ch9-2

Harry Woo

2020 6 16

# Package Loading  
rm(list = ls())  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(rpart)  
library(rattle)

## Loading required package: tibble

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.  
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(rpart.plot)  
library(RColorBrewer)  
library(gt)  
  
# Data Loading  
data(iris)  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

# 전처리  
firis <- iris %>% rename(SL = Sepal.Length, SW = Sepal.Width,   
 PL = Petal.Length, PW = Petal.Width, SP = Species) %>%   
 mutate(SP = factor(SP, labels = c("st","vc","vg")))  
head(firis) %>% gt()

SL

SW

PL

PW

SP

5.1

3.5

1.4

0.2

st

4.9

3.0

1.4

0.2

st

4.7

3.2

1.3

0.2

st

4.6

3.1

1.5

0.2

st

5.0

3.6

1.4

0.2

st

5.4

3.9

1.7

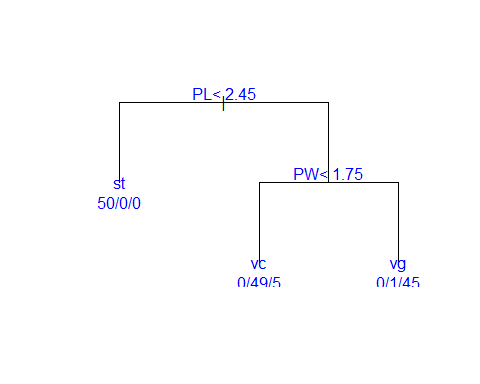
0.4

st

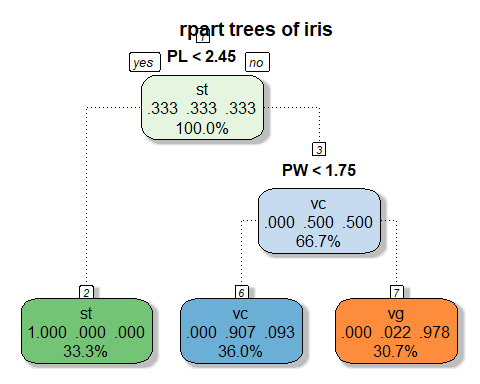
# rpart 함수를 이용하여 나무모형을 수행한 후, 나무모형의 도형화 결과를 보이고 결과를 해석하시오.   
# 나무모형을 이용하여 분류한 결과와 원래 그룹과의 분류표를 구하고, 오분류율을 구하시오.   
  
tree <- rpart(SP ~ ., data = firis)  
tree

## n= 150   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 150 100 st (0.33333333 0.33333333 0.33333333)   
## 2) PL< 2.45 50 0 st (1.00000000 0.00000000 0.00000000) \*  
## 3) PL>=2.45 100 50 vc (0.00000000 0.50000000 0.50000000)   
## 6) PW< 1.75 54 5 vc (0.00000000 0.90740741 0.09259259) \*  
## 7) PW>=1.75 46 1 vg (0.00000000 0.02173913 0.97826087) \*

plot(tree, uniform = TRUE, compress = TRUE, margin = 0.1)  
text(tree, use.n = TRUE, col = "blue")



fancyRpartPlot(tree, main = "rpart trees of iris \n", caption = "",   
 type = 1, digits = 3)



prediction <- predict(tree, newdata = firis, type = "class")  
  
confu\_mat <- table(firis$SP, prediction, dnn = c("Actual", "Predicted"))  
confu\_mat

## Predicted  
## Actual st vc vg  
## st 50 0 0  
## vc 0 49 1  
## vg 0 5 45

error <- 1 - sum(diag(confu\_mat)) / sum(confu\_mat)  
error

## [1] 0.04

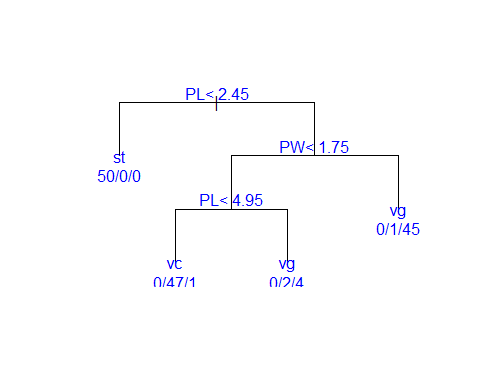
summary(tree)

