Quiz 3

Aaron Brown

May 2, 2016

Question 1

```
#load the cell segmentation data
library(AppliedPredictiveModeling)
data(segmentationOriginal)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
require(rattle)
```

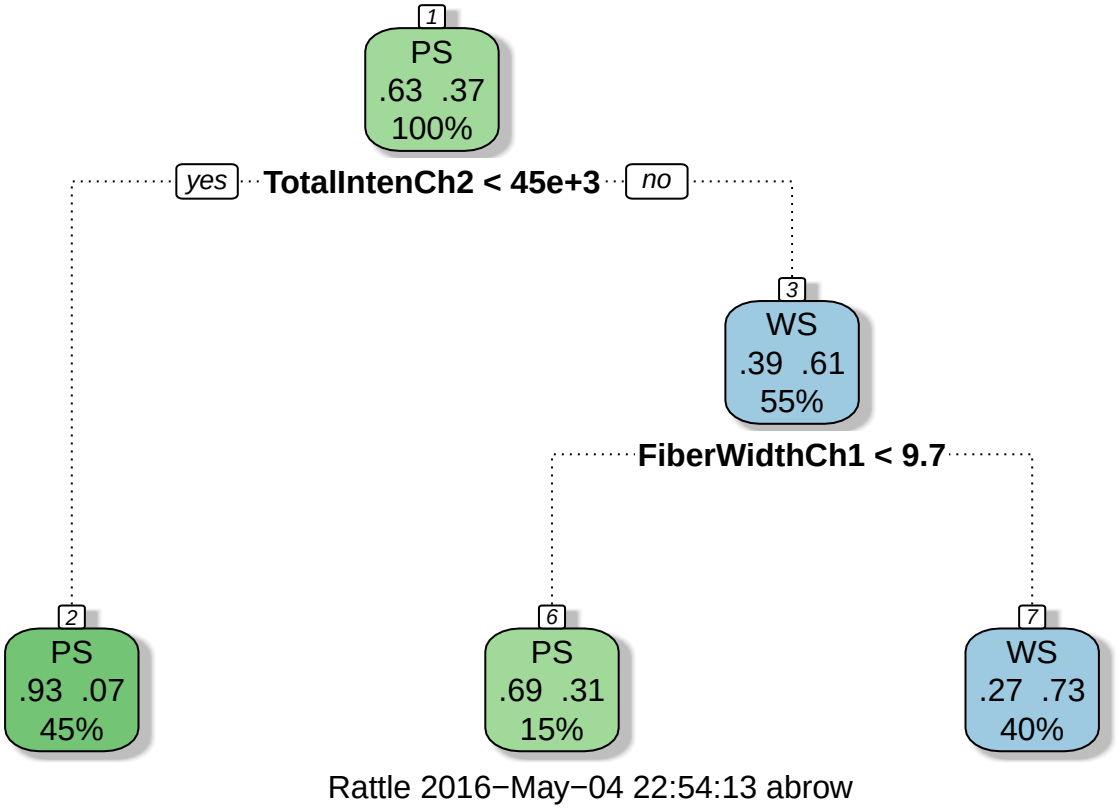
```
## Loading required package: rattle
```

```
## Rattle: A free graphical interface for data mining with R.
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
# require(rpart.plot)

set.seed(125)

split <- segmentationOriginal$Case == 'Train'

training = segmentationOriginal[split,]
testing = segmentationOriginal[!split,]

modelFit = train(Class ~ ., method = "rpart", data = training, control = rpart.control(maxdepth = 4L))
```

```
## Loading required package: rpart
```

```
fancyRpartPlot(modelFit$finalModel)
```



Answer:
A. PS
B. WS
C. PS
D. Not possible to predict

Question 2

Answer: (See Slide 8/8 on cross-validation in Week 1) The bias is larger and the variance is smaller. Under leave one out cross validation K is equal to the sample size.

Question 3

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

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

modelFit = train(Area ~ ., method = 'rpart', data = olive)
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```

```
a3 <- predict(modelFit, newdata)
```

Answer: 'r a3'

Question 4

```
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,]

set.seed(13234)

modelFit = train(chd ~ age + alcohol + obesity + tobacco + typea + ldl, method = 'glm',
                 family = 'binomial', data = trainSA)
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to
## do classification? If so, use a 2 level factor as your outcome column.
```

```
train.pred = predict(modelFit, trainSA)
test.pred  = predict(modelFit, testSA)

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

missClass(testSA$chd,  test.pred)
```

```
## [1] 0.3116883
```

```
missClass(trainSA$chd, train.pred)
```

```
## [1] 0.2727273
```

Answer:

Question 5

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

set.seed(33833)

vowel.train$y = as.factor(vowel.train$y)
vowel.test$y = as.factor(vowel.test$y)

modelFit = train(y ~ ., method = 'rf', data = vowel.train)
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-12
```

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

```
##
## Attaching package: 'randomForest'
```

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

```r
varImp(modelFit$finalModel)
```

```
##       Overall
## x.1  79.00581
## x.2  78.37953
## x.3  36.54356
## x.4  36.41000
## x.5  54.16992
## x.6  47.06475
## x.7  34.62493
## x.8  41.08484
## x.9  37.41012
## x.10 34.40001
```

Answer: According to the quiz, the correct answer of the order begins x.2, x.1,...

My answer begins with x.1, x.2,... and ends the same. I'm not sure why...