Practical Machine Learning Course Project

Lotte Sluyser

16 september 2017

Introduction to the project

In this project the goal was to use data from accelerometers on the belt, forearm, arm and dumbell to quantify how well 6 participants do a particular activity. Participants were asked to perform barbell lifts correctly and incorrectly in 5 different ways. Goal was to predict the manner in which they did the exercise, the "classe" variable in the training set. This report describes how the model was built, how cross validation was used and the expected out of sample error. The model was used to predict 20 different test cases.

Loading the data

```
fileUrl<- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
download.file(fileUrl,destfile="train.csv", method="curl")
training = read.csv("~/Desktop/coursera/Datascience_cursus_8/train.csv")
fileUrl2<- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
download.file(fileUrl2,destfile="test.csv", method="curl")
testing = read.csv("~/Desktop/coursera/Datascience_cursus_8/test.csv")
dim(training);dim(testing)
```

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

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

Data processing

It is checked how many missing values (na's) there are per column:

```
sapply(training, function(x) sum(is.na(x)))
```

```
##                        X                user_name     raw_timestamp_part_1
##                        0                        0                        0
##     raw_timestamp_part_2           cvtd_timestamp               new_window
##                        0                        0                        0
##               num_window                roll_belt               pitch_belt
##                        0                        0                        0
##                yaw_belt         total_accel_belt       kurtosis_roll_belt
##                        0                        0                        0
##      kurtosis_picth_belt        kurtosis_yaw_belt       skewness_roll_belt
##                        0                        0                        0
##     skewness_roll_belt.1        skewness_yaw_belt            max_roll_belt
##                        0                        0                    19216
##           max_picth_belt             max_yaw_belt            min_roll_belt
##                    19216                        0                    19216
##           min_pitch_belt             min_yaw_belt      amplitude_roll_belt
##                    19216                        0                    19216
##     amplitude_pitch_belt       amplitude_yaw_belt     var_total_accel_belt
```

```
##                   19216                       0                   19216
##            avg_roll_belt        stddev_roll_belt            var_roll_belt
##                   19216                   19216                   19216
##           avg_pitch_belt       stddev_pitch_belt           var_pitch_belt
##                   19216                   19216                   19216
##             avg_yaw_belt         stddev_yaw_belt             var_yaw_belt
##                   19216                   19216                   19216
##             gyros_belt_x            gyros_belt_y            gyros_belt_z
##                       0                       0                       0
##             accel_belt_x            accel_belt_y            accel_belt_z
##                       0                       0                       0
##            magnet_belt_x           magnet_belt_y           magnet_belt_z
##                       0                       0                       0
##                roll_arm               pitch_arm                 yaw_arm
##                       0                       0                       0
##          total_accel_arm           var_accel_arm            avg_roll_arm
##                       0                   19216                   19216
##          stddev_roll_arm            var_roll_arm           avg_pitch_arm
##                   19216                   19216                   19216
##         stddev_pitch_arm           var_pitch_arm             avg_yaw_arm
##                   19216                   19216                   19216
##           stddev_yaw_arm             var_yaw_arm             gyros_arm_x
##                   19216                   19216                       0
##             gyros_arm_y             gyros_arm_z             accel_arm_x
##                       0                       0                       0
##             accel_arm_y             accel_arm_z            magnet_arm_x
##                       0                       0                       0
##            magnet_arm_y            magnet_arm_z       kurtosis_roll_arm
##                       0                       0                       0
##       kurtosis_picth_arm        kurtosis_yaw_arm       skewness_roll_arm
##                       0                       0                       0
##       skewness_pitch_arm        skewness_yaw_arm            max_roll_arm
##                       0                       0                   19216
##            