PracticalMachineLearning-001predictionMotivation

esdeewhy

July 21, 2015

BACKGROUND

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://groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

DATA PREPARATION AND PROCESSING

```
## Needed library
library(randomForest) # to build the model

## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.

library(caret)

## Loading required package: lattice
## Loading required package: ggplot2

testing <- read.csv("pml-testing.csv")
training <- read.csv("pml-training.csv")
dim(training)

## [1] 19622   160
```

CLEANING THE DATA

The data consists of 160 columns and 19622 rows. We will clean the data in order to get a data set of 52 possible predictors.

```
# Remove any nonquantitative variables
testingData <- training[-c(1:7)]
trainingtestingQtty <- testingData[, sapply(testingData, is.numeric)]
trainingtestingQtty$classe <- training$classe
```

```
# Removing all columsn with only NA values
trainingtestingQtty <- trainingtestingQtty[,!
colSums(is.na(trainingtestingQtty)) >= 19216]

dim(trainingtestingQtty)

## [1] 19622    53
```

MODELING

We will use Random forests in order to generate the prediction model from the training data set.

```
trainingSet <- trainingtestingQtty
# Random forest to find the classification model with training.train
data set.
Rmodel <- randomForest(classe ~ ., data=trainingSet, ntree=50)
print(Rmodel)

##
## Call:
##  randomForest(formula = classe ~ ., data = trainingSet, ntree = 50)
##                Type of random forest: classification
##                      Number of trees: 50
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 0.44%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 5578    0    0    1    1 0.0003584229
## B   14 3775    8    0    0 0.0057940479
## C    0   16 3401    5    0 0.0061367621
## D    0    1   30 3183    2 0.0102611940
## E    0    0    1    8 3598 0.0024951483
```

PREDICTION

The testing data set consits of 20 rows and 160 variables. The question is to know if we can accurately predict what exercise routine (variable classe) is being accomplished by a set of quantitative measurements. The testing data set does not contain the classe variable but will be predicted from the quantitative variables present in the data set.

The same variables used in the model generated from the training set must also be used to predict the exercise routine in the testing data sets. This means we must follow the same cleanup workflow on the testing data set.

```
# Remove any nonquantitative variables
testingSub <- testing[-c(1:7)]
testingtestingQtty <- testingSub[, sapply(testingSub, is.numeric)]
```

```
## Add back the Classe variable
testingtestingQtty$classe <- testing$classe

# Data cleanup

## Remove all columsn that are only NA values
testingtestingQtty <- testingtestingQtty[,!
colSums(is.na(testingtestingQtty)) >= 19216]

pred <- as.character(predict(Rmodel, testingtestingQtty))

prediction<- predict(Rmodel, testingtestingQtty)
print(prediction)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  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
```