Machine Learning Project

Ionas Kelepouris

8 March 2018

# I. Introduction

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>

# II. Background

The main goal of the project is to predict the manner in which 6 participants performed some exercise as described below. This is the “classe” variable in the training set. The machine learning algorithm described here is applied to the 20 test cases available in the test data and the predictions are submitted in appropriate format to the Course Project Prediction Quiz for automated grading.

# Data Loading and Exploratory Analysis

## a) Dataset

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 <http://groupware.les.inf.puc-rio.br/har>. Full source:

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.

## b) Used Libraries

First, we load the libraries and and we set the working direction.

library(knitr)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(rpart)  
library(rpart.plot)  
library(rattle)

## Rattle: A free graphical interface for data science with R.  
## Version 5.1.0 Copyright (c) 2006-2017 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

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:rattle':  
##   
## importance

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(corrplot)

## corrplot 0.84 loaded

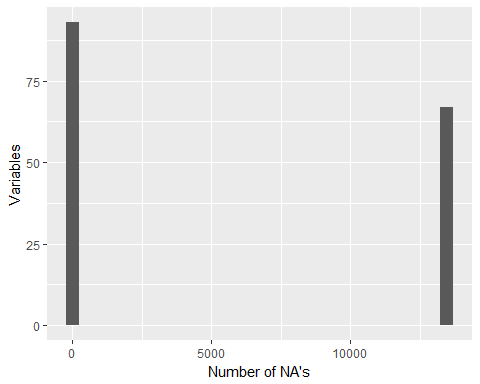
## c) Data Loading and Cleaning

# Download the datasets  
  
training = read.csv('pml-training.csv')  
testing = read.csv('pml-testing.csv')  
  
# Create Partitions from the training dataset  
set.seed(111)  
inTrain = createDataPartition(training$classe , p = 0.7 , list = FALSE)  
  
train\_set = training[inTrain , ]  
test\_set = training[-inTrain , ]

In our case 70% of the data available will be used in the *train\_set* variable and 30% in the *test set*. There are 13737 observations in our training set and 5885 in the test set and they both have 160 variables. There are a few variables with a lot of missing values that can be removed.

qplot(sapply(train\_set ,function(x) sum(is.na(x))) , xlab = "Number of NA's" , ylab = 'Variables')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



There are a lot of variables with 13449 missing values. So we exclude them.

most\_na = sapply(train\_set , function(x) sum(is.na(x))) > 13000  
train\_set = train\_set[ , most\_na == FALSE]  
test\_set = test\_set[ , most\_na == FALSE]

Now we have *93 variable*.