max_picth_arm            max_yaw_arm            min_roll_arm
##                   19216                   19216                   19216
##            min_pitch_arm            min_yaw_arm      amplitude_roll_arm
##                   19216                   19216                   19216
##      amplitude_pitch_arm       amplitude_yaw_arm           roll_dumbbell
##                   19216                   19216                       0
##           pitch_dumbbell           yaw_dumbbell   kurtosis_roll_dumbbell
##                       0                       0                       0
##  kurtosis_picth_dumbbell   kurtosis_yaw_dumbbell   skewness_roll_dumbbell
##                       0                       0                       0
##  skewness_pitch_dumbbell   skewness_yaw_dumbbell       max_roll_dumbbell
##                       0                       0                   19216
##       max_picth_dumbbell        max_yaw_dumbbell       min_roll_dumbbell
##                   19216                       0                   19216
##       min_pitch_dumbbell        min_yaw_dumbbell  amplitude_roll_dumbbell
##                   19216                       0                   19216
## amplitude_pitch_dumbbell  amplitude_yaw_dumbbell     total_accel_dumbbell
##                   19216                       0                       0
##       var_accel_dumbbell        avg_roll_dumbbell     stddev_roll_dumbbell
##                   19216                   19216                   19216
##        var_roll_dumbbell       avg_pitch_dumbbell    stddev_pitch_dumbbell
```

```
##                    19216                    19216                    19216
##       var_pitch_dumbbell         avg_yaw_dumbbell       stddev_yaw_dumbbell
##                    19216                    19216                    19216
##         var_yaw_dumbbell         gyros_dumbbell_x         gyros_dumbbell_y
##                    19216                        0                        0
##         gyros_dumbbell_z         accel_dumbbell_x         accel_dumbbell_y
##                        0                        0                        0
##         accel_dumbbell_z        magnet_dumbbell_x        magnet_dumbbell_y
##                        0                        0                        0
##        magnet_dumbbell_z             roll_forearm            pitch_forearm
##                        0                        0                        0
##              yaw_forearm    kurtosis_roll_forearm   kurtosis_picth_forearm
##                        0                        0                        0
##     kurtosis_yaw_forearm    skewness_roll_forearm   skewness_pitch_forearm
##                        0                        0                        0
##     skewness_yaw_forearm         max_roll_forearm        max_picth_forearm
##                        0                    19216                    19216
##          max_yaw_forearm         min_roll_forearm        min_pitch_forearm
##                        0                    19216                    19216
##          min_yaw_forearm   amplitude_roll_forearm  amplitude_pitch_forearm
##                        0                    19216                    19216
##    amplitude_yaw_forearm      total_accel_forearm        var_accel_forearm
##                        0                        0                    19216
##         avg_roll_forearm      stddev_roll_forearm         var_roll_forearm
##                    19216                    19216                    19216
##        avg_pitch_forearm     stddev_pitch_forearm        var_pitch_forearm
##                    19216                    19216                    19216
##          avg_yaw_forearm       stddev_yaw_forearm          var_yaw_forearm
##                    19216                    19216                    19216
##          gyros_forearm_x          gyros_forearm_y          gyros_forearm_z
##                        0                        0                        0
##          accel_forearm_x          accel_forearm_y          accel_forearm_z
##                        0                        0                        0
##         magnet_forearm_x         magnet_forearm_y         magnet_forearm_z
##                        0                        0                        0
##                   classe
##                        0
```

There are many columns with 19216 na's. These are removed.

```
a<-Filter(function(x) sum(is.na(x)) < 19216, training)
b<-Filter(function(x) sum(is.na(x)) < 19216, testing)
```

NearZerovar is applied to remove predictors with only 1 value

```
a_nzv<- nearZeroVar(a)
new_a<- a[,-a_nzv]
b_nzv<- nearZeroVar(b)
new_b<- b[,-b_nzv]
dim(new_a);dim(new_b)
```

```
## [1] 19622    59
```

```
## [1] 20 59
```

There are 59 variables left. The first 6 variables are deleted as they are descriptive and no measures

```
trainnew<-new_a[,-c(1:6)]
testnew <-new_b[,-c(1:6)]
```

The validation set is built by splitting off 30% of the trainingset:

```
inTrain<- createDataPartition(y=trainnew$classe,p=0.7,list=FALSE)
trainset<- trainnew[inTrain,]
validationset<- trainnew[-inTrain,]
```

Building the model

