Practical machine learning

Anteneh

July 14, 2015

First we load the necessary packages and set the seed inorder to get reproduceable results and set our working directory.

```
library(AppliedPredictiveModeling)
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(rattle)
```

```
## Loading required package: RGtk2
## Rattle: A free graphical interface for data mining with R.
## Version 3.5.0 Copyright (c) 2006-2015 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(rpart.plot)
```

```
## Loading required package: rpart
```

```
library(randomForest)
```

```
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

Our data may have NA, blank and #DIV/0!. To get rid of these we defined vector of na.strings and replace by NA.Because both data sets contain columns with all missing values, we will delete these and this help us to clean a data.

```
## [1] 19622   160
```

```
## [1]  20 160
```

After that we have checked the dimension of the new data set and then delete columns with missing values.

The first 7 columns such as user_name, raw_timestamp_part_1, raw_timestamp_part_2 cvtd_timestamp, new_window, and num_window are unnecessary for predicting our project, we delete all these variables.

Our new training data set contains 53 variables and 19622 observations where as the testing data set contains 53 variables and 20 observations.

```
## [1] 19622    53
```

```
## [1] 20 53
```

Since the data we are working is too large to performe an algorithm, the given training data set partitioned into two: Training data set into two data sets, 60% for myTraining, and subTest 40%. This can be performed with random sampling without replacement.

```
## [1] 11776    53
```

```
## [1] 7846   53
```

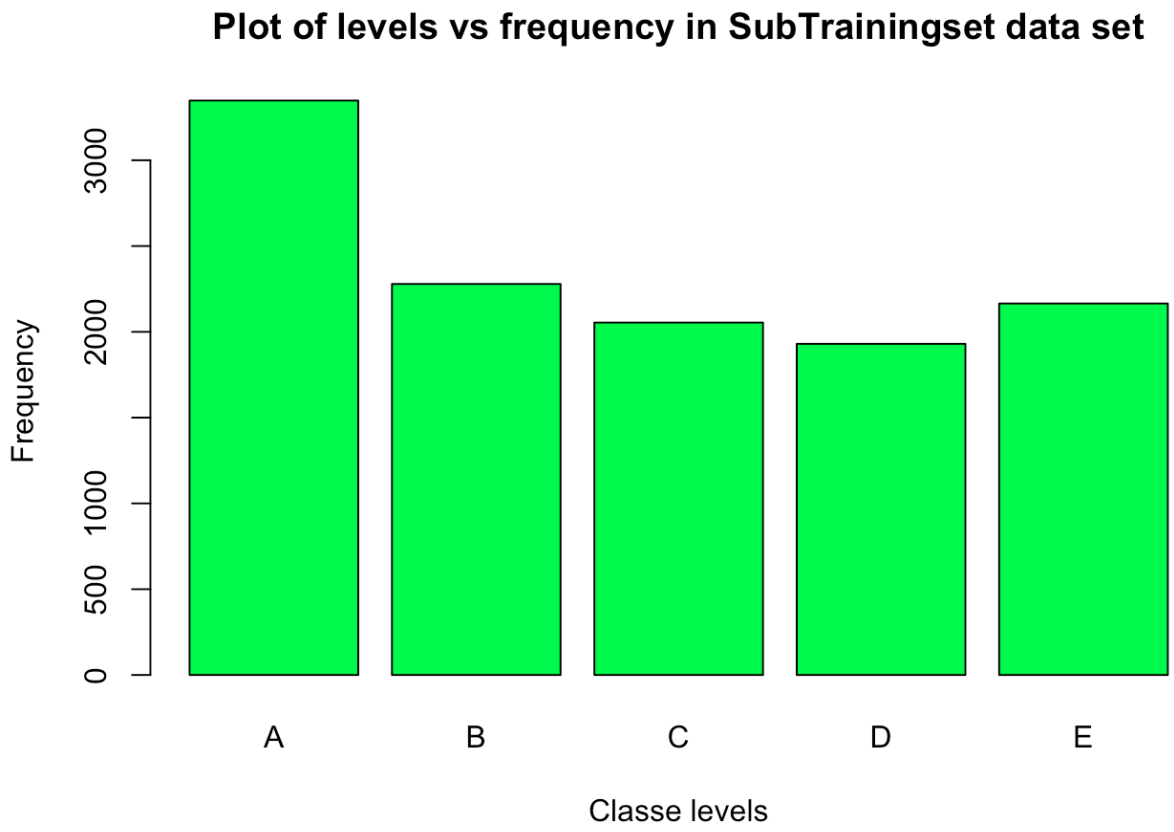
```
##    roll_belt pitch_belt yaw_belt total_accel_belt gyros_belt_x
## 2       1.41       8.07    -94.4                3         0.02
## 6       1.45       8.06    -94.4                3         0.02
## 7       1.42       8.09    -94.4                3         0.02
## 8       1.42       8.13    -94.4                3         0.02
## 9       1.43       8.16    -94.4                3         0.02
## 11      1.45       8.18    -94.4                3         0.03
##    gyros_belt_y gyros_belt_z accel_belt_x accel_belt_y accel_belt_z
## 2             0        -0.02          -22            4           22
## 6             0        -0.02          -21            4           21
## 7             0        -0.02          -22            3           21
## 8             0        -0.02          -22            4           21
## 9             0        -0.02          -20            2           24
## 11            0        -0.02          -21            2           23
##    magnet_belt_x magnet_belt_y magnet_belt_z roll_arm pitch_arm yaw_arm
## 2             -7           608          -311     -128      22.5    -161
## 6              0           603          -312     -128      22.0    -161
## 7             -4           599          -311     -128      21.9    -161
## 8             -2           603          -313     -128      21.8    -161
## 9              1           602          -312     -128      21.7    -161
## 11            -5           596          -317     -128      21.5    -161
##    total_accel_arm gyros_arm_x gyros_arm_y gyros_arm_z accel_arm_x
## 2               34        0.02       -0.02       -0.02        -290
## 6               34        0.02       -0.03        0.00        -289
## 7               34        0.00       -0.03        0.00        -289
## 8               34        0.02       -0.02        0.00        -289
## 9               34        0.02       -0.03       -0.02        -288
## 11              34        0.02       -0.03        0.00        -290
##    accel_arm_y accel_arm_z magnet_arm_x magnet_arm_y magnet_arm_z
## 2          110        -125         -369          337          513
## 6          111        -122         -369          342          513
## 7          111        -125         -373          336          509
## 8          111        -124         -372          338          510
## 9          109        -122         -369          341          518
## 11         110        -123         -366          339          509
##    roll_dumbbell pitch_dumbbell yaw_dumbbell total_accel_dumbbell
## 2       13.13074      -70.63751    -84.71065                   37
## 6       13.38246      -70.81759    -84.46500                   37
## 7       13.12695      -70.24757    -85.09961                   37
## 8       12.75083      -70.34768    -85.09708                   37
## 9       13.15463      -70.42520    -84.91563                   37
## 11      13.13074      -70.63751    -84.71065                   37
##    gyros_dumbbell_x gyros_dumbbell_y gyros_dumbbell_z accel_dumbbell_x
## 2                 0            -0.02                0             -233
## 6                 0            -0.02                0             -234
## 7                 0            -0.02                0             -232
```

