Practical Machine Learning project

Igor Laberov

Friday, April 24, 2015

Overview

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://groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

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

Exploratory Data analysis

- a lot of almost empty columns
- Correlation analysis has shown that there are no correlation between single columns and classe (maximum cor-value=0.08). See also Figures 1 and 2.
- There are columns that are strongly correlated (>0.8)

Data cleaning

- Columns with empty data were excluded
- New column 'userCode' was added

Algorithm choice

Several approaches were probed, they gave different values of Accuracy. At the end was selected the method with the best results.

-Linear regression - 42%

-PCA - 67%

-Tree partition(caret) - 96.9%

-Tree partition(rpart) - 92.6%

-Random forests - 94%

As a result,Tree partition method from caret package was chosen. In order to achieve the best results, mutually correlated columns were replaced by one representative column (See Figure 3). As a result number of significant columns was decreased from 55 to 39. Rpart was run for size-120 length and produced 96.9% accuracy. See interval and statistical error below.

Cross-validation

Cross-validation was done on 20-fold basis, while choosing the fit model with the best Accuracy results. The set that was provided for training purposes was splitted each time in 75/25(training/test) proportion.

```
## Loading required package: lattice
## Loading required package: ggplot2
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## [1] "Predicted classe for testing set:"
```

```
##  [1] "C" "A" "A" "A" "A" "E" "D" "B" "A" "A" "B" "C" "B" "A" "E" "E" "A"
## [18] "B" "B" "B"
```

Prediction model

These is a "winning" model and its characteristics. Eventually, this model provided 100% accuracy on 20-records test set.

```
cm
```

```
## [[1]]
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3    4    5
##          1 1364   20    7    0    4
##          2   34  860   47    6    2
##          3    3   18  814   19    1
##          4    0    0   21  777    6
##          5    0    3   12   31  855
##
## Overall Statistics
##
##                Accuracy : 0.9523
##                  95% CI : (0.9459, 0.9581)
##     No Information Rate : 0.2857
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9397
##  Mcnemar's Test P-Value : NA
```

```
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity            0.9736   0.9545   0.9034   0.9328   0.9850
## Specificity            0.9912   0.9778   0.9898   0.9934   0.9886
## Pos Pred Value         0.9778   0.9062   0.9520   0.9664   0.9489
## Neg Pred Value         0.9895   0.9896   0.9785   0.9863   0.9968
## Prevalence             0.2857   0.1837   0.1837   0.1699   0.1770
## Detection Rate         0.2781   0.1754   0.1660   0.1584   0.1743
## Detection Prevalence   0.2845   0.1935   0.1743   0.1639   0.1837
## Balanced Accuracy      0.9824   0.9661   0.9466   0.9631   0.9868
```

