Machine Learning - Decision Trees Exercises/Solution

Jan-Philipp Kolb

02 Juni, 2019

Exercise - rpart Kyphosis

Consider the Kyphosis data frame

1) Which variables are in the `kyphosis` dataset
2) Build a tree to classify Kyphosis from Age, Number and Start.

Consider the tree build above.

3) Which variables are used to explain Kyphosis presence?
4) How many observations contain the terminal nodes.

Consider the Kyphosis data frame.

5) Build a tree using the first 60 observations of kyphosis.
6) Predict the kyphosis presence for the other 21 observations.
7) Which is the misclassification rate (prediction error)

The dataset kyphosis

The dataset contains (1):

- Kyphosis: a factor with levels absent present indicating if a kyphosis (a type of deformation) was present after the operation.
- Age: in months.
- Number: the number of vertebrae involved.
- Start: the number of the first (topmost) vertebra operated on.

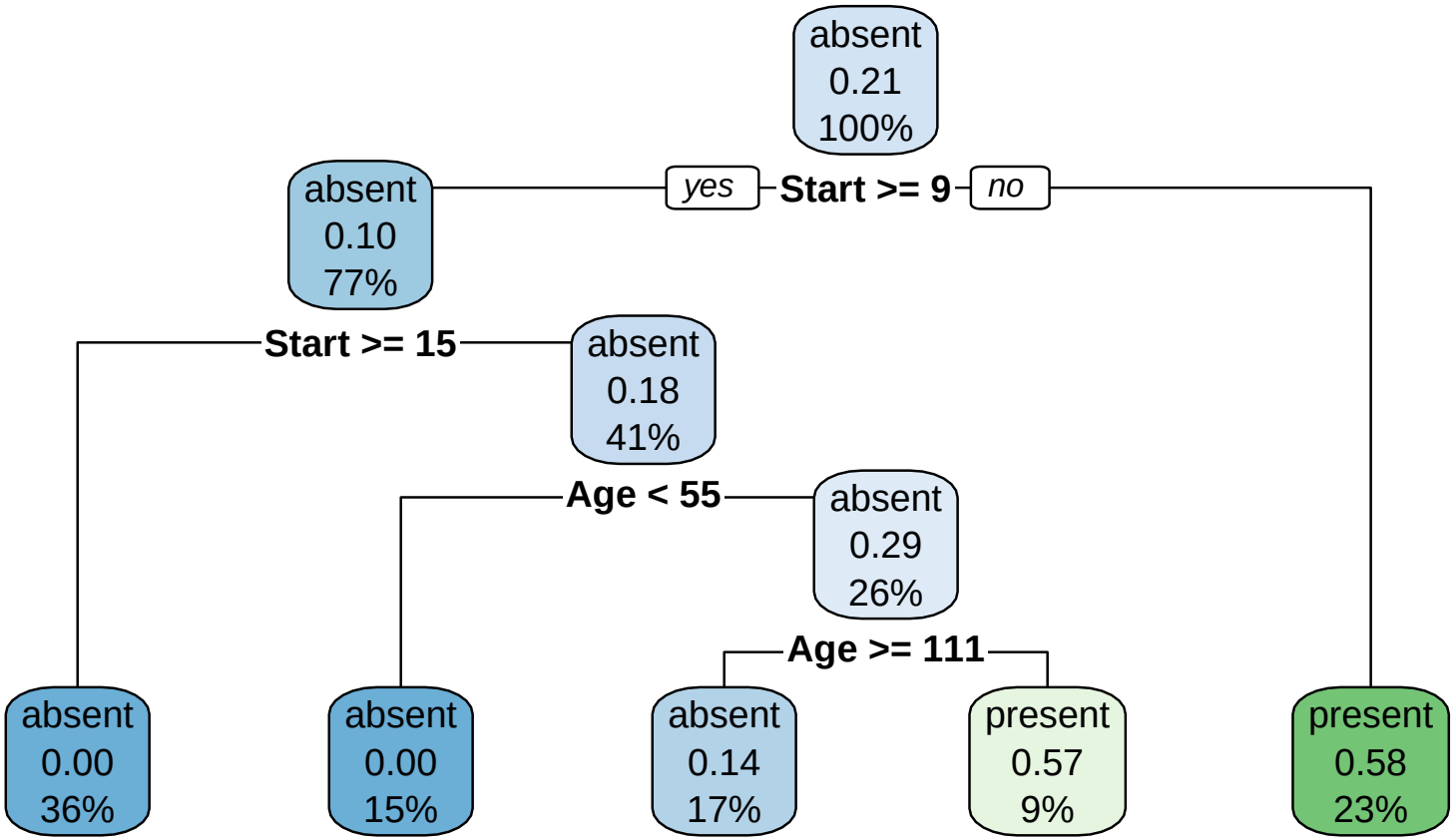
Build the tree (2)

```
library('rpart')
TREE <- rpart(Kyphosis ~ Age + Number + Start,
              data=kyphosis,method="class")
TREE
```

```
## n= 81
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 81 17 absent (0.79012346 0.20987654)
##    2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
##      4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
##      5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
##       10) Age< 55 12  0 absent (1.00000000 0.00000000) *
##       11) Age>=55 21  6 absent (0.71428571 0.28571429)
##         22) Age>=111 14  2 absent (0.85714286 0.14285714) *
##         23) Age< 111 7  3 present (0.42857143 0.57142857) *
```