## Call:  
## rpart(formula = SP ~ ., data = firis)  
## n= 150   
##   
## CP nsplit rel error xerror xstd  
## 1 0.50 0 1.00 1.18 0.05017303  
## 2 0.44 1 0.50 0.82 0.06096994  
## 3 0.01 2 0.06 0.07 0.02583280  
##   
## Variable importance  
## PW PL SL SW   
## 34 31 21 14   
##   
## Node number 1: 150 observations, complexity param=0.5  
## predicted class=st expected loss=0.6666667 P(node) =1  
## class counts: 50 50 50  
## probabilities: 0.333 0.333 0.333   
## left son=2 (50 obs) right son=3 (100 obs)  
## Primary splits:  
## PL < 2.45 to the left, improve=50.00000, (0 missing)  
## PW < 0.8 to the left, improve=50.00000, (0 missing)  
## SL < 5.45 to the left, improve=34.16405, (0 missing)  
## SW < 3.35 to the right, improve=19.03851, (0 missing)  
## Surrogate splits:  
## PW < 0.8 to the left, agree=1.000, adj=1.00, (0 split)  
## SL < 5.45 to the left, agree=0.920, adj=0.76, (0 split)  
## SW < 3.35 to the right, agree=0.833, adj=0.50, (0 split)  
##   
## Node number 2: 50 observations  
## predicted class=st expected loss=0 P(node) =0.3333333  
## class counts: 50 0 0  
## probabilities: 1.000 0.000 0.000   
##   
## Node number 3: 100 observations, complexity param=0.44  
## predicted class=vc expected loss=0.5 P(node) =0.6666667  
## class counts: 0 50 50  
## probabilities: 0.000 0.500 0.500   
## left son=6 (54 obs) right son=7 (46 obs)  
## Primary splits:  
## PW < 1.75 to the left, improve=38.969400, (0 missing)  
## PL < 4.75 to the left, improve=37.353540, (0 missing)  
## SL < 6.15 to the left, improve=10.686870, (0 missing)  
## SW < 2.45 to the left, improve= 3.555556, (0 missing)  
## Surrogate splits:  
## PL < 4.75 to the left, agree=0.91, adj=0.804, (0 split)  
## SL < 6.15 to the left, agree=0.73, adj=0.413, (0 split)  
## SW < 2.95 to the left, agree=0.67, adj=0.283, (0 split)  
##   
## Node number 6: 54 observations  
## predicted class=vc expected loss=0.09259259 P(node) =0.36  
## class counts: 0 49 5  
## probabilities: 0.000 0.907 0.093   
##   
## Node number 7: 46 observations  
## predicted class=vg expected loss=0.02173913 P(node) =0.3066667  
## class counts: 0 1 45  
## probabilities: 0.000 0.022 0.978

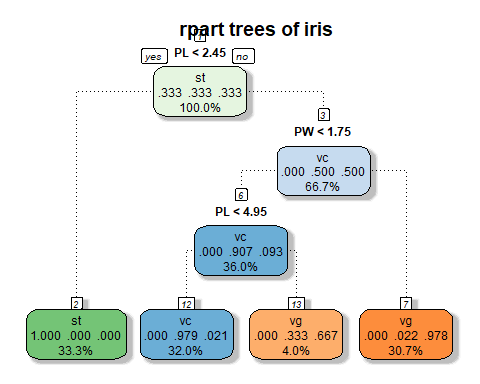
# 정지규칙에서 가지를 나누는 최소 자료의 수를 5로 하여 나무모형을 수행하고 앞의 (1)의 결과와 비교하시오  
  
my\_control <- rpart.control(minsplit = 5)  
tree2 <- rpart(SP ~ ., data = firis, method = "class", control = my\_control)  
tree2

## n= 150   
##   
## node), split, n, loss, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 150 100 st (0.33333333 0.33333333 0.33333333)   
## 2) PL< 2.45 50 0 st (1.00000000 0.00000000 0.00000000) \*  
## 3) PL>=2.45 100 50 vc (0.00000000 0.50000000 0.50000000)   
## 6) PW< 1.75 54 5 vc (0.00000000 0.90740741 0.09259259)   
## 12) PL< 4.95 48 1 vc (0.00000000 0.97916667 0.02083333) \*  
## 13) PL>=4.95 6 2 vg (0.00000000 0.33333333 0.66666667) \*  
## 7) PW>=1.75 46 1 vg (0.00000000 0.02173913 0.97826087) \*

plot(tree2, uniform = TRUE, compress = TRUE, margin = 0.1)  
text(tree2, use.n = TRUE, col = "blue")



fancyRpartPlot(tree2, main = "rpart trees of iris \n", caption = "",   
 type = 1, digits = 3)



prediction2 <- predict(tree2, newdata = firis, type = "class")  
  
confu\_mat2 <- table(firis$SP, prediction2, dnn = c("Actual", "Predicted"))  
confu\_mat2