Figures

Figure 1 - Classe vs Total accel Belt

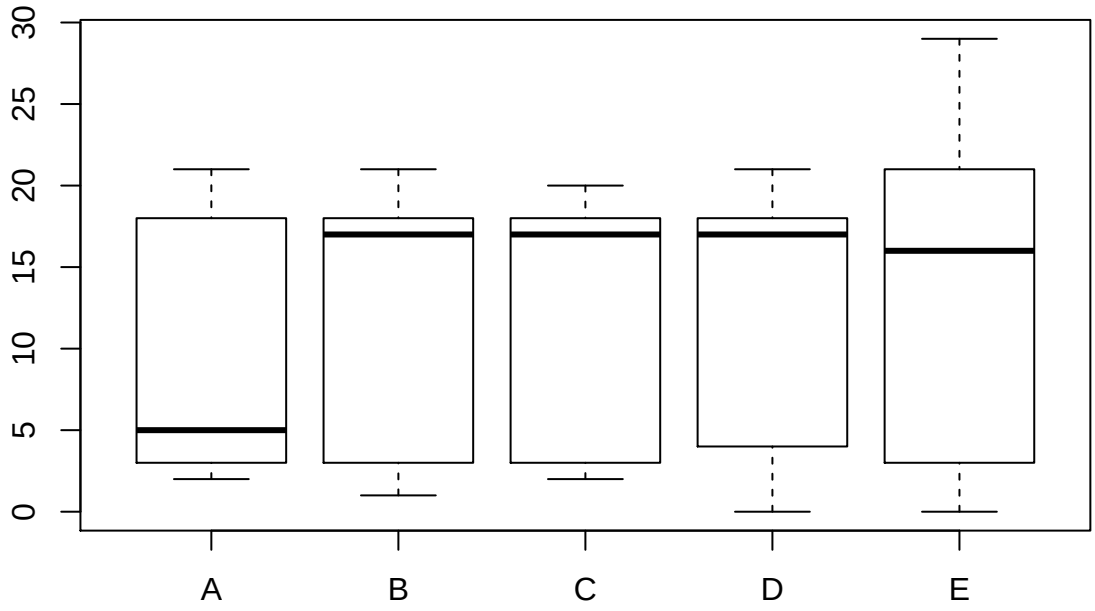


Figure 2 total_accel_dumbbell vs classe

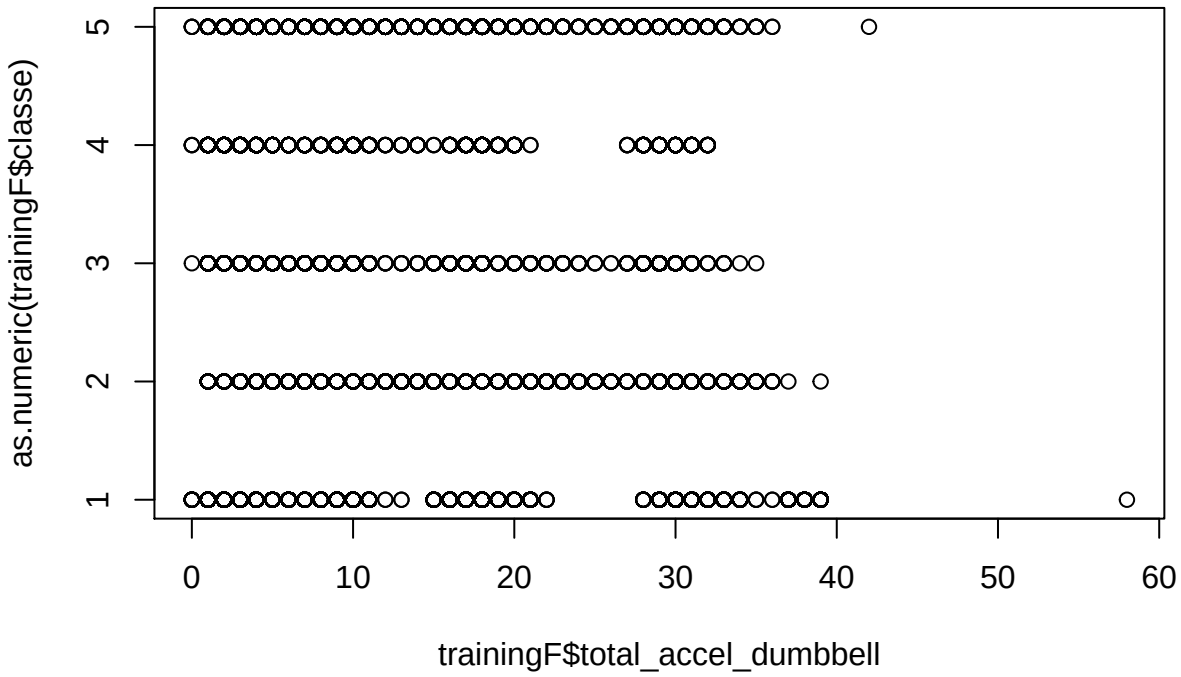
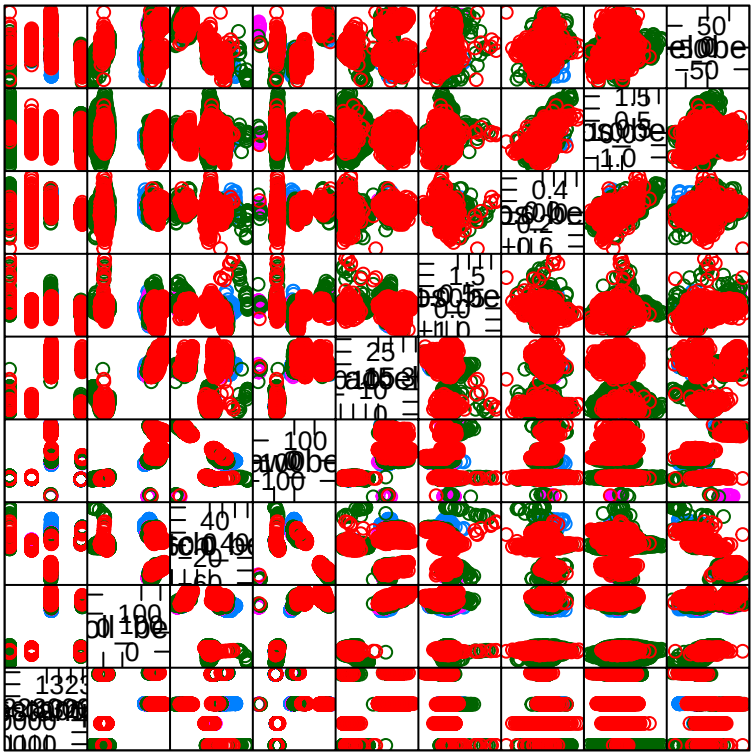


Figure 3 corrleation between first 8 significant columns



Scatter Plot Matrix