p8106_hw4_yg2625

Yue Gu

April 21, 2019

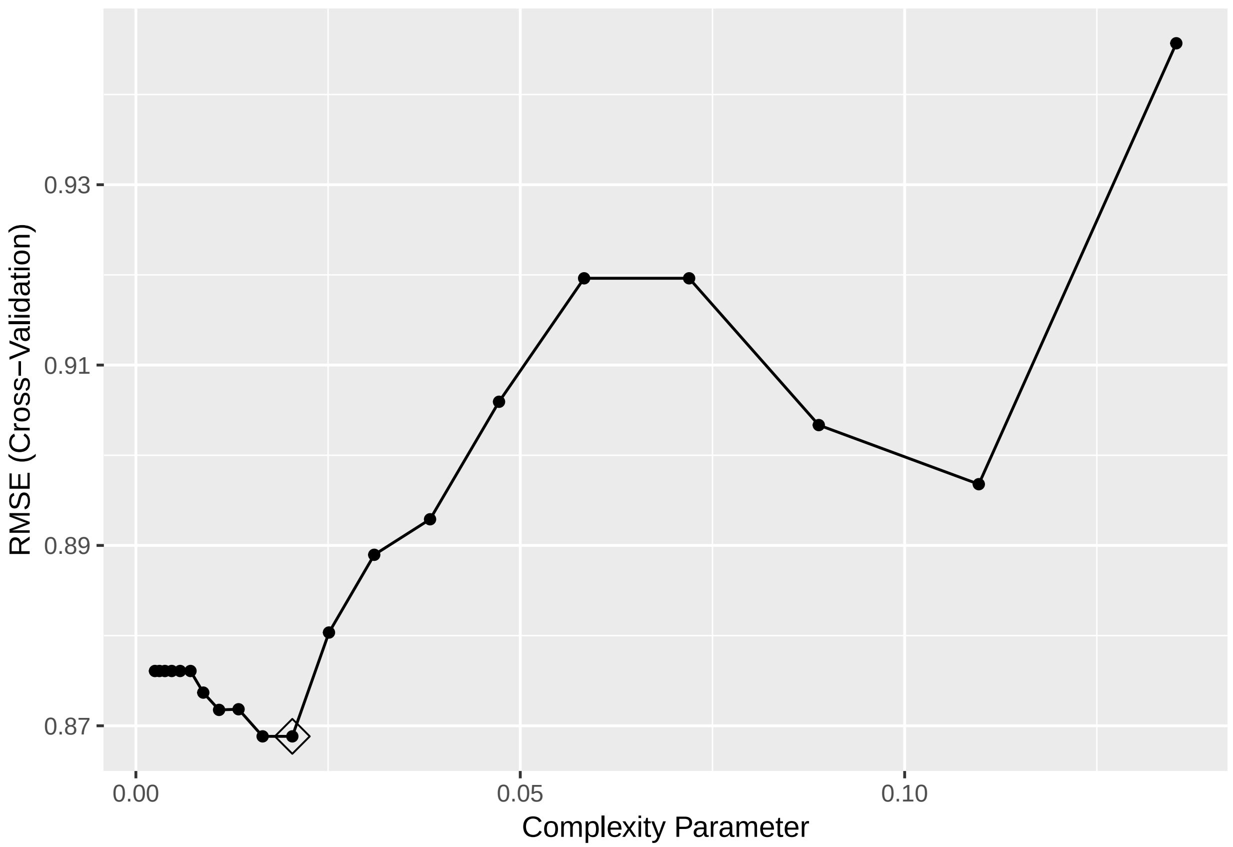
1. This problem involves the Prostate data in the lasso2 package (see L5.Rmd). Use set.seed() for reproducible results.

load data

```
data("Prostate")
pros_data = Prostate%>%
  janitor::clean_names()
```

(a) Fit a regression tree with lpsa as the response and the other variables as predictors. Use cross-validation to determine the optimal tree size. Which tree size corresponds to the lowest cross-validation error? Is this the same as the tree size obtained using the 1 SE rule?