Plot the result

```
rpart.plot::rpart.plot(TREE)
```



Answers

3) Which variables are used to explain Kyphosis presence?
   - The variables are Start and Age
4) How many observations contain the terminal nodes.
   - *denotes terminal nodes. The nodes have 29, 12, 14, 7 and 19 observations

Consider the Kyphosis data frame.

5) Build a tree using the first 60 observations of kyphosis.

```
TREE <- rpart(Kyphosis ~ Age + Number + Start,
              data=kyphosis[1:60,],method="class")
```

6) Predict the kyphosis presence for the other 21 observations.

```
PR <- predict(TREE,kyphosis[61:81,],type='class')
```

7) Which is the misclassification rate (prediction error)

```
test <- kyphosis$Kyphosis[61:81]
table(PR,test)
```

```
##          test
## PR        absent present
##   absent      14       2
##   present      3       2
```

```
(rate <- 100*length(which(PR!=test))/length(PR))
```

```
## [1] 23.80952
```

Exercise rpart - iris

Consider the iris data frame

1) Build a tree to classify Species from the other variables.
2) Plot the trees, add nodes information.

Consider the tree build before

3) Prune the the using median complexity parameter (cp) associated to the tree.
4) Plot in the same window, the pruned and the original tree.
5) In which terminal nodes is clasified each oobservations of `iris`?
6) Which Specie has a flower of `Petal.Length` greater than 2.45 and `Petal.Width` less than 1.75.

Solution - rpart - iris (I)

1) Build a tree to classify Species from the other variables.

