Quiz3

Jason Liu

Oct 6, 2016

Load Packages

```
library(caret)
library(AppliedPredictiveModeling)
library(rpart)
library(ElemStatLearn)
library(pgmm)
library(rpart.plot)
library(randomForest)
```

Question 1

We load the data based on the instructions on the quiz page.

```
library(AppliedPredictiveModeling)
data(segmentationOriginal)
library(caret)
```

*Instructions: 1. Subset the data to a training set and testing set based on the Case variable in the data set.

2. Set the seed to 125 and fit a CART model with the rpart method using all predictor variables and default caret settings.

3. In the final model what would be the final model prediction for cases with the following variable values:

a. TotalIntench2 = 23,000; FiberWidthCh1 = 10; PerimStatusCh1=2

b. TotalIntench2 = 50,000; FiberWidthCh1 = 10;VarIntenCh4 = 100

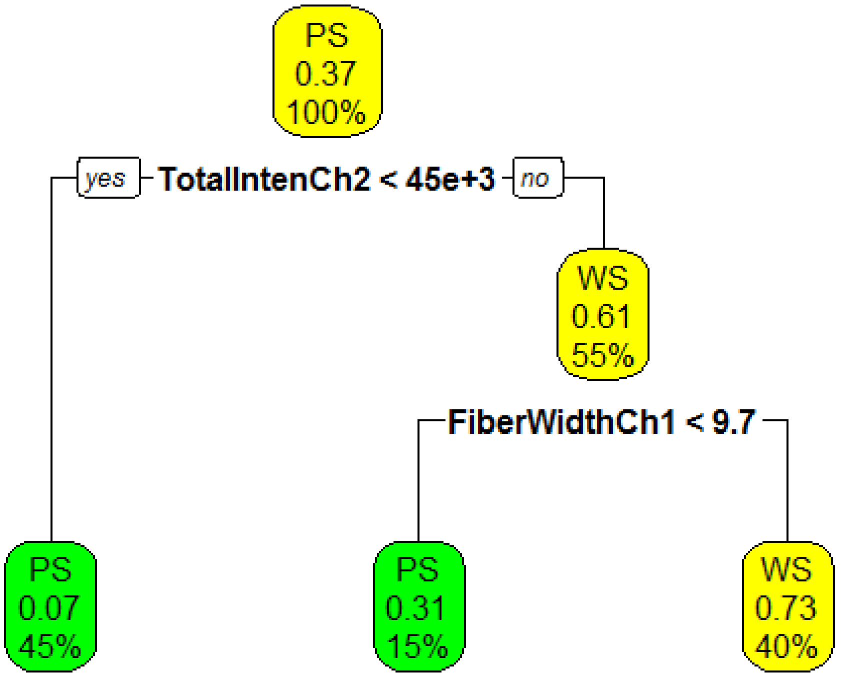
c. TotalIntench2 = 57,000; FiberWidthCh1 = 8;VarIntenCh4 = 100

d. FiberWidthCh1 = 8;VarIntenCh4 = 100; PerimStatusCh1=2

```
train<-segmentationOriginal[segmentationOriginal$Case=='Train',]
test <-segmentationOriginal[segmentationOriginal$Case=='Test',]
set.seed(125)
training<- train[,-c(1,2)]
modelfit<- train(training[,-1],training$Class,method='rpart')
print(modelfit$finalModel)
```

```
## n= 1009
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 1009 373 PS (0.63032706 0.36967294)
##   2) TotalIntenCh2< 45323.5 454  34 PS (0.92511013 0.07488987) *
##   3) TotalIntenCh2>=45323.5 555 216 WS (0.38918919 0.61081081)
##     6) FiberWidthCh1< 9.673245 154  47 PS (0.69480519 0.30519481) *
##     7) FiberWidthCh1>=9.673245 401 109 WS (0.27182045 0.72817955) *
```

```
rpart.plot(modelfit$finalModel,box.col=c("yellow", "green"))
```



Based on the rpart plot, we can find that:

A: PS B: ws *C: PS* D: Not explanable by this algorithm

Question 2

This is a question about the basic knowledge.

Rules:

Smaller K leads to greater bias but smaller variance, Larger K leads to smaller bias but larger variance. Under leave one out of class K should be equal to sample size.

Question 3

We load the data based on the instructions on the quiz page.

```
library(pgmm)
data(olive)
olive = olive[,-1]
```

Instructions: These data contain information on 572 different Italian olive oils from multiple regions in Italy. Fit a classification tree where Area is the outcome variable. Then predict the value of area for the following data frame using the tree command with all defaults.

```
fitmodel2<-train(olive[,-1],olive$Area,method='rpart')
newdata = as.data.frame(t(colMeans(olive)))
prediction2<- predict(fitmodel2,newdata = newdata)
```

In this case, the prediction result is a value with 2 decimals. Taken the average values of the predictors, the prediction result becomes strange as it does not belong to any existing class.

Question 4

We load the data based on the instructions on the quiz page.

```
library(ElemStatLearn)
data(SAheart)
set.seed(8484)
train = sample(1:dim(SAheart)[1],size=dim(SAheart)[1]/2,replace=F)
trainSA = SAheart[train,]
testSA = SAheart[-train,]
missClass = function(values,prediction){sum(((prediction > 0.5)*1) !=
values)/length(values)}
```

Instructions: Then set the seed to 13234 and fit a logistic regression model (method="glm", be sure to specify family="binomial") with Coronary Heart Disease (chd) as the outcome and age at onset, current alcohol consumption, obesity levels, cumulative tabacco, type-A behavior, and low density lipoprotein cholesterol as predictors.

```
set.seed(13234)
modelfit3<- train(chd ~ age + alcohol + obesity + tobacco + typea +
ldl,trainSA,method='glm',family='binomial')
Prediction3<- predict(modelfit3,newdata=testSA[,-10])
Prediction_In<- predict(modelfit3,newdata = trainSA[,-10])
Training_Missclass<-missClass(trainSA$chd,Prediction_In)
Test_Misclass<- missClass(testSA$chd,Prediction3)
Training_Missclass

## [1] 0.2727273
```

```
Test_Misclass

## [1] 0.3116883
```

The answers are provided above.

Question 5

We load the data based on the instructions on the quiz page.

```
library(ElemStatLearn)
data(vowel.train)
data(vowel.test)
```

Instructions: Set the variable y to be a factor variable in both the training and test set. Then set the seed to 33833. Fit a random forest predictor relating the factor variable y to the remaining variables. Read about variable importance in random forests here: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr The caret package uses by default the Gini importance.

```
set.seed(33833)
vowel.train$y<-as.factor(vowel.train$y)
vowel.test$y<- as.factor(vowel.test$y)
modelfit4<- randomForest(y~.,data=vowel.train)
order(varImp(modelfit4), decreasing = T)

##  [1]  2  1  5  6  8  4  9  3  7 10
```

The importance of the variables can be calculated by 'VarImp' in caret package and the answer is presented above.