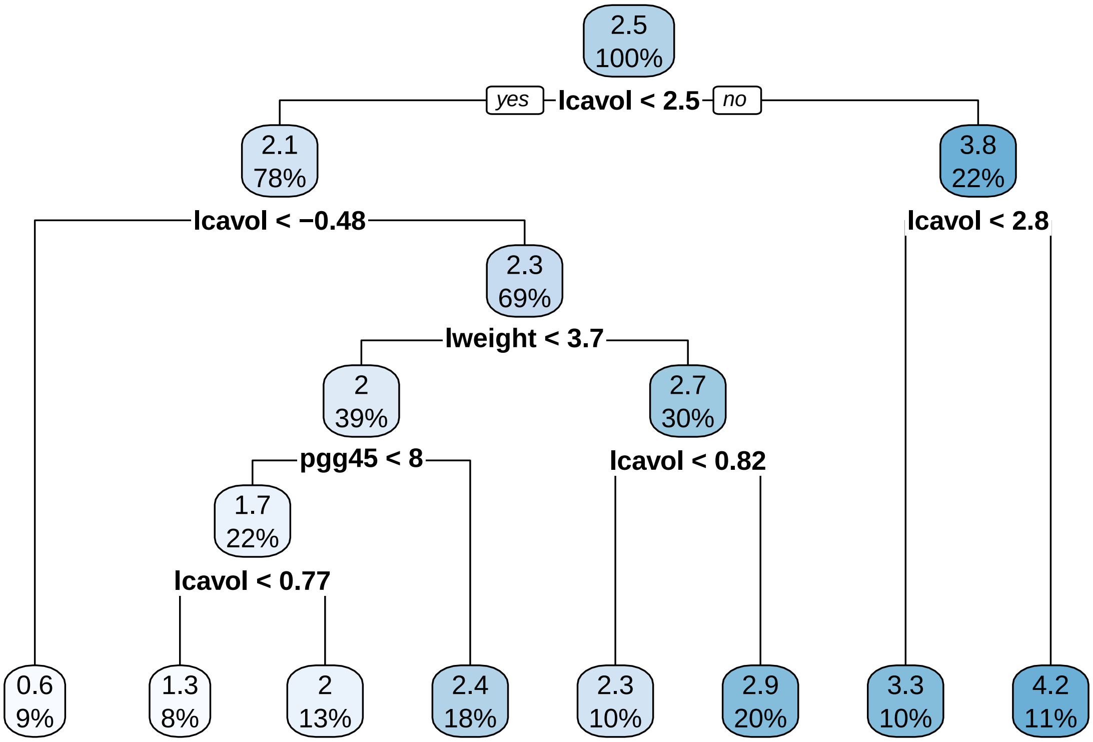
```
# use cross-validation through caret
ctrl <- trainControl(method = "cv")

# tune over cp, method = "rpart"
rpart.fit <- train(lpsa ~ ., pros_data,
                   method = "rpart",
                   tuneGrid = data.frame(cp = exp(seq(-6,-2, length = 20))),
                   trControl = ctrl)
ggplot(rpart.fit, highlight = TRUE)
```



```
# cptable showed that the optimal tree size is 8
rpart.fit$finalModel$cptable
```

```
##            CP nsplit rel error
## 1 0.34710828      0 1.0000000
## 2 0.18464743      1 0.6528917
## 3 0.05931585      2 0.4682443
## 4 0.03475635      3 0.4089284
## 5 0.03460901      4 0.3741721
## 6 0.02156368      5 0.3395631
## 7 0.02146995      6 0.3179994
## 8 0.00000000      7 0.2965295
```

```
rpart.plot(rpart.fit$finalModel)
```