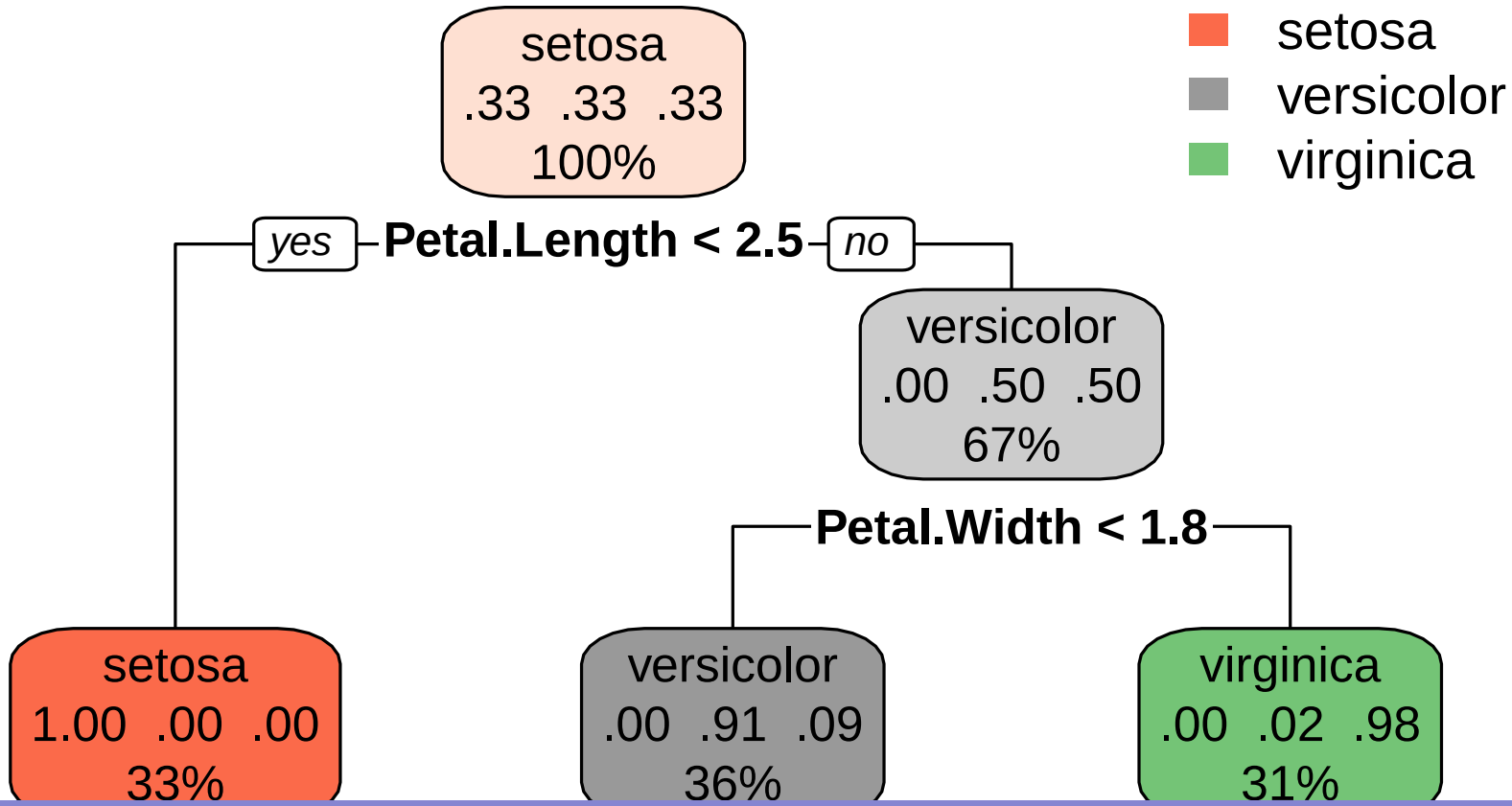
```
TREE2 <- rpart(Species ~ ., data=iris,method="class")
TREE2
```

```
## n= 150
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 150 100 setosa (0.33333333 0.33333333 0.33333333)
##   2) Petal.Length< 2.45 50   0 setosa (1.00000000 0.00000000
##   3) Petal.Length>=2.45 100  50 versicolor (0.00000000 0.5000
##     6) Petal.Width< 1.75 54   5 versicolor (0.00000000 0.9074
##     7) Petal.Width>=1.75 46   1 virginica (0.00000000 0.02173
```

Solution - rpart - iris (II)

2) Plot the trees, add nodes information.

```
library(rpart.plot)
rpart.plot(TREE2)
```



Solution - rpart - iris (III)

