Quiz 3

Question 1

For this quiz we will be using several R packages. R package versions change over time, the right answers have been checked using the following versions of the packages.

* AppliedPredictiveModeling: v1.1.6
* caret: v6.0.47
* ElemStatLearn: v2012.04-0
* pgmm: v1.1
* rpart: v4.1.8

If you aren’t using these versions of the packages, your answers may not exactly match the right answer, but hopefully should be close.

Load the cell segmentation data from the AppliedPredictiveModeling package using the commands:

**library**(AppliedPredictiveModeling)

data(segmentationOriginal)

**library**(caret)

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:
4. TotalIntench2 = 23,000; FiberWidthCh1 = 10; PerimStatusCh1=2
5. TotalIntench2 = 50,000; FiberWidthCh1 = 10;VarIntenCh4 = 100
6. TotalIntench2 = 57,000; FiberWidthCh1 = 8;VarIntenCh4 = 100
7. FiberWidthCh1 = 8;VarIntenCh4 = 100; PerimStatusCh1=2

Answer

subset <- split(segmentationOriginal, segmentationOriginal$Case)

set.seed(125)

modCART <- rpart(Class ~ ., data=subset$Train)

modCART

## 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)

## 4) IntenCoocASMCh3< 0.6021832 447 27 PS (0.93959732 0.06040268) \*

## 5) IntenCoocASMCh3>=0.6021832 7 0 WS (0.00000000 1.00000000) \*

## 3) TotalIntenCh2>=45323.5 555 216 WS (0.38918919 0.61081081)

## 6) FiberWidthCh1< 9.673245 154 47 PS (0.69480519 0.30519481)

## 12) AvgIntenCh1< 323.9243 139 33 PS (0.76258993 0.23741007) \*

## 13) AvgIntenCh1>=323.9243 15 1 WS (0.06666667 0.93333333) \*

## 7) FiberWidthCh1>=9.673245 401 109 WS (0.27182045 0.72817955)

## 14) ConvexHullAreaRatioCh1>=1.173618 63 26 PS (0.58730159 0.41269841)

## 28) VarIntenCh4>=172.0165 19 2 PS (0.89473684 0.10526316) \*

## 29) VarIntenCh4< 172.0165 44 20 WS (0.45454545 0.54545455)

## 58) KurtIntenCh3< 4.05017 24 8 PS (0.66666667 0.33333333)

## 116) Cell< 2.083968e+08 17 2 PS (0.88235294 0.11764706) \*

## 117) Cell>=2.083968e+08 7 1 WS (0.14285714 0.85714286) \*

## 59) KurtIntenCh3>=4.05017 20 4 WS (0.20000000 0.80000000) \*

## 15) ConvexHullAreaRatioCh1< 1.173618 338 72 WS (0.21301775 0.78698225) \*

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

testA <- segmentationOriginal[0,]

testA[1,c("TotalIntenCh2", "FiberWidthCh1", "PerimStatusCh1")] <- c(23000, 10, 2)

predict(modCART, testA, type="prob")

## PS WS

## 1 0.9395973 0.06040268

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

testB <- segmentationOriginal[0,]

testB[1,c("TotalIntenCh2", "FiberWidthCh1", "VarIntenCh4")] <- c(50000, 10, 100)

predict(modCART, testB, type="prob")

## PS WS

## 1 0.2130178 0.7869822

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

testC <- segmentationOriginal[0,]

testC[1,c("TotalIntenCh2", "FiberWidthCh1", "VarIntenCh4")] <- c(57000, 8, 100)

predict(modCART, testC, type="prob")

## PS WS

## 1 0.7625899 0.2374101

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

testD <- segmentationOriginal[0,]

testD[1,c("FiberWidthCh1", "VarIntenCh4","PerimStatusCh1")] <- c(8, 100, 2)

predict(modCART, testD, type="prob")

## PS WS

## 1 0.9395973 0.06040268

Best fitting answer: a. PS b. WS c. PS d. Not possible to predict

Question 2

If K is small in a K-fold cross validation is the bias in the estimate of out-of-sample (test set) accuracy smaller or bigger? If K is small is the variance in the estimate of out-of-sample (test set) accuracy smaller or bigger. Is K large or small in leave one out cross validation?

Answer

The bias is larger and the variance is smaller. Under leave one out cross validation K is equal to the sample size.

Question 3

Load the olive oil data using the commands:

**library**(pgmm)

data(olive)

olive = olive[,-1]

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

newdata = as.data.frame(t(colMeans(olive)))

What is the resulting prediction? Is the resulting prediction strange? Why or why not?

Answer

Since we’re not doing any testing, just use all the data :-)

set.seed(125)

modCART2 <- rpart(Area ~ ., data=olive)

modCART2

## n= 572

##

## node), split, n, deviance, yval

## \* denotes terminal node

##

## 1) root 572 3171.32000 4.599650

## 2) Eicosenoic>=6.5 323 176.82970 2.783282

## 4) Oleic>=7770.5 19 16.10526 1.315789 \*

## 5) Oleic< 7770.5 304 117.25000 2.875000 \*

## 3) Eicosenoic< 6.5 249 546.51410 6.955823

## 6) Linoleic>=1053.5 98 21.88776 5.336735 \*

## 7) Linoleic< 1053.5 151 100.99340 8.006623

## 14) Oleic< 7895 95 23.72632 7.515789 \*

## 15) Oleic>=7895 56 15.55357 8.839286 \*

predict(modCART2, newdata)

## 1

## 2.875

2.783 (or 2.875 in my case). It is strange because Area should be a qualitative variable - but tree is reporting the average value of Area as a numeric variable in the leaf predicted for newdata

Question 4

Load the South Africa Heart Disease Data and create training and test sets with the following code:

**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,]

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. Calculate the misclassification rate for your model using this function and a prediction on the “response” scale:

missClass = **function**(values,prediction){sum(((prediction > 0.5)\*1) != values)/length(values)}

What is the misclassification rate on the training set? What is the misclassification rate on the test set? ### Answer

set.seed(13234)

fit <- train(chd ~ age + alcohol + obesity + tobacco + typea + ldl, data=trainSA, method="glm", family="binomial")

predictTrainSA <- predict(fit)

missClass(trainSA$chd,predictTrainSA)

## [1] 0.2727273

predictTestSA <- predict(fit, testSA)

missClass(testSA$chd,predictTestSA)

## [1] 0.3116883

Question 5

Load the vowel.train and vowel.test data sets:

**library**(ElemStatLearn)

data(vowel.train)

data(vowel.test)

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.

Calculate the variable importance using the varImp function in the caret package. What is the order of variable importance?

Answer

vowel.train$y <- as.factor(vowel.train$y)

vowel.test$y <- as.factor(vowel.test$y)

*#set.seed(33833)*

modRF <- train(y ~ ., data=vowel.train, method="rf")

res <- predict(modRF,vowel.test)

varImp(modRF)

## rf variable importance

##

## Overall

## x.2 100.000

## x.1 96.575

## x.5 43.649

## x.6 31.117

## x.8 25.257

## x.4 13.137

## x.3 9.204

## x.9 9.071

## x.7 4.573

## x.10 0.000