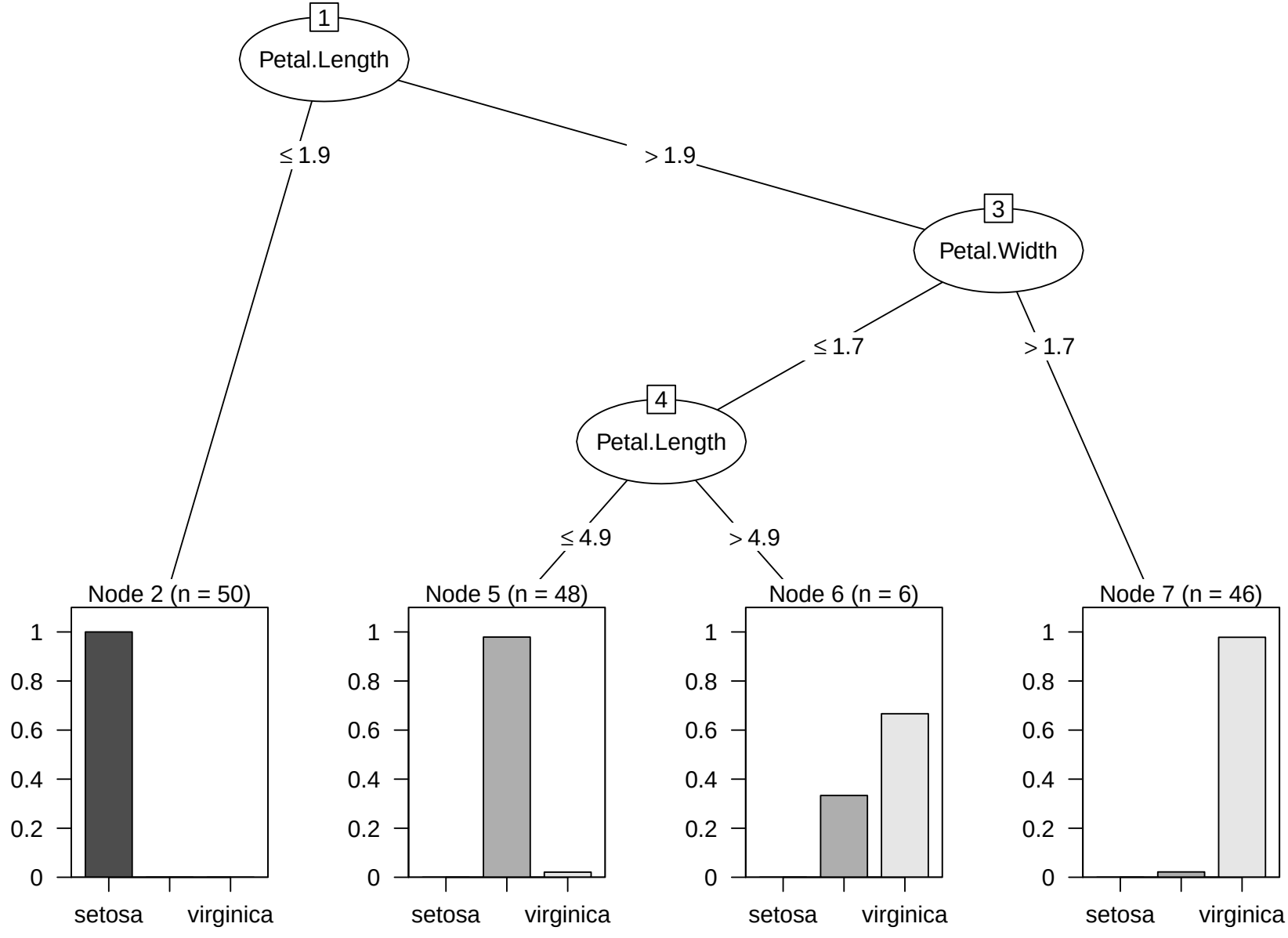
We will see one way to randomize the rows of a dataset.

We will also see how to make plots of trees.