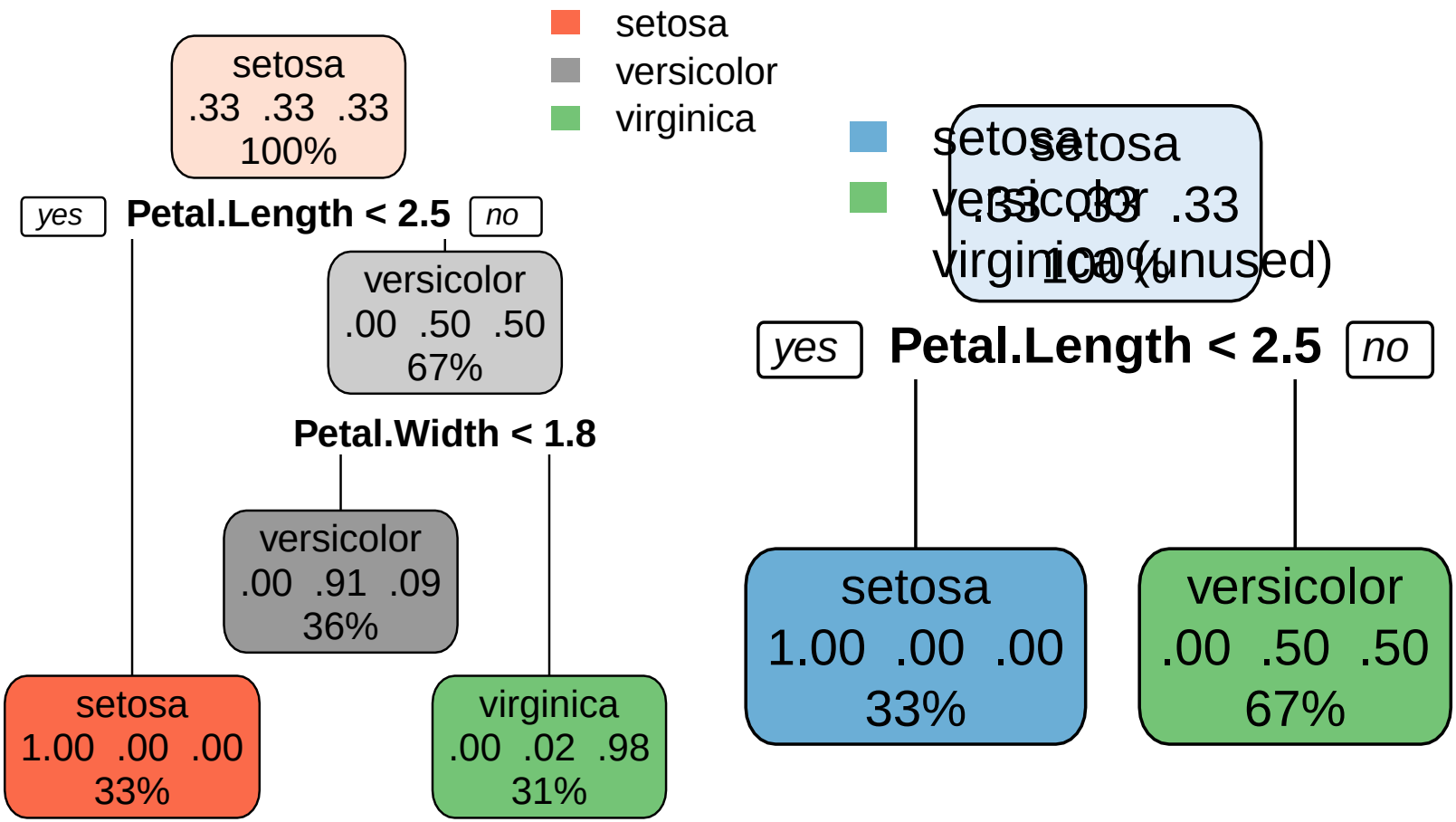
3) Prune the the using median complexity parameter (cp) associated to the tree.

```
TP <- prune(TREE2,cp=median(TREE2$cptable[,'CP']))
```

4) Plot in the same window, the pruned and the original tree.

```
par(mfrow=c(1,2))
rpart.plot(TREE2)
rpart.plot(TP)
```

The plotted results



Solution - rpart - iris (IV)

5) In which terminal nodes is clasified each observations of iris?

```
TREE2$where
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##   2   2   2   2   2   2   2   2   2   2   2   2   2   2   4
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69
##   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87
##   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
##  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 1
##   4   4   4   4   4   4   4   4   4   4   5   5   5   5   5
## 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 1
##   5   5   5   5   5   5   5   5   5   5   5   4   5   5   5
## 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 1
```

Solution - rpart - iris (V)

6) Which species has a flower of `Petal.Length` greater than 2.45 and `Petal.Width` less than 1.75.

```
print('versicolor')
```

```
## [1] "versicolor"
```

```
unique(iris[iris$Petal.Length>2.45 &
              iris$Petal.Width<1.75,"Species"])
```

```
## [1] versicolor virginica
## Levels: setosa versicolor virginica
```