Practical Machine Learning Course Peer Graded Assignment

Luis David Bedon Gomez

27 9 2017

1. Executive summary

The following report describes two machine learning models to predict the way in which exercises were done after the data collected by Ugulino et al. in "Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements".

Ugulino et al. consider 5 activity classes, gathered from 4 subjects wearing accelerometers mounted on four different parts of the body and provide a public domain dataset comprising 165,633 samples.

This dataset was loaded. After an exploratory data analysis, 52 covariates were selected to predict the activity classes described in the dataset.

A random-tree-model, a random-forest-model and a boosting model were applied strategically changing parameters and comparing the testing accuracy.

The random-forest-model and the boosting model trained over all 53 variables showed a very well prediction accuracy of 99,6%.

Based on this model, the predictions asked in the course were done.

2. Data Processing

2.1 Loading the Data

The datasets were loaded from the given URLs and imported in R using *read.csv*(). In this analysis, the libraries *caret* and *gbm* were used.

```
library(caret)
library(gbm)
library(knitr)
#library(ggplot2)
#library(dplyr)

# Download the data
setwd("~/Coursera/08PracticalMachineLearning")
if(file.exists("pml-training.csv")==FALSE){
  urlTrain<-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
  urlTest<-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
  download.file(urlTrain,"pml-training.csv")
  download.file(urlTest,"pml-testing.csv")
}

# Load the data
pml_training<-read.csv("pml-training.csv",stringsAsFactors = FALSE)
pml_testing<-read.csv("pml-testing.csv",stringsAsFactors = FALSE)
```

2.2 Exploratory Data Analysis

Examinating the data

The imported files consist in an training dataset *pml_training* with 19622 rows and 160 columns, and a quiz dataset *pml_testing* with 20 rows and the same number of columns.

	pml_training	pml_testing
No. of Measurements	19622	20
No. of Covariates	160	160

In order to avoid any type of overfitting, the dataset for the quiz will be not inspected.

The structure of *pml_training* shows diverse types of columns, mostly of class *int*, *num*, but also *chr*, some of them with a considerable amount of NA's:

```
## 'data.frame':    19622 obs. of  160 variables:
##  $ X                       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ user_name               : chr  "carlitos" "carlitos" "carlitos" "carlitos" ...
##  $ raw_timestamp_part_1    : int  1323084231 1323084231 1323084231 1323084232 1323084232 1323084232 1
##  $ raw_timestamp_part_2    : int  788290 808298 820366 120339 196328 304277 368296 440390 484323 484
##  $ cvtd_timestamp          : chr  "05/12/2011 11:23" "05/12/2011 11:23" "05/12/2011 11:23" "05/12/20
##  $ new_window              : chr  "no" "no" "no" "no" ...
##  $ num_window              : int  11 11 11 12 12 12 12 12 12 12 ...
##  $ roll_belt               : num  1.41 1.41 1.42 1.48 1.48 1.45 1.42 1.42 1.43 1.45 ...
##  $ pitch_belt              : num  8.07 8.07 8.07 8.05 8.07 8.06 8.09 8.13 8.16 8.17 ...
##  $ yaw_belt                : num  -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 -94.4 ...
##  $ total_accel_belt        : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ kurtosis_roll_belt      : chr  "" "" "" "" ...
##  $ kurtosis_picth_belt     : chr  "" "" "" "" ...
##  $ kurtosis_yaw_belt       : chr  "" "" "" "" ...
##  $ skewness_roll_belt      : chr  "" "" "" "" ...
##  $ skewness_roll_belt.1    : chr  "" "" "" "" ...
##  $ skewness_yaw_belt       : chr  "" "" "" "" ...
##  $ max_roll_belt           : num  NA NA NA NA NA NA NA NA NA NA ...
##   [list output truncated]
```

Column *classe*

Of special interest is the column *classe*, which contains the group factors and constitutes the outcome to predict. This column consists of characters "A" to "E" and contains no NA's:

```
##  chr [1:19622] "A" "A" "A" "A" "A" "A" "A" "A" "A" "A" ...
```

```
##
##    A    B    C    D    E
## 5580 3797 3422 3216 3607
```

2.3 Preselecting Covariates

As we saw above, not every column represents signal to be used for prediction. To select the right covariates, we first use the function $nearZeroVar()$ from the *caret*-package. It returns a vector with the index of the

near-zero columns and we create a new variable *pml_training1* eliminating the near-zero columns. This reduces the column-number by 60.

Parallel, we create also a new variable *pml_testing1* replicating the changes in the training data, but only without further examination of the testing set.

```
# Preprocessing and analizing possible covariates -> Class Video W206.mp4
## Near Zero Predictors
zeroCovar<-nearZeroVar(pml_training,saveMetrics = 0)
length(zeroCovar)
```

```
## [1] 60
```

```
pml_training1<-pml_training[,-zeroCovar]
pml_testing1<-pml_testing[,-zeroCovar]
```

Remaining columns containing NA's and strings are also removed, as well as columns with heading information, leading to a pair of variables *pml_trainig2* and *pml_testing2* with each 53 columns.

3. Prediction Models

In order to firm the knowledge gained in the course, 3 prediction models were used:

- a CART model,
- a Random Tree model and
- a Boosting model.

For every model several parameters were changed and the accuracy of the predictions to the testing data *pml_testing*2 registered, among with the processing time to train the model.

The results can be seen in the following table. The code used for each model can be found in the Appendix.

```
\begin{table}
\begin{tabular}{|c|c|c|c|c|c|}

Prediction Model & Subpartition   &   No. of covariates & Additional & Time & Accuracy vs.\\ \hline
                 &  pml_training2 &                    & Parameters &  & pml_testing2 \\ \hline
                 &  in train/test &    &           &     &  \\ \hline
CART             &      100/0    & 3  &       3   &   &   \\ \hline
CART             &      75/25    & 3  &       3   &   & \\ \hline
CART             &      75/25    & 52  &      53   &   & \\ \hline
Random Tree      &      75/25    & 3  &   ntree=10 & 17s  & 50.6%  \\ \hline
Random Tree      &      75/25    & 52  &   ntree=10 & 2.11min  & 99.6%  \\ \hline
Boosting         &      75/25    & 52  &    shrinkage=0.01   &   xs   & \\
                 &               &     &    ntree=100        &       & \\ \hline
Boosting         &      75/25    & 52  &    shrinkage 0.7   &   xs   & 99.55\% \\ \hline
\end{tabular}
\end{table}
```

4. Results

From the results in the table presented in section 3 we corroborate several concepts tough in class:

- It is extremely important to avoid overfitting by making subpartitions of the training data. We see this in the first CART model with the overfitted data and an accuracy of only ... %. Choosing the data partition improved the final accuracy to ... %
- The predictions rely on the quality of the collected data and small amounts of data can not be compensated by better algorithms. The prediction accuracy is enormously better taking the 53 columns than only a few of them.
- Both Random Forest and Boosting reach a very high accuracy of 99.x% and 99.y% respectively. The training time for the RF-method was 2.11min compared to only xs for boosting.

The prediction for the quiz was doing two times, one with the RF-model and the Boosting-model. Both methods give the following results:

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Prediction	B	A	B	A	A	E	D	B	A	A	B	C	B	A	E	E	A	B	B	B

Appendix

A1