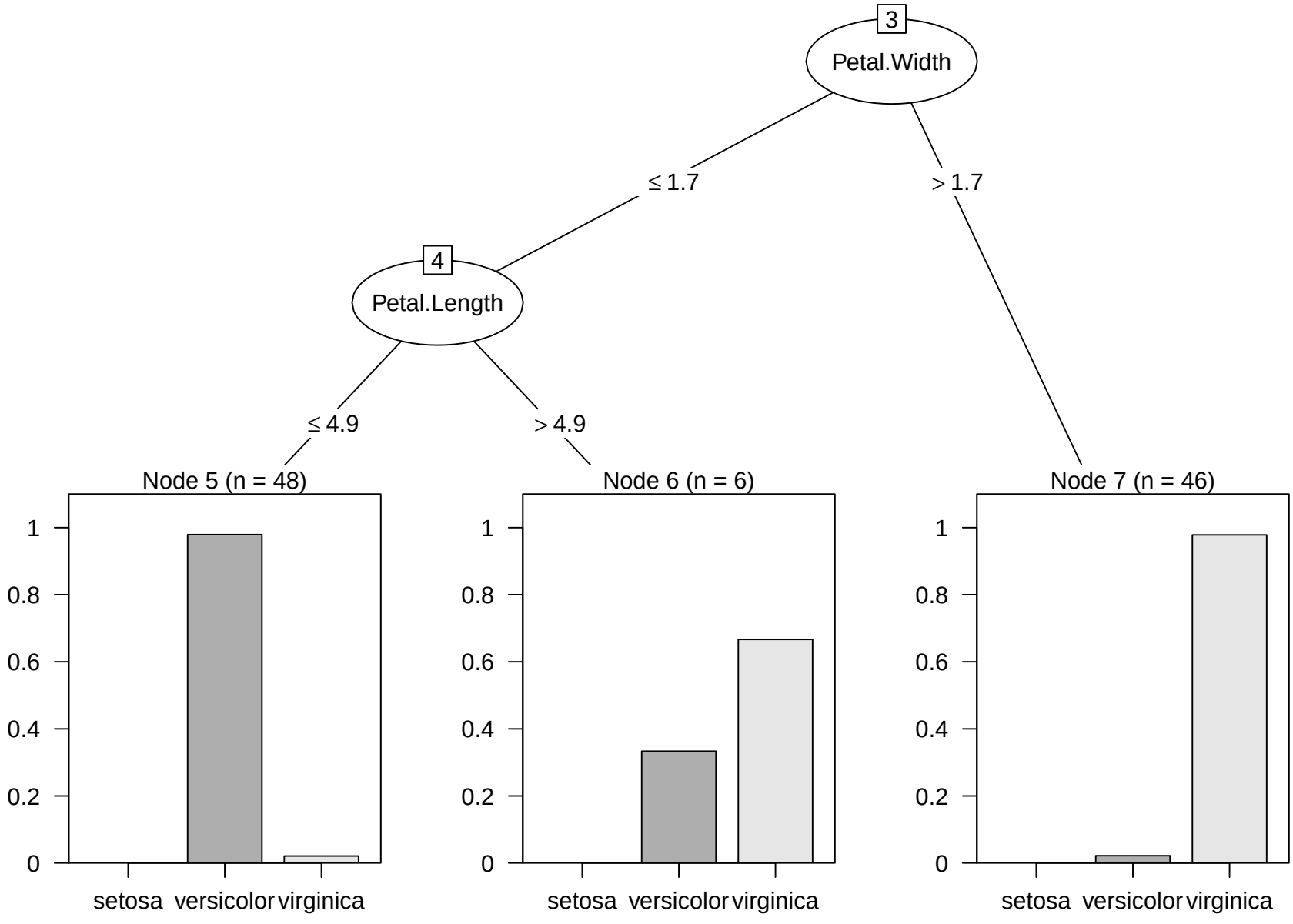
Package C5.0 in R

Here is the link to the documentation of R Package C5.0

plot for iris data



sub plot for iris data



plot the C5.0 tree

In the book there is no tree plotted. Why?

I do not know. Seems trees are nice to plot.

plots of trees from rpart

Here is another blog post about making nice plots of trees.

Revolution Analytics: Plotting trees

Example – identifying risky bank loans using C5.0 decision trees

2007-2008 bad years for the banking industry.

How to identify risky loans?

Who can get a loan and who cannot? Why?

Decision trees are very nice and give the model in plain language.

Banks use decision trees to try and minimize potential losses, i.e., minimize making bad loans.

Step 1 – collecting the data

The credit data is from the UCI Machine Learning Data Repository.

Note that the data is from Germany and the currency is in Deutsche Marks (DM).

Step 2 – exploring and preparing the data

To randomize the dataset the following R code is used to randomize the index.

```
set.seed(123)

train_sample <- sample(1000, 900)

credit_train <- credit[train_sample, ]

credit_test <- credit[-train_sample, ]
```

Step 3 – train the model

Run the model, see the output.

It is here we would like to see a plot of the tree.

Also, it should be noted that there is a tendency for decision trees to overfit the model to the training data. So the error rates on the training data may be overly optimistic.

So it is important to evaluate the decision tree on the test dataset.

Step 4 – evaluating model performance

Using the training data evaluate the model.

Model only predicted 50% of the defaulted loans. Not so good.

Step 5 – improving model performance

Adaptive Boosting

This is a process in which many trees are built and the trees vote on the best class for each example.

According to the author, "... boosting is rooted in the notion that by combining a number of **weak learners**, you can create a **team that is much stronger** that any one of the learners alone."

Step 5 – improving model performance

Cost

rattle

It would be good to have an easy way to get the plot of the tree.

Try rattle