```
## 8                 0            -0.02                0             -234
## 9                 0            -0.02                0             -232
## 11                0            -0.02                0             -233
##    accel_dumbbell_y accel_dumbbell_z magnet_dumbbell_x magnet_dumbbell_y
## 2                47             -269              -555                296
## 6                48             -269              -558                294
## 7                47             -270              -551                295
## 8                46             -272              -555                300
## 9                47             -269              -549                292
## 11               47             -269              -564                299
##    magnet_dumbbell_z roll_forearm pitch_forearm yaw_forearm
## 2                -64         28.3         -63.9        -153
## 6                -66         27.9         -63.9        -152
## 7                -70         27.9         -63.9        -152
## 8                -74         27.8         -63.8        -152
## 9                -65         27.7         -63.8        -152
## 11               -64         27.6         -63.8        -152
##    total_accel_forearm gyros_forearm_x gyros_forearm_y gyros_forearm_z
## 2                   36            0.02            0.00           -0.02
## 6                   36            0.02           -0.02           -0.03
## 7                   36            0.02            0.00           -0.02
## 8                   36            0.02           -0.02            0.00
## 9                   36            0.03            0.00           -0.02
## 11                  36            0.02           -0.02           -0.02
##    accel_forearm_x accel_forearm_y accel_forearm_z magnet_forearm_x
## 2              192             203            -216              -18
## 6              193             203            -215               -9
## 7              195             205            -215              -18
## 8              193             205            -213               -9
## 9              193             204            -214              -16
## 11             193             205            -214              -17
##    magnet_forearm_y magnet_forearm_z classe
## 2               661              473      A
## 6               660              478      A
## 7               659              470      A
## 8               660              474      A
## 9               653              476      A
## 11              657              465      A
```

