Excercise Predition Model

jassalak

January 22, 2017

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, we use data from accelerometers on the belt, forearm, arm, and dumbbell of 6 participants [@velloso2013]. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. Given data from accelerometers, the goal is to predict the class of action which is one of the following.

- exactly according to the specification (A)
- throwing elbows to the front (B)
- lifting the dumbbell only halfway (C)
- lowering the dumbbell only halfway (D)
- throwing the hips to the front (E).

[This assignment is performed for Johns Hopkins University Data Science Specialization-Course#8-Week#4]

Environment Preparation

```
rm(list = ls())
#Removes anything in the Environment

setwd("C:/Users/akash/Desktop/StatsCourses/JHU_Specialization/Course8/w4")
#Sets the working directory

library(knitr)
opts_chunk$set(eval = TRUE, echo = TRUE, warning = FALSE,
               tidy = TRUE, results = "hold", cache = TRUE)
#Knitr global options

set.seed(2222)
#Sets the overall seed for reproducibility
```

Load necessary packages

```
library(lattice)
library(ggplot2)
library(caret)
library(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
```

```
library(rpart)
library(rpart.plot)
```

Data Processing

Upload the Training and Test datasets

```
TrainSet <- read.csv("C:/Users/akash/Desktop/StatsCourses/JHU_Specialization/Course8/w4/pml-training.cs
    na.strings = c("NA", "#DIV/0!", ""))
TestSet <- read.csv("C:/Users/akash/Desktop/StatsCourses/JHU_Specialization/Course8/w4/pml-testing.csv"
    na.strings = c("NA", "#DIV/0!", ""))
```

Delete columns with no values in them (both data sets)

```
TrainSet <- TrainSet[, colSums(is.na(TrainSet)) == 0]
TestSet <- TestSet[, colSums(is.na(TestSet)) == 0]
```

Remove irrelevant variables (both data sets)

user_name, raw_timestamp_part_1, raw_timestamp_part_,2 cvtd_timestamp, new_window, and num_window (columns 1 to 7)

```
TrainSet <- TrainSet[, -c(1:7)]
TestSet <- TestSet[, -c(1:7)]
```

Check for near zero variance (TrainSet)

```
nzv <- nearZeroVar(TrainSet, saveMetrics = TRUE)
if (any(nzv$nzv)) nzv else message("No variables with near zero variance")
```

```
## No variables with near zero variance
```

Split the TrainSet into SubTrainSet and SubTestSet; this allows for cross-validation (Train Set)

```
subsamples <- createDataPartition(y = TrainSet$classe, p = 0.6, list = FALSE)
SubTrainSet <- TrainSet[subsamples, ]
SubTestSet <- TrainSet[-subsamples, ]
```

Sample the data (SubTestSet)

```
head(SubTrainSet)
dim(SubTrainSet)
summary(SubTrainSet)
```

First Model (DecisionTree)

```
m1 <- rpart(classe ~ ., data = SubTrainSet, method = "class")
# DecisionTree Model creation

p1 <- predict(m1, SubTrainSet, type = "class")
# Predicting m1

rpart.plot(m1, main = "Classification Tree", extra = 102, under = TRUE, faclen = 0)
# Plotting m1
```

Second Model (randomForest)

```
m2 <- randomForest(classe ~ ., data = SubTrainSet, method = "class")
# RandomForest Model creation

p2 <- predict(m2, SubTestSet, type = "class")
# Predicting m1
```

Prediction on TestSet

After performing a Confusion Matrix on both Models, it was determined that Model2 (RandomForest) had a more accurate prediction (accuracy: 0.994). The out-of-sample error rate is 0.006, or 0.6%. This number is calculated by (1-(accuracy for predictions made against validation set)). With a greater than 99% accuracy on the validation set, and a Test Set of 20cases, we can expect few samples to be misclassified.

```
pfinal <- predict(m2, TestSet, type = "class")
pfinal
```

```
##  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
```

Submission

Creating .txt files of the TestDataSet answers

```
setwd("C:/Users/akash/Desktop/StatsCourses/JHU_Specialization/Course8/w4/answers")
pml_write_files = function(x) {
    n = length(x)
    for (i in 1:n) {
```

```
        filename = paste0("problem_id_", i, ".txt")
        write.table(x[i], file = filename, quote = FALSE, row.names = TRUE,
            col.names = TRUE)
    }
}

pml_write_files(pfinal)
```

References:

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013.

Read more: http://groupware.les.inf.puc-rio.br/har#ixzz4WcKtuSfy