

MACHINE LEARNING - DECISION TREES

EXERCISES/SOLUTION

Jan-Philipp Kolb

03 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 and 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.

```
data(kyphosis,package = "rpart")
```

```
dplyr::glimpse(kyphosis)
```

```
## Observations: 81
```

```
## Variables: 4
```

```
## $ Kyphosis <fct> absent, absent, present, absent, absent, abs
```

```
## $ Age      <int> 71, 158, 128, 2, 1, 1, 61, 37, 113, 59, 82,
```

```
## $ Number   <int> 3, 3, 4, 5, 4, 2, 2, 3, 2, 6, 5, 3, 5, 4, 3,
```

```
## $ Start    <int> 5, 14, 5, 1, 15, 16, 17, 16, 16, 12, 14, 16,
```

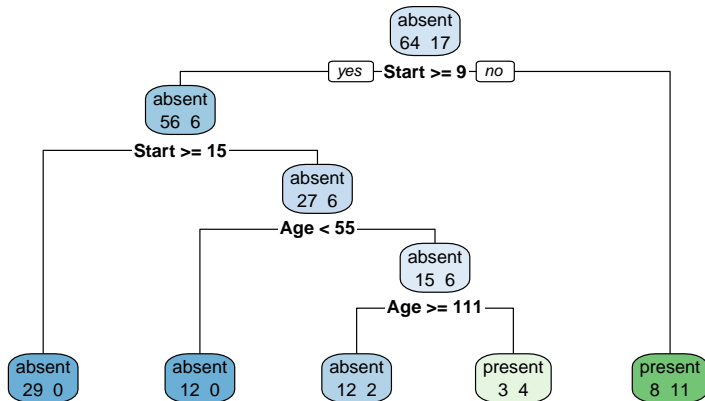
BUILD THE TREE (2)

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

## 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) *
##    3) Start< 8.5 19  8 present (0.42105263 0.57894737) *
```

PLOT THE RESULT

```
rpart.plot::rpart.plot(TREE,extra=1)
```



ANSWERS

3) Which variables are used to explain Kyphosis presence?

- ▶ The variables are Start and Age

4) How many observations contain the terminal nodes.

- ▶ The terminal nodes have 29, 12, 12, 3 and 8 observations

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

- ▶ `y` is a factor \Rightarrow we choose `method="class"`

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

FURTHER ANSWERS

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 tree 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 classified each observations of iris?
- 6) Which Species 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
```

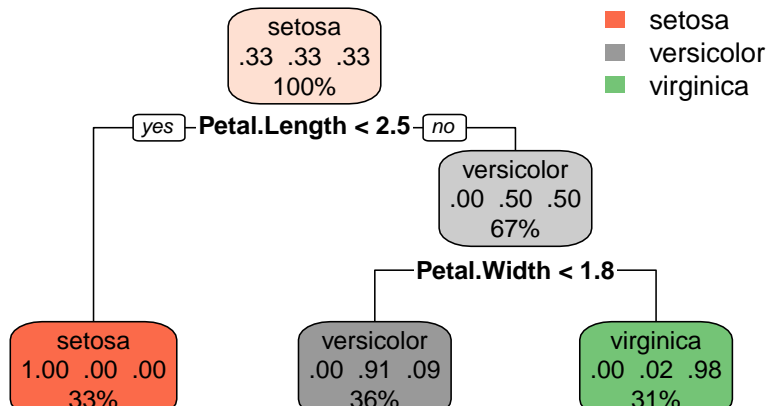
```
##     6) Petal.Width< 1.75 54    5 versicolor (0.00000000 0.9074
```

```
##     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)
```



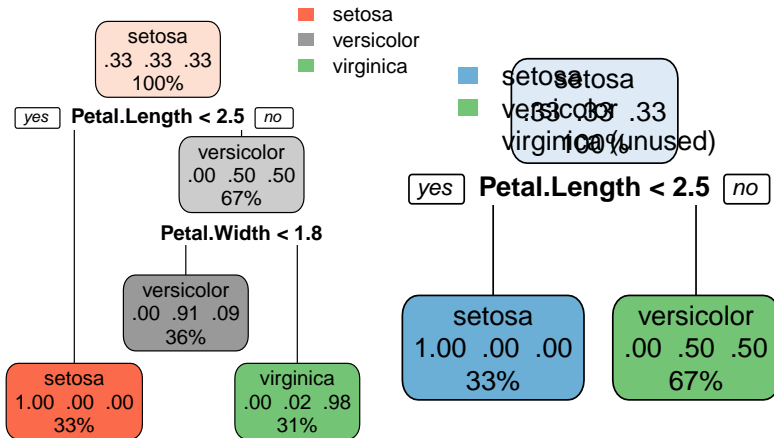
SOLUTION - RPART - IRIS (III)

- 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)
```



SOLUTION - RPART - IRIS (IV)

5) In which terminal nodes is classified 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
##	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
##	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

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
```