## Predicted  
## Actual st vc vg  
## st 50 0 0  
## vc 0 47 3  
## vg 0 1 49

error2 <- 1 - sum(diag(confu\_mat2)) / sum(confu\_mat2)  
error2

## [1] 0.02666667

summary(tree2)

## Call:  
## rpart(formula = SP ~ ., data = firis, method = "class", control = my\_control)  
## n= 150   
##   
## CP nsplit rel error xerror xstd  
## 1 0.50 0 1.00 1.16 0.05127703  
## 2 0.44 1 0.50 0.62 0.06031031  
## 3 0.02 2 0.06 0.07 0.02583280  
## 4 0.01 3 0.04 0.10 0.03055050  
##   
## Variable importance  
## PW PL SL SW   
## 34 32 20 14   
##   
## Node number 1: 150 observations, complexity param=0.5  
## predicted class=st expected loss=0.6666667 P(node) =1  
## class counts: 50 50 50  
## probabilities: 0.333 0.333 0.333   
## left son=2 (50 obs) right son=3 (100 obs)  
## Primary splits:  
## PL < 2.45 to the left, improve=50.00000, (0 missing)  
## PW < 0.8 to the left, improve=50.00000, (0 missing)  
## SL < 5.45 to the left, improve=34.16405, (0 missing)  
## SW < 3.35 to the right, improve=19.03851, (0 missing)  
## Surrogate splits:  
## PW < 0.8 to the left, agree=1.000, adj=1.00, (0 split)  
## SL < 5.45 to the left, agree=0.920, adj=0.76, (0 split)  
## SW < 3.35 to the right, agree=0.833, adj=0.50, (0 split)  
##   
## Node number 2: 50 observations  
## predicted class=st expected loss=0 P(node) =0.3333333  
## class counts: 50 0 0  
## probabilities: 1.000 0.000 0.000   
##   
## Node number 3: 100 observations, complexity param=0.44  
## predicted class=vc expected loss=0.5 P(node) =0.6666667  
## class counts: 0 50 50  
## probabilities: 0.000 0.500 0.500   
## left son=6 (54 obs) right son=7 (46 obs)  
## Primary splits:  
## PW < 1.75 to the left, improve=38.969400, (0 missing)  
## PL < 4.75 to the left, improve=37.353540, (0 missing)  
## SL < 6.15 to the left, improve=10.686870, (0 missing)  
## SW < 2.45 to the left, improve= 3.555556, (0 missing)  
## Surrogate splits:  
## PL < 4.75 to the left, agree=0.91, adj=0.804, (0 split)  
## SL < 6.15 to the left, agree=0.73, adj=0.413, (0 split)  
## SW < 2.95 to the left, agree=0.67, adj=0.283, (0 split)  
##   
## Node number 6: 54 observations, complexity param=0.02  
## predicted class=vc expected loss=0.09259259 P(node) =0.36  
## class counts: 0 49 5  
## probabilities: 0.000 0.907 0.093   
## left son=12 (48 obs) right son=13 (6 obs)  
## Primary splits:  
## PL < 4.95 to the left, improve=4.4490740, (0 missing)  
## PW < 1.35 to the left, improve=0.9971510, (0 missing)  
## SL < 4.95 to the right, improve=0.6894587, (0 missing)  
## SW < 2.65 to the right, improve=0.2500139, (0 missing)  
##   
## Node number 7: 46 observations  
## predicted class=vg expected loss=0.02173913 P(node) =0.3066667  
## class counts: 0 1 45  
## probabilities: 0.000 0.022 0.978   
##   
## Node number 12: 48 observations  
## predicted class=vc expected loss=0.02083333 P(node) =0.32  
## class counts: 0 47 1  
## probabilities: 0.000 0.979 0.021   
##   
## Node number 13: 6 observations  
## predicted class=vg expected loss=0.3333333 P(node) =0.04  
## class counts: 0 2 4  
## probabilities: 0.000 0.333 0.667

# 오분류율이 개선되었다.