Practical learning machine - Final project

Manuela Escobar

### 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. In this project, our goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants, and the participants were asked to perform barbell lifts correctly and incorrectly in 5 different ways.

The goal of our project is to predict the manner in which they did the exercise. This is the “class” variable in the training set. We may use any of the other variables to predict with. We should create a report describing how we built our model, how we used cross validation, what we think the expected out of sample error is, and why we made the choices we did. We will also use our prediction model to predict 20 different test cases.

##### Download libraries and reading data

library(ggplot2)  
library(lattice)  
library(caret)  
library(rpart)  
library(RColorBrewer)  
library(rattle)

## Loading required package: tibble

## Loading required package: bitops

## Rattle: A free graphical interface for data science with R.  
## Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(e1071)  
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(mlbench)  
library(parallel)  
library(doParallel)

## Loading required package: foreach

## Loading required package: iterators

library(corrplot)

## corrplot 0.84 loaded

library(gbm)

## Loaded gbm 2.1.8

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:randomForest':  
##   
## combine

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

train <- read.csv('./data/pml-training.csv', header=T)  
validation <- read.csv('./data/pml-testing.csv', header=T)  
dim(train)  
dim(validation)

##### Cleaning data from NA values

traindb<- Filter(function(x) !any(is.na(x)), train)  
validationdb <- Filter(function(x) !any(is.na(x)), validation)  
traindb <- as.data.frame(traindb)  
validationdb <- as.data.frame(validationdb)  
dim(traindb)

## [1] 19622 93

dim(validationdb)

## [1] 20 60

##### Making train data base of the same length of validation base

classe <- traindb$classe  
trainRemove <- grepl("^X|timestamp|window", names(traindb))  
traindb <- traindb[, !trainRemove]  
trainCleaned <- traindb[, sapply(traindb, is.numeric)]  
trainCleaned$classe <- classe  
testRemove <- grepl("^X|timestamp|window", names(validationdb))  
validationdb <- validationdb[, !testRemove]  
validCleaned <- validationdb[, sapply(validationdb, is.numeric)]  
dim(trainCleaned)

## [1] 19622 53

dim(validCleaned)

## [1] 20 53

##### Creating test data set

traindb <- trainCleaned[createDataPartition(trainCleaned$classe, p=0.7, list=FALSE), ]  
testdb <- trainCleaned[-(createDataPartition(trainCleaned$classe, p=0.7, list=FALSE)), ]  
nrow(traindb)

## [1] 13737

nrow(testdb)

## [1] 5885

##### Trainning the model with random forest technique and decission trees

cluster <- makeCluster(detectCores() - 1)   
registerDoParallel(cluster)  
fitControl <- trainControl(method = "cv", number = 2, allowParallel = TRUE)  
system.time(modFit <- train(classe ~ ., method="rf",data=traindb,trControl = fitControl))

## user system elapsed   
## 50.015 0.857 118.908

stopCluster(cluster)  
registerDoSEQ()  
  
predict <- predict(modFit, newdata=testdb)  
  
vi <- varImp(modFit)$importance  
vi[head(order(unlist(vi), decreasing = TRUE), 5L), , drop = FALSE]

## Overall  
## roll\_belt 100.00000  
## pitch\_forearm 60.19416  
## yaw\_belt 55.78411  
## magnet\_dumbbell\_y 47.24668  
## pitch\_belt 45.76193

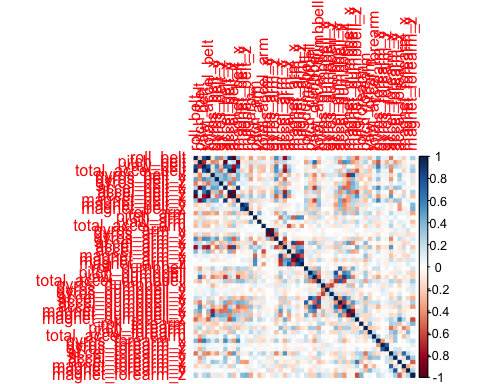
##### Predicting values for validation data

finalprediction <- predict(modFit, newdata=validCleaned)  
finalprediction

## [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

##### Figures

corrPlot <- cor(traindb[, -length(names(traindb))])  
corrplot(corrPlot, method="color")

 ##### Source of the data Velloso,Bulling, Gellersen,Ugulino, And Fuks. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI, Germany: ACM SIGCHI, 2013.