Practical Machine Learning Project

1. Introduction

The objective of this project is to apply the practical machine learning concepts and formulas learnt in the course to solve a particular prediction problem. In this case, the problem consists of predicting how well individuals perform an exercise, in particular, lifting a dumbbell. The individuals that participated in the experiment used sensors in their arm, forearm and belt and the dumbbell had a sensor as well. Therefore, by implementing a machine learning algorithm to this sensor data it is possible to predict how likely an individual will perform correctly or not an exercise without the need of a person to supervise it.

1. Data Analysis
   1. Data depuration

The first step in building the database to use for training is depurating the data. The initial data had 159 variables, however, many of these variables had empty cells, NA cells, and division by zero errors. As those variables had a very low variability, if none at all, they were deleted from the database. The final database comprised 55 variables, including the Id of the observation, the participant name, the classe variable which is the variable we are trying to predict, as well as the following variables for each fo the four sensors (arm, forearm, belt, dumbbell):

* Roll, pitch, yaw,
* Gyros in x,y,z axis
* Acceleration in x,y,z axis
* Magnet in x,y,z axis
* Total acceleration

The TimeStamp and the number-of-window/new-window variables were excluded from the database. Although the accumulated time that a person has been making an exercise may have an impact on how correctly it is being performed, because of increased expertise (long-term effect) or tiredness symptoms (short-term effect), these time-related effects can be approximated by the sensor data: a tired person will show more deviations from the correct execution-axis, etc. Also, it is difficult to relate these variables with the performance, since participants were asked to perform the exercise correctly and uncorrectly in different times and days, therefore including this variable could lead to biased estimations and overfitting.

It is worth noting that the Id of the observation and the participant name are included in the database, but are not used in the training process, because using them can lead to overfitting.

* 1. Exploratory Analysis

Figure 1 showas a preliminary scatter plot between the classe variable (on “y” axis) and several of the belt variables was made (see Figure 1). This graph was done with the purpose to show the entangled relationships that the data exhibits. The green dots show classe C, the pink classe B and the blue Classe A, and while the green dots are more concentrated, the blue ones (exercises done correctly) have more dispersion. This could mean that exercises done incorrectly tend to exhibit less variation in their movements. However, as the green and pink dots are a subset of the blue dots, this could difficult the ability of the model to predict incorrect movements.

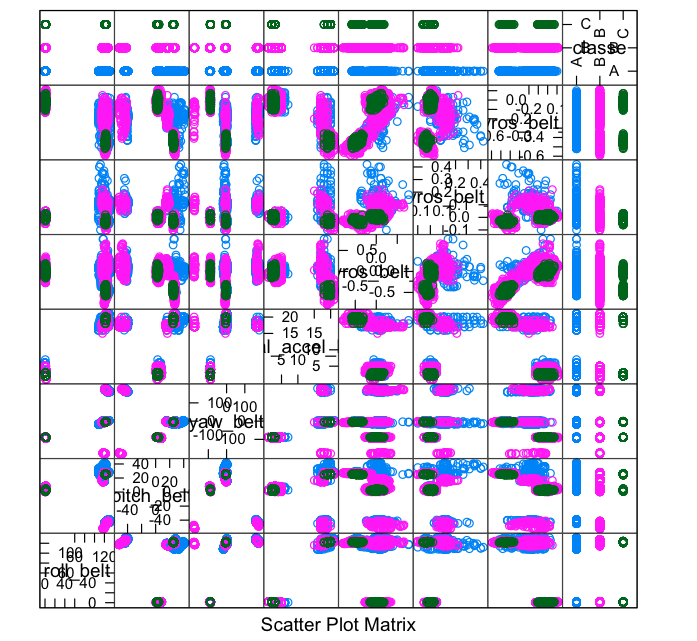


Figure 1: Scatter plot of “classe” and roll\_belt, pitch\_belt, yaw\_belt, total\_accel\_belt, gyros\_belt\_x, gyros\_belt\_y, gyros\_belt\_z (blue: classe A, pink: classe B, green: classe B)

1. Estimation
   1. Method selection

The training process consists of selecting the most adequate training method, setting the train control functions and whether to use or not of pre processing techniques.

The random forest method was used to train the data for the following reasons:

* It is a classification type of problem.
* There are 52 potential variables, and it is difficult to capture the relationships among them, since they are complex and hard to understand as was seen in the data analysis section. As the random forest method builds trees that differ both in the resampled data and the variables used, these combination of trees can capture more adequately the relationships between variables than bagging, although at higher computational time.

Regarding the train control functions, a 20 fold cross validation procedure was used. Initially a repeated cross validation was tried, but the computational times in estimating the random forest were so high that a simple cross validation was used preferred. 20 folds were chosen (twice the deault 10 value) because the data is large (more than 10,000 observations and 52 variables) and was considered that a good balance between estimators bias and variance could be achieved.

Finally, the data was not preprocessed with transformations or principal components analysis.

* 1. Performance indicators

The following figure shows a summary of the model fit. The optimal random forest is composed of two trees and the accuracy in the training set is 0.998. The accuracy is very high which can be indicative of overfitting. A ensemble between a random forest and boosting could be tested, as well as a more thorough procedure to eliminate variables from the model using as a basis the correlation between variables and the exploratory graphs. Also, the a principal component analysis could be useful to reduce the number of variables in the model, as from construction some of them are related (for instance, total acceleration with acceleration in x,y,z axis).