```
##    roll_belt pitch_belt yaw_belt total_accel_belt gyros_belt_x
## 1       1.41       8.07    -94.4                3         0.00
## 3       1.42       8.07    -94.4                3         0.00
## 4       1.48       8.05    -94.4                3         0.02
## 5       1.48       8.07    -94.4                3         0.02
## 10      1.45       8.17    -94.4                3         0.03
## 14      1.42       8.21    -94.4                3         0.02
##    gyros_belt_y gyros_belt_z accel_belt_x accel_belt_y accel_belt_z
## 1          0.00        -0.02          -21            4           22
## 3          0.00        -0.02          -20            5           23
## 4          0.00        -0.03          -22            3           21
## 5          0.02        -0.02          -21            2           24
## 10         0.00         0.00          -21            4           22
## 14         0.00        -0.02          -22            4           21
##    magnet_belt_x magnet_belt_y magnet_belt_z roll_arm pitch_arm yaw_arm
## 1             -3           599          -313     -128      22.5    -161
## 3             -2           600          -305     -128      22.5    -161
## 4             -6           604          -310     -128      22.1    -161
## 5             -6           600          -302     -128      22.1    -161
## 10            -3           609          -308     -128      21.6    -161
## 14            -8           598          -310     -128      21.4    -161
##    total_accel_arm gyros_arm_x gyros_arm_y gyros_arm_z accel_arm_x
## 1               34        0.00        0.00       -0.02        -288
## 3               34        0.02       -0.02       -0.02        -289
## 4               34        0.02       -0.03        0.02        -289
## 5               34        0.00       -0.03        0.00        -289
## 10              34        0.02       -0.03       -0.02        -288
## 14              34        0.02        0.00       -0.03        -288
##    accel_arm_y accel_arm_z magnet_arm_x magnet_arm_y magnet_arm_z
## 1          109        -123         -368          337          516
## 3          110        -126         -368          344          513
## 4          111        -123         -372          344          512
## 5          111        -123         -374          337          506
## 10         110        -124         -376          334          516
## 14         111        -124         -371          331          523
##    roll_dumbbell pitch_dumbbell yaw_dumbbell total_accel_dumbbell
## 1       13.05217      -70.49400    -84.87394                   37
## 3       12.85075      -70.27812    -85.14078                   37
## 4       13.43120      -70.39379    -84.87363                   37
## 5       13.37872      -70.42856    -84.85306                   37
## 10      13.33034      -70.85059    -84.44602                   37
## 14      13.41048      -70.99594    -84.28005                   37
##    gyros_dumbbell_x gyros_dumbbell_y gyros_dumbbell_z accel_dumbbell_x
## 1              0.00            -0.02             0.00             -234
## 3              0.00            -0.02             0.00             -232
## 4              0.00            -0.02            -0.02             -232
```

