**BU.510.650 Homework #5**

**Data Analytics Page 1 of 2**

**Spring 2021 The Johns Hopkins University**

**Arnab Bisi Carey Business School**

**Homework #5**

Due: 03/07/21, 11:59pm

1. For the Smarket data set from “ISLR” library, using the 2005 data as the test data and the remaining as the training data, what is your recommended value of K in the K-Nearest Neighbors (KNN) approach and why? (2 points)

199, because it has the highest accuracy

acc.list <- **c**()  
**for** (i **in** 1**:**(**nrow**(Smarket.train) **/** 2)) {  
 knn.pred <- **knn**(Smarket.train, Smarket.test, Direction.train, k = i)  
 acc.list[i] <- **mean**(knn.pred **==** Direction.test)  
}

**plot**(acc.list)

Chart, scatter chart

Description automatically generated

**which.max**(acc.list)

## [1] 119

acc.list[**which.max**(acc.list)]

## [1] 0.6111111

1. Using Carseats data set from “ISLR” library and installing “tree” library, run the R codes provided in the file “Carseats-DecisionTree-Example.R” to demonstrate decision tree method and submit your output here. (2 points)

[Carseats-DecisionTree-Example.docx](file:///C:\Users\duter\AppData\Roaming\Microsoft\Word\Carseats-DecisionTree-Example.docx)

1. In this question, you will use the K-Nearest Neighbors (KNN) algorithm to predict whether a passenger will survive or not.
2. Run the KNN algorithm to predict the response variable Survived for each passenger in the test data. Do this for K = 1, 3, and 5. According to these predictions for K = 1, 3, and 5, what is the proportion of passengers in the test data that will survive? (2 points)

0.3161435, 0.3049327, 0.3004484

Titanic.knn.1 <-  
 knn(  
 subset(Titanic.train, select = -Survived),  
 subset(Titanic.test, select = -Survived),  
 Titanic.train$Survived,  
 k = 1  
 )  
  
table(Titanic.knn.1, Titanic.test$Survived)

##   
## Titanic.knn.1 0 1  
## 0 248 57  
## 1 28 113

length(which(Titanic.knn.1 == 1)) / length(Titanic.knn.1)

## [1] 0.3161435

Titanic.knn.3 <-  
 knn(  
 subset(Titanic.train, select = -Survived),  
 subset(Titanic.test, select = -Survived),  
 Titanic.train$Survived,  
 k = 3  
 )  
  
table(Titanic.knn.3, Titanic.test$Survived)

##   
## Titanic.knn.3 0 1  
## 0 248 62  
## 1 28 108

length(which(Titanic.knn.3 == 1)) / length(Titanic.knn.3)

## [1] 0.3049327

Titanic.knn.5 <-  
 knn(  
 subset(Titanic.train, select = -Survived),  
 subset(Titanic.test, select = -Survived),  
 Titanic.train$Survived,  
 k = 5  
 )  
  
table(Titanic.knn.5, Titanic.test$Survived)

##   
## Titanic.knn.5 0 1  
## 0 248 64  
## 1 28 106

length(which(Titanic.knn.5 == 1)) / length(Titanic.knn.5)

## [1] 0.3004484

1. For each K, compute the accuracy of predictions for the test data. Which K works best in this case?

mean(Titanic.knn.1 == Titanic.test$Survived)

## [1] 0.809417

mean(Titanic.knn.3 == Titanic.test$Survived)

## [1] 0.7982063

mean(Titanic.knn.5 == Titanic.test$Survived)

## [1] 0.793722

(1 point)

1. In this question, you will estimate a decision tree for the AutoLoss data. The data file for this question, AutoLoss-DT.csv, is slightly different from the previous data file. In particular, instead of the actual loss amount for each vehicle, it has a column called HighLoss, which indicates whether the loss is high (“Yes”) or low (“No”) for each vehicle. Our goal is to create a decision tree that predicts whether the loss for a vehicle will be high or low.
2. Fit a decision tree to the entire data, with HighLoss as the response and all other variables as predictors. Plot the tree (including the names of predictors in the plot) and answer the following questions: Which predictors are used at the nodes of the tree? How many terminal nodes (leaves) does the tree have? (1 point)

"Height" "Weight" "Horsepower" "NumDoors" "Price" "Length" "EngineSize"

14 nodes

AutoLoss.tree <- tree(AutoLoss$HighLoss ~ ., AutoLoss)  
  
summary(AutoLoss.tree)

##   
## Classification tree:  
## tree(formula = AutoLoss$HighLoss ~ ., data = AutoLoss)  
## Variables actually used in tree construction:  
## [1] "Height" "Weight" "Horsepower" "NumDoors" "Price"   
## [6] "Length" "EngineSize"  
## Number of terminal nodes: 14   
## Residual mean deviance: 0.3839 = 57.21 / 149   
## Misclassification error rate: 0.09202 = 15 / 163

1. Determine the best tree size, using cross-validation and pruning. (See how we accomplished this in TASK 7 of Carseats example.) Plot the tree you obtained (including the names of predictors in the plot). (1 point)

AutoLoss.tree.cv <- cv.tree(AutoLoss.tree, FUN = prune.misclass)  
AutoLoss.tree.cv

## $size  
## [1] 14 11 8 6 4 2 1  
##   
## $dev  
## [1] 32 32 31 35 34 51 74  
##   
## $k  
## [1] -Inf 0.0 1.0 2.5 3.0 8.5 28.0  
##   
## $method  
## [1] "misclass"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

AutoLoss.tree.prune <-  
 prune.misclass(AutoLoss.tree, best = AutoLoss.tree.cv$size[which.min(AutoLoss.tree.cv$dev)])

plot(AutoLoss.tree.prune)  
text(AutoLoss.tree.prune, pretty = 0)

Diagram

Description automatically generated

1. Use the best tree to answer the following question (you do not need to use R for this): Suppose my car fits the description shown below. Will this car incur a high loss or not?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| FuelType | Aspiration | NumDoors | BodyStyle | DriveWheels | Length | Width | Height |
| gas | std | four | wagon | 4wd | 177 | 71 | 61 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Weight | EngineSize | Horsepower | PeakRPM | Citympg | Price |
| 3527 | 122 | 241 | 5000 | 22 | 25000 |

data = data.frame(  
 "gas",  
 "std",  
 "four",  
 "wagon",  
 "4wd",  
 177.00,  
 71.00,  
 61.00,  
 3527.00,  
 122.00,  
 241.00,  
 5000.00,  
 22.00,  
 25000.00,  
 ''  
)  
names(data) <- c(colnames(AutoLoss))  
AutoLoss <- rbind(AutoLoss, data)  
  
predict(AutoLoss.tree.prune,  
 AutoLoss[nrow(AutoLoss), ],  
 type = "class")

## [1] No

(1 point)