ProblemSet8.R

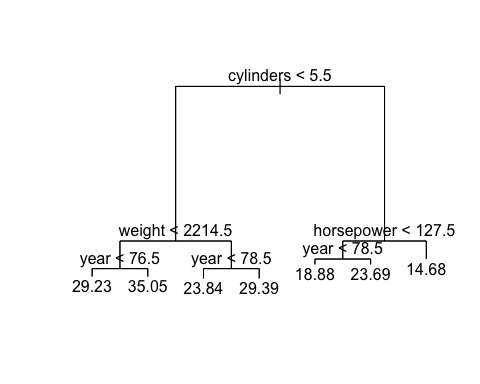
DavidAarhus

2020-05-08

# Name: David Aarhus  
# Problem Set 8  
  
rm(list = ls()) #removing all variables  
  
# load Auto dataset  
library(ISLR)  
data(Auto)  
  
#==Question a and b  
set.seed(310) #seed set  
indx <- sample(1:nrow(Auto), size = 0.70\*nrow(Auto),replace = FALSE)  
auto\_train <- Auto[indx,] #creates train set  
auto\_test <- Auto[-indx,] #creates test set  
  
#==Question c  
library("tree")  
  
#regression tree on mpg, using all variables except vehicle name  
tree <- tree(mpg ~ cylinders + displacement + horsepower + weight +  
 acceleration + year + origin,  
 data = auto\_train)  
  
#==Question d  
# plot the tree and add text to objects  
plot(tree) #plots tree  
text(tree) #labels tree  
  
#==Question e  
#Describe tree  
# There are alot of cars with less than 5.5 cylinders  
# Also there seems to be only 1 car that has horse power greater than 127.5 and is over 5.5 cylinders  
  
#==Question f  
preds\_test <- predict(tree, newdata = auto\_test) #test predictions  
preds\_train <- predict(tree, newdata = auto\_train) #train predictions  
  
# calculate RMSE  
library("caret")

## Loading required package: lattice

## Loading required package: ggplot2



RMSE(preds\_test, auto\_test$mpg) # -> 3.987852

## [1] 3.987852

RMSE(preds\_train, auto\_train$mpg) # -> 2.87646

## [1] 2.87646

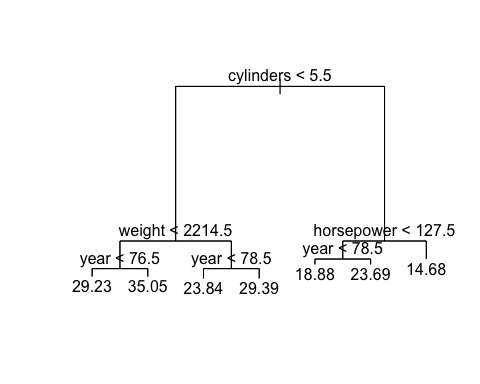
#==Question g  
cv.tree <- cv.tree(tree)  
cv.tree

## $size  
## [1] 7 6 5 4 3 2 1  
##   
## $dev  
## [1] 2873.978 3293.560 3913.297 4395.654 5504.904 7421.759 15633.827  
##   
## $k  
## [1] -Inf 313.8217 468.4140 574.5242 1059.1337 1624.1752 9090.8373  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

best\_tree\_indx <- which.min(cv.tree$dev) #finds best tree size  
cv.tree$size[best\_tree\_indx]

## [1] 7

# -> 7  
  
#==Question 1h  
pruned\_tree\_auto <- prune.tree(tree, best = 7) #prunes tree   
  
plot(pruned\_tree\_auto) #plots pruned tree  
text(pruned\_tree\_auto) #labels pruned tree



preds\_test\_pruned <- predict(pruned\_tree\_auto, newdata = auto\_test) #predictions from train tree  
preds\_train\_pruned <- predict(pruned\_tree\_auto, newdata = auto\_train) #predictions from test tree  
  
RMSE(preds\_test\_pruned, auto\_test$mpg) # 3.987852

## [1] 3.987852

RMSE(preds\_train\_pruned, auto\_train$mpg) # 2.87646

## [1] 2.87646

#==Question 1i  
# estimate a random forest  
library("randomForest")

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

bag <- randomForest(mpg ~ cylinders + displacement + horsepower + weight +  
 acceleration + year + origin,,  
 data = auto\_train,  
 mtry = 4,  
 ntree = 500,   
 importance = TRUE)  
  
# The mtry parameter is the number of variables available for splitting each node.   
# This comes into factor when you may have limits on memory and  
# computational limitations.   
  
#==Question 1j  
preds\_bag\_test <- predict(bag, newdata = auto\_test) #test predictions  
preds\_bag\_train <- predict(bag, newdata = auto\_train) #train predictions  
  
# RMSE  
RMSE(preds\_bag\_test, auto\_test$mpg) # -> 3.041922

## [1] 3.041922

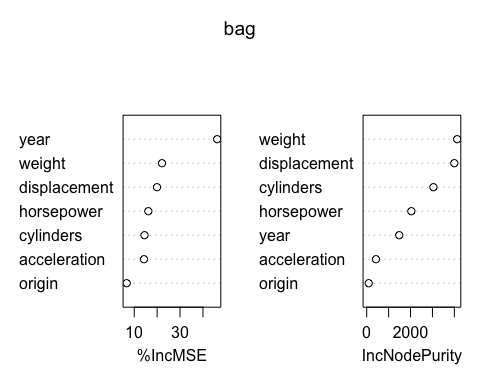
RMSE(preds\_bag\_train, auto\_train$mpg) # -> 1.269536

## [1] 1.269536

#==Question 1k  
importance(bag)

## %IncMSE IncNodePurity  
## cylinders 14.482331 3045.40382  
## displacement 19.942927 3998.19168  
## horsepower 16.158239 2037.04020  
## weight 22.114092 4128.74044  
## acceleration 14.260476 422.89089  
## year 46.212884 1485.36658  
## origin 6.710638 92.18818

varImpPlot(bag)



#==Question 1l  
# The tree and pruned tree models appeared to perform the same with similar RMSE scores  
# However, the randomForest model came out on top with the lowest RMSE score for both  
# the training set and test set. I think this is because since a randomForest is made  
# up of several trees, this allows the model to be more precise and accurate which is  
# shown in the RMSE scores.