```
## 5              0.00            -0.02             0.00             -233
## 10             0.00            -0.02             0.00             -235
## 14             0.02            -0.02            -0.02             -234
##    accel_dumbbell_y accel_dumbbell_z magnet_dumbbell_x magnet_dumbbell_y
## 1                47             -271              -559               293
## 3                46             -270              -561               298
## 4                48             -269              -552               303
## 5                48             -270              -554               292
## 10               48             -270              -558               291
## 14               48             -268              -554               295
##    magnet_dumbbell_z roll_forearm pitch_forearm yaw_forearm
## 1                -65         28.4         -63.9        -153
## 3                -63         28.3         -63.9        -152
## 4                -60         28.1         -63.9        -152
## 5                -68         28.0         -63.9        -152
## 10               -69         27.7         -63.8        -152
## 14               -68         27.2         -63.9        -151
##    total_accel_forearm gyros_forearm_x gyros_forearm_y gyros_forearm_z
## 1                   36            0.03            0.00           -0.02
## 3                   36            0.03           -0.02            0.00
## 4                   36            0.02           -0.02            0.00
## 5                   36            0.02            0.00           -0.02
## 10                  36            0.02            0.00           -0.02
## 14                  36            0.00           -0.02           -0.03
##    accel_forearm_x accel_forearm_y accel_forearm_z magnet_forearm_x
## 1              192             203            -215              -17
## 3              196             204            -213              -18
## 4              189             206            -214              -16
## 5              189             206            -214              -17
## 10             190             205            -215              -22
## 14             193             202            -214              -14
##    magnet_forearm_y magnet_forearm_z classe
## 1               654              476      A
## 3               658              469      A
## 4               658              469      A
## 5               655              473      A
## 10              656              473      A
## 14              659              478      A
```

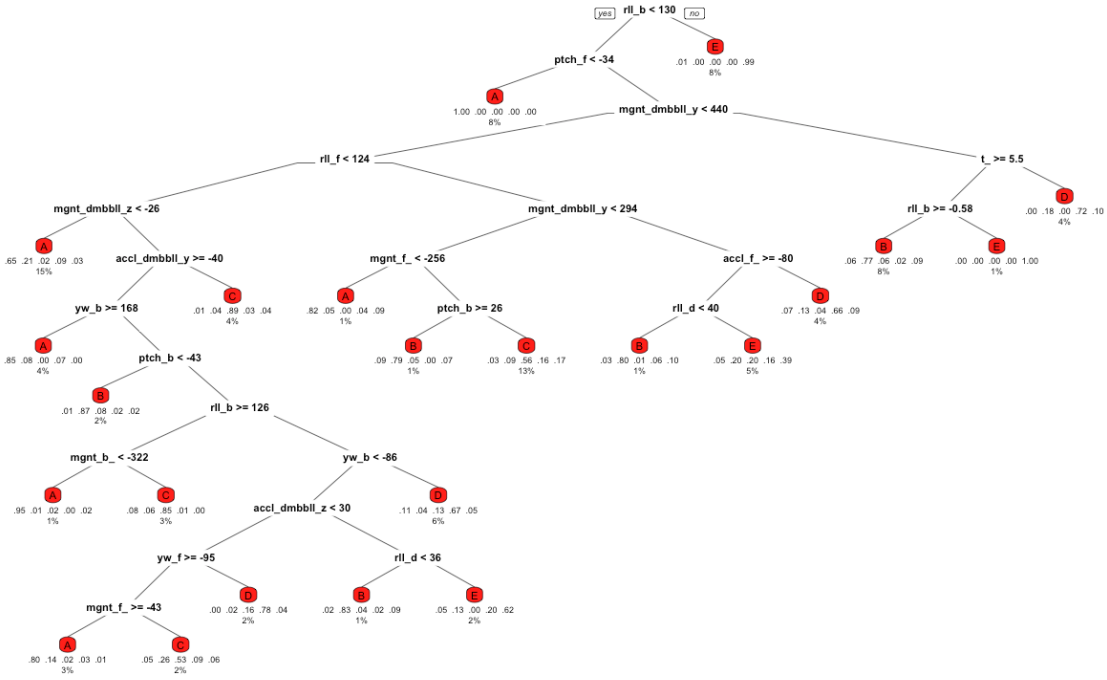
When we look at variable “classe”, it contains 5 levels: A, B, C, D and E. A plot of the outcome variable will allow us to see the frequency of each levels in the SubTrainingset data set. As we can see in the figure below level A has more than 4000 occurrences than other levels.



For predicition we used Decision Tree and Random Forest prediction modesl.

1. Predicting using Decision Tree and Testing the results on SubTestingset data set

Plot of the Decision Tree



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2015  316   30  119   48
##          B   46  809   63   29   74
##          C   68  165 1120  220  190
##          D   77  121   78  817   81
##          E   26  107   77  101 1049
##
## Overall Statistics
##
##                Accuracy : 0.7405
##                  95% CI : (0.7307, 0.7502)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6705
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9028   0.5329   0.8187   0.6353   0.7275
## Specificity            0.9086   0.9665   0.9007   0.9456   0.9514
## Pos Pred Value         0.7971   0.7924   0.6353   0.6959   0.7713
## Neg Pred Value         0.9592   0.8961   0.9592   0.9297   0.9394
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2568   0.1031   0.1427   0.1041   0.1337
## Detection Prevalence   0.3222   0.1301   0.2247   0.1496   0.1733
## Balanced Accuracy      0.9057   0.7497   0.8597   0.7904   0.8394
```

2. Predicting using Random Forest and Test the results on SubTestingset data set.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2231    6    0    0    0
##          B    1 1504   14    0    0
##          C    0    8 1354   22    0
##          D    0    0    0 1261    3
##          E    0    0    0    3 1439
##
## Overall Statistics
##
##                Accuracy : 0.9927
##                  95% CI : (0.9906, 0.9945)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9908
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9996   0.9908   0.9898   0.9806   0.9979
## Specificity            0.9989   0.9976   0.9954   0.9995   0.9995
## Pos Pred Value         0.9973   0.9901   0.9783   0.9976   0.9979
## Neg Pred Value         0.9998   0.9978   0.9978   0.9962   0.9995
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2843   0.1917   0.1726   0.1607   0.1834
## Detection Prevalence   0.2851   0.1936   0.1764   0.1611   0.1838
## Balanced Accuracy      0.9992   0.9942   0.9926   0.9901   0.9987
```

As shown above the accuracy for Random Forest model is 0.9927 where as for for Decision Tree model is 0.7405. Therefore Random Forest algorithm is choosen because it performed better than Decision Trees. From our cross-validation data none of the test samples will be missclassified.

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  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
```

```
# Write files for submission
pml_write_files = function(x){
      n = length(x)
      for(i in 1:n){
            filename = paste0("problem_id_",i,".txt")
            write.table(x[i],file=filename,quote=FALSE,row.names=FALSE,col.name
s=FALSE)
      }
}

pml_write_files(predictedresult)
```

References

1. http://www.jstatsoft.org/v28/i05/paper (http://www.jstatsoft.org/v28/i05/paper)

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.