1 point

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 to predict Class 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

TIP: Plot the resulting tree and to use the plot to answer this question.

- a. PS
- b. Not possible to predict

c. PS

d. Not possible to predict

a. PS

b. WS

c. PS

d. Not possible to predict

a. PS

b. PS

c. PS

d. Not possible to predict

a. PS

b. WS

c. PS

d. WS

Quiz 3

Quiz, 5 questions

1 point

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?

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

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

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

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

---

1 point

3. Load the olive oil data using the commands:

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

(NOTE: If you have trouble installing the **pgmm** package, you can download the -code-olive-/code- dataset here: olive_data.zip. After unzipping the archive, you can load the file using the -code-load()-/code- function in R.)

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?

- 0.005291005 0 0.994709 0 0 0 0 0 0. The result is strange because Area is a numeric variable and we should get the average within each leaf.
- 4.59965. There is no reason why the result is strange.
- 2.783. 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
- 0.005291005 0 0.994709 0 0 0 0 0 0. There is no reason why the result is strange.

---

1 point

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?

- Test Set Misclassification: 0.31

  Training Set: 0.27

- Test Set Misclassification: 0.35

  Training Set: 0.31

- Test Set Misclassification: 0.27

  Training Set: 0.31

- Test Set Misclassification: 0.43

  Training Set: 0.31

---

1 point

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?

[NOTE: Use randomForest() specifically, not caret, as there's been some issues reported with that approach. 11/6/2016]

- The order of the variables is:

  x.10, x.7, x.9, x.5, x.8, x.4, x.6, x.3, x.1,x.2

- The order of the variables is:

  x.2, x.1, x.5, x.8, x.6, x.4, x.3, x.9, x.7,x.10

- The order of the variables is:

  x.1, x.2, x.3, x.8, x.6, x.4, x.5, x.9, x.7,x.10

- The order of the variables is:

  x.2, x.1, x.5, x.6, x.8, x.4, x.9, x.3, x.7,x.10

☐ I, **Ange Liu**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code

Submit Quiz