Practical Machine Learning Project (Coursera) - Quantified Data Analysis

Description:

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Data

The training data for this project are available here:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv

The test data are available here:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

The data for this project come from this source: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har. If you use the document you create for this class for any purpose please cite them as they have been very generous in allowing their data to be used for this kind of assignment.

Peer Review Portion

Your submission for the Peer Review portion should consist of a link to a Github repo with your R markdown and compiled HTML file describing your analysis. Please constrain the text of the writeup to < 2000 words and the number of figures to be less than 5. It will make it easier for the graders if you submit a repo with a gh-pages branch so the HTML page can be viewed online (and you always want to make it easy on graders.

Course Project Prediction Quiz Portion Apply your machine learning algorithm to the 20 test cases available in the test data above and submit your predictions in appropriate format to the Course Project Prediction Quiz for automated grading.

Environment Setup and Analysis

```
trainingURL <-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
testingURL <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
trainingFile <- "./data/pml-training.csv"
```

```
testingFile  <- "./data/pml-testing.csv"
if (!file.exists("./data")) {
  dir.create("./data")
}
if (!file.exists(trainingFile)) {
  download.file(trainingURL, destfile=trainingFile)
}
if (!file.exists(testingFile)) {
  download.file(testingURL, destfile=testingFile)
}
```

Data dimentions

```
trainRaw <- read.csv("./data/pml-training.csv")
testRaw <- read.csv("./data/pml-testing.csv")
dim(trainRaw)
```

```
## [1] 19622   160
```

```
dim(testRaw)
```

```
## [1]  20 160
```

Pre-processing

```
train <- trainRaw[, 6:ncol(trainRaw)]
```

Split the data into 80% training and 20% testing set

```
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
set.seed(23954)
inTrain <- createDataPartition(y = train$classe, p = 0.8, list = F)
training <- train[inTrain, ]
testing <- train[-inTrain, ]
```

Remove the NAs

```
nzv <- nearZeroVar(train, saveMetrics = T)
keepFeat <- row.names(nzv[nzv$nzv == FALSE, ])
training <- training[, keepFeat]
training <- training[, colSums(is.na(training)) == 0]
dim(training)
```

```
## [1] 15699    54
```

Model Building

```
controlRf <- trainControl(method="cv", 5)
modelRf <- train(classe ~ ., data=training, method="rf", trControl=controlRf, ntree=250)
modelRf
```

```
## Random Forest
##
## 15699 samples
##    53 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 12560, 12559, 12559, 12559, 12559
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9954138  0.9941987
##   27    0.9974522  0.9967773
##   53    0.9952864  0.9940370
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 27.
```

Prediction and the confusion matrix.

```
predRf <- predict(modelRf, newdata = testing)
confusionMatrix(
  factor(predRf),
  factor(testing$classe)
)$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 1116    1    0    0    0
##          B    0  758    1    0    0
##          C    0    0  683    1    0
##          D    0    0    0  642    1
##          E    0    0    0    0  720
```

Quiz answers

```
predRfTest <- predict(modelRf, newdata = testRaw)
predRfTest
```

```
##  [1] B A B A A E D B A A B C B A E E A B B B
## Levels: A B C D E
```