

Machine Learning Assignment - Prediction Exercise

Bolaji Ogundipe

12/1/2019

Synopsis

This project work is part of exxercise to fulfill requirement of Machine Learning course part of coursera Data Science Specializaton program. It aim to predict the manner in which 6 participant, with data from accelerometers on the belt, forearm, arm, and dumbbell quantifies how well of a particular activity they do. This is the “classe” variable in the trainng set. The prediction model (algorithm) are applied to predict 20 different test cases available in the test data and submitted in appropriate format to the Coursera Project Prediction Quiz for automated grading.

Background to study

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 activity relatively inexpensively. These type of devices are part of the quantified self movement – a group of enthusiasts who take 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).

Exploratory Analysis of 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 is from this source: <http://web.archive.org/web/20161224072740/http://groupware.les.inf.puc-rio.br/har>.

But for the generosity of 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. There would'nt have been this data to perform this analysis. I thank you guys!

General overview of the data by the author;

Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: exactly according to the specification (Class A), throwing the elbows to the f(Class B), lifting the dumbbell only halfway (Class C), lowering the dumbbell only halfway (Class D) and throwing the hips to the front (Class E).

Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes. Participants were supervised by an experienced weight lifter to make sure the execution complied to the manner they were supposed to simulate. The exercises were performed by six male participants aged between 20-28 years, with little weight lifting experience. We made sure that all participants could easily simulate the mistakes in a safe and controlled manner by using a relatively light dumbbell (1.25kg).

Setting work environment

```
setwd("~/DScoursera/Machine Learning/ML project_work")
set.seed(12345)
library(caret)

## Loading required package: lattice
## Loading required package: ggplot2
library(corrplot)

## corrplot 0.84 loaded
library(rattle)

## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
library(randomForest)

## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:rattle':
##
##     importance
## The following object is masked from 'package:ggplot2':
##
##     margin
library(mice)

##
## Attaching package: 'mice'
## The following objects are masked from 'package:base':
##
##     cbind, rbind
library(knitr)
library(rpart.plot)

## Loading required package: rpart
library(rpart)
```

Loading and Cleaning Data

```
urlTrain <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
urlTest <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

training <- read.csv(url(urlTrain))
test <- read.csv(url(urlTest))
```

```
inTrain <- createDataPartition(training$classe, p=0.7, list = FALSE)
```

```
train <- training[inTrain, ]  
test <- training[-inTrain, ]
```

```
dim(train)
```

```
## [1] 13737 160
```

```
dim(test)
```

```
## [1] 5885 160
```

```
#removing almost zero variables
```

```
NZV <- nearZeroVar(train)
```

```
train <- train[, -NZV]
```

```
test <- test[, -NZV]
```

```
dim(train)
```

```
## [1] 13737 104
```

```
dim(test)
```

```
## [1] 5885 104
```

```
#removing Na values
```

```
NAs <- sapply(train, function(x) mean(is.na(x))) >0.95
```

```
train <- train[, NAs==FALSE]
```

```
test <- test[, NAs==FALSE]
```

```
dim(train)
```

```
## [1] 13737 59
```

```
dim(test)
```

```
## [1] 5885 59
```

```
#removing variables use for identification only
```

```
train <- train[, -(1:5)]
```

```
test <- test[, -(1:5)]
```

```
dim(train)
```

```
## [1] 13737 54
```