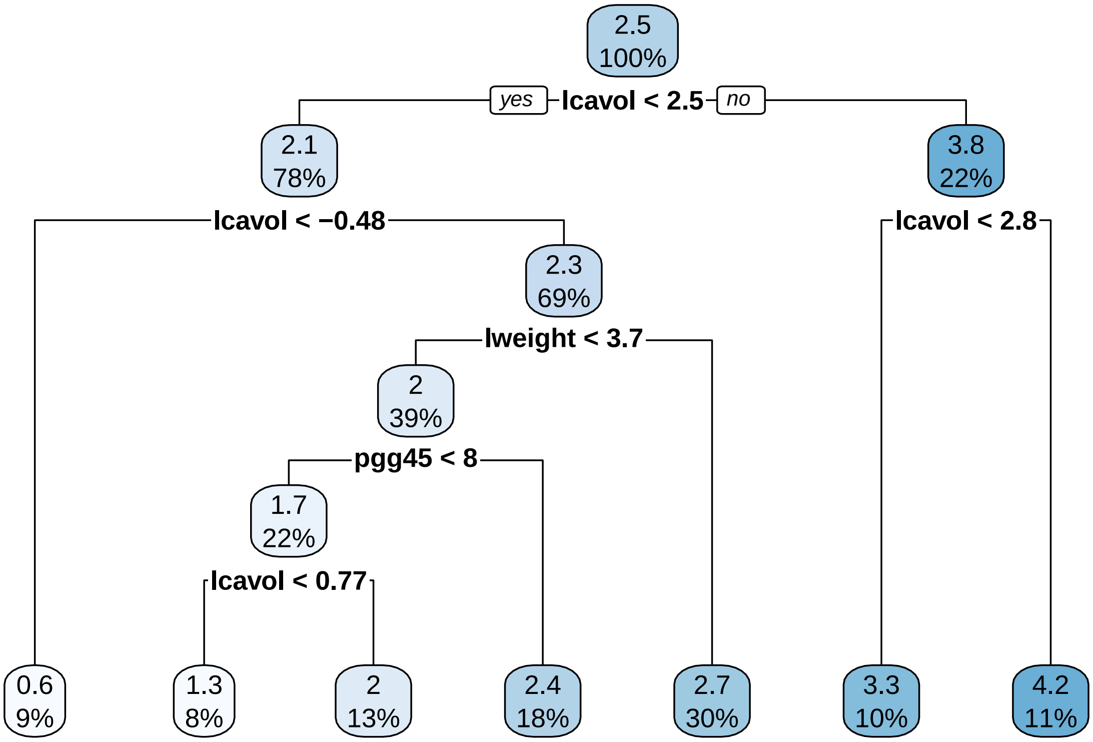
```r
# fit regression tree with default cp=0.01
tree1 = rpart(lpsa ~ ., pros_data)
# show cptable
cpTable = printcp(tree1)
```

```
##
## Regression tree:
## rpart(formula = lpsa ~ ., data = pros_data)
##
## Variables actually used in tree construction:
## [1] lcavol  lweight pgg45
##
## Root node error: 127.92/97 = 1.3187
##
## n= 97
##
##          CP nsplit rel error  xerror     xstd
## 1 0.347108      0   1.00000 1.04175 0.165040
## 2 0.184647      1   0.65289 0.82741 0.111422
## 3 0.059316      2   0.46824 0.62521 0.077455
## 4 0.034756      3   0.40893 0.59266 0.067608
## 5 0.034609      4   0.37417 0.58888 0.065364
## 6 0.021564      5   0.33956 0.57487 0.062840
## 7 0.021470      6   0.31800 0.57298 0.065930
## 8 0.010000      7   0.29653 0.59558 0.070062
```

```
# prune the tree based on cptable
minErr = which.min(cpTable[,4]);minErr
```

```
## 7
## 7
```

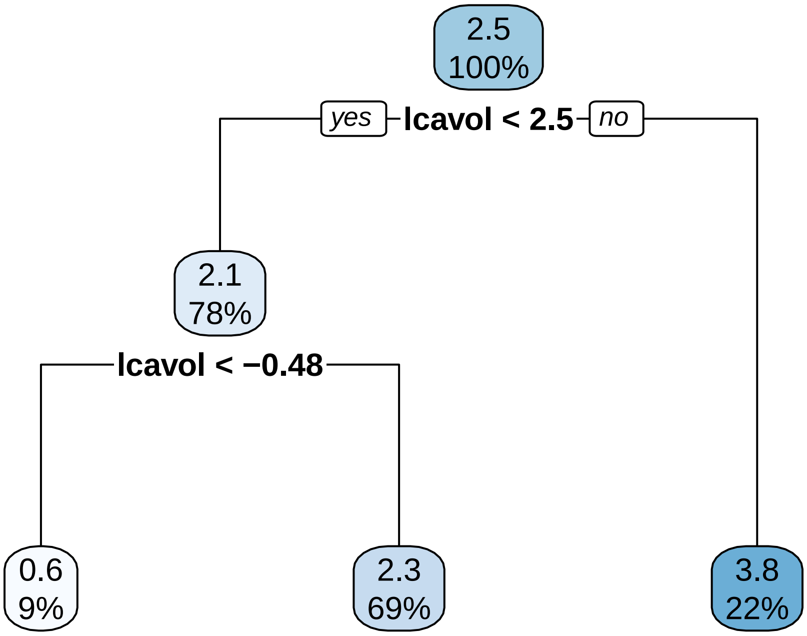
```
# minimum cross-validation error, use cp=8 with minimum CV error
tree2 <- prune(tree1, cp = cpTable[minErr,1])
rpart.plot(tree2)
```



```
# 1SE rule, use cp=4 with 1SE rule
min_1se = cpTable[cpTable[,4] < cpTable[minErr,4] + cpTable[minErr,5],1][1]; min_1se
```

```
##          3
## 0.05931585
```

```
tree3 <- prune(tree1, cp = min_1se)
rpart.plot(tree3)
```



Based on the result, cross-validation showed that the optimal tree size is 8 while 1SE obtained optimal tree size as 3. Hence, 1SE rule generates tree with smaller size.