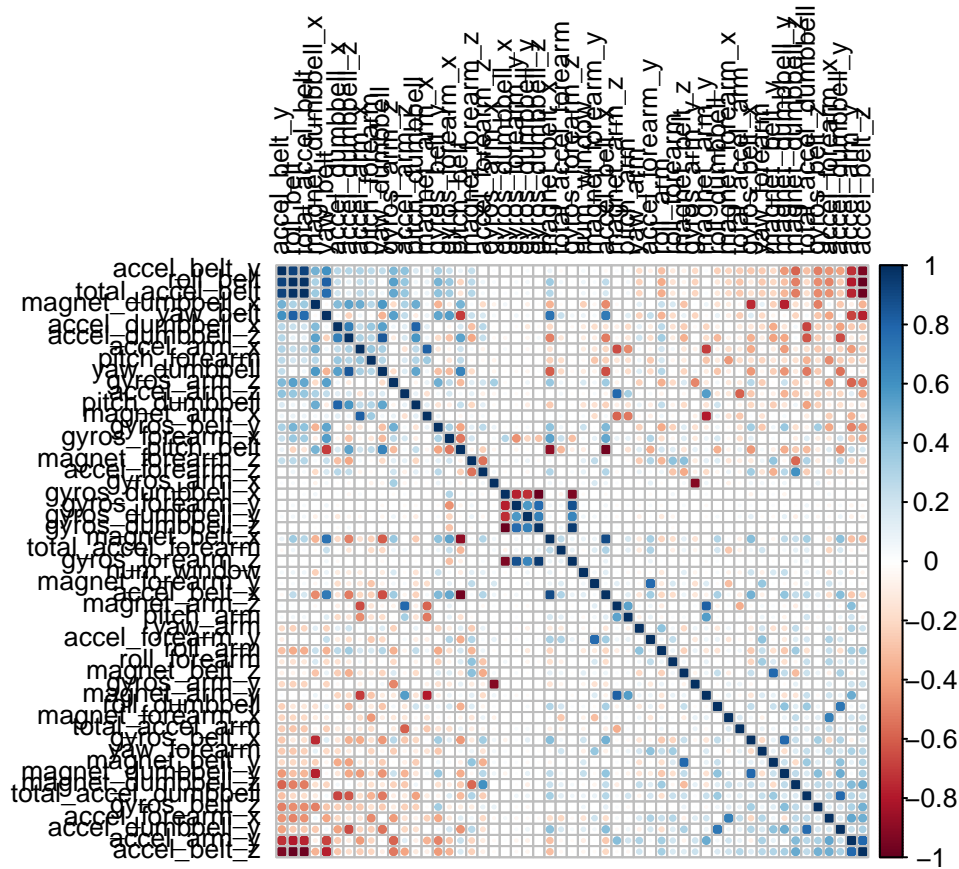
```
dim(test)
```

```
## [1] 5885 54
```

Cleaning procedure reduced the variables to 54. Below is correlation matrix of the variables. In the graph, dark-color areas shows the highly correlated variables. I would have carried out a PCA to further reduce the number of variables but it is few variables that are highly correlated in this case.

```
corMatrix <- cor(train[, -54])
```

```
corrplot(corMatrix, order = "FPC", method = "circle", type = "full", tl.cex = 0.8, tl.col = rgb(0, 0, 0,
```



Prediction Models Random Forest

```
set.seed(3245)
controlRF <- trainControl(method="cv", number=3, verboseIter=FALSE)

modfitRF <- train(classe ~., data=train, method = "rf", trControl = controlRF)

modfitRF$finalModel
```

```
##
## Call:
## randomForest(x = x, y = y, mtry = param$mtry)
##           Type of random forest: classification
##           Number of trees: 500
## No. of variables tried at each split: 27
##
##           OOB estimate of  error rate: 0.19%
## Confusion matrix:
##      A   B   C   D   E class.error
## A 3905   0   0   0   1 0.0002560164
## B   7 2647   3   1   0 0.0041384500
## C   0   4 2392   0   0 0.0016694491
## D   0   0   7 2245   0 0.0031083481
## E   0   0   0   3 2522 0.0011881188
```

```
#prediction on test data
predictRF <- predict(modfitRF, newdata = test)
confmat <- confusionMatrix(predictRF, test$classe)
```

```
confmat
```

```
## Confusion Matrix and Statistics
```

```
##
```

```
##           Reference
```

```
## Prediction    A    B    C    D    E
##           A 1674    1    0    0    0
##           B    0 1138    1    0    0
##           C    0    0 1025    1    0
##           D    0    0    0  963    2
##           E    0    0    0    0 1080
```

```
##
```

```
## Overall Statistics
```

```
##
```

```
##           Accuracy : 0.9992
```

```
##           95% CI : (0.998, 0.9997)
```

```
## No Information Rate : 0.2845
```

```
## P-Value [Acc > NIR] : < 2.2e-16
```

```
##
```

```
##           Kappa : 0.9989
```

```
##
```

```
## McNemar's Test P-Value : NA
```

```
##
```

```
## Statistics by Class:
```

```
##
```

```
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity      1.0000  0.9991  0.9990  0.9990  0.9982
## Specificity      0.9998  0.9998  0.9998  0.9996  1.0000
## Pos Pred Value   0.9994  0.9991  0.9990  0.9979  1.0000
## Neg Pred Value   1.0000  0.9998  0.9998  0.9998  0.9996
## Prevalence       0.2845  0.1935  0.1743  0.1638  0.1839
## Detection Rate   0.2845  0.1934  0.1742  0.1636  0.1835
## Detection Prevalence 0.2846  0.1935  0.1743  0.1640  0.1835
## Balanced Accuracy 0.9999  0.9995  0.9994  0.9993  0.9991
```

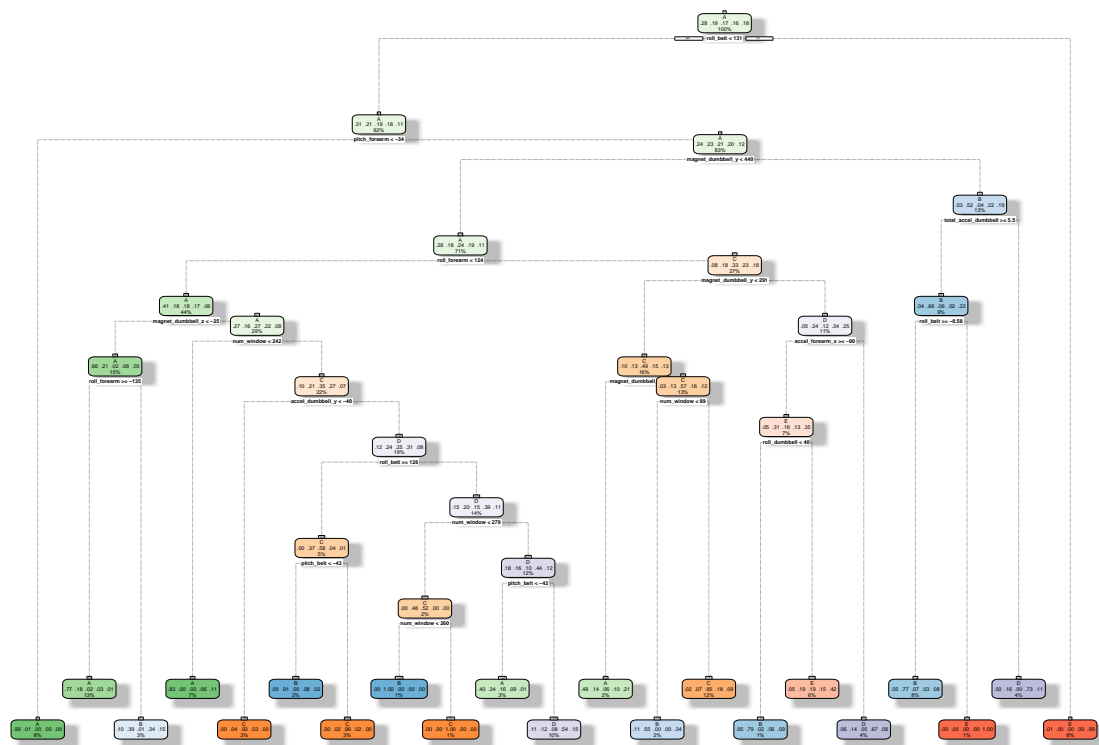
```
Decision Tree
```

```
set.seed(3245)
```

```
modfitDT <- rpart(classe ~., data = train, method = "class")
```

```
fancyRpartPlot(modfitDT)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



Rattle 2019-Dec-10 11:42:08 bogun

predict on test data

```
predictDT <- predict(modfitDT, newdata = test, type = "class")
confmatDT <- confusionMatrix(predictDT, test$classe)
confmatDT
```

Confusion Matrix and Statistics

##

Reference

Prediction	A	B	C	D	E
A	1502	201	59	66	74
B	58	660	37	64	114
C	4	66	815	129	72
D	90	148	54	648	126
E	20	64	61	57	696

##

Overall Statistics

##

```
## Accuracy : 0.7342
## 95% CI : (0.7228, 0.7455)
## No Information Rate : 0.2845
## P-Value [Acc > NIR] : < 2.2e-16
```

##

```
## Kappa : 0.6625
```

##

```
## McNemar's Test P-Value : < 2.2e-16
```

##

```
## Statistics by Class:
```

##	Class: A	Class: B	Class: C	Class: D	Class: E
## Sensitivity	0.8973	0.5795	0.7943	0.6722	0.6433
## Specificity	0.9050	0.9425	0.9442	0.9151	0.9579
## Pos Pred Value	0.7897	0.7074	0.7505	0.6079	0.7751
## Neg Pred Value	0.9568	0.9033	0.9560	0.9344	0.9226
## Prevalence	0.2845	0.1935	0.1743	0.1638	0.1839
## Detection Rate	0.2552	0.1121	0.1385	0.1101	0.1183
## Detection Prevalence	0.3232	0.1585	0.1845	0.1811	0.1526
## Balanced Accuracy	0.9011	0.7610	0.8693	0.7936	0.8006

apply model to data

```
predictTest <- predict(modfitRF, newdata = test)
predictTest
```

[illegible]

[illegible]

[illegible]

```
## [5440] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5477] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5514] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5551] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5588] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5625] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5662] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5699] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5736] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5773] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5810] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5847] E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E E
## [5884] E E
## Levels: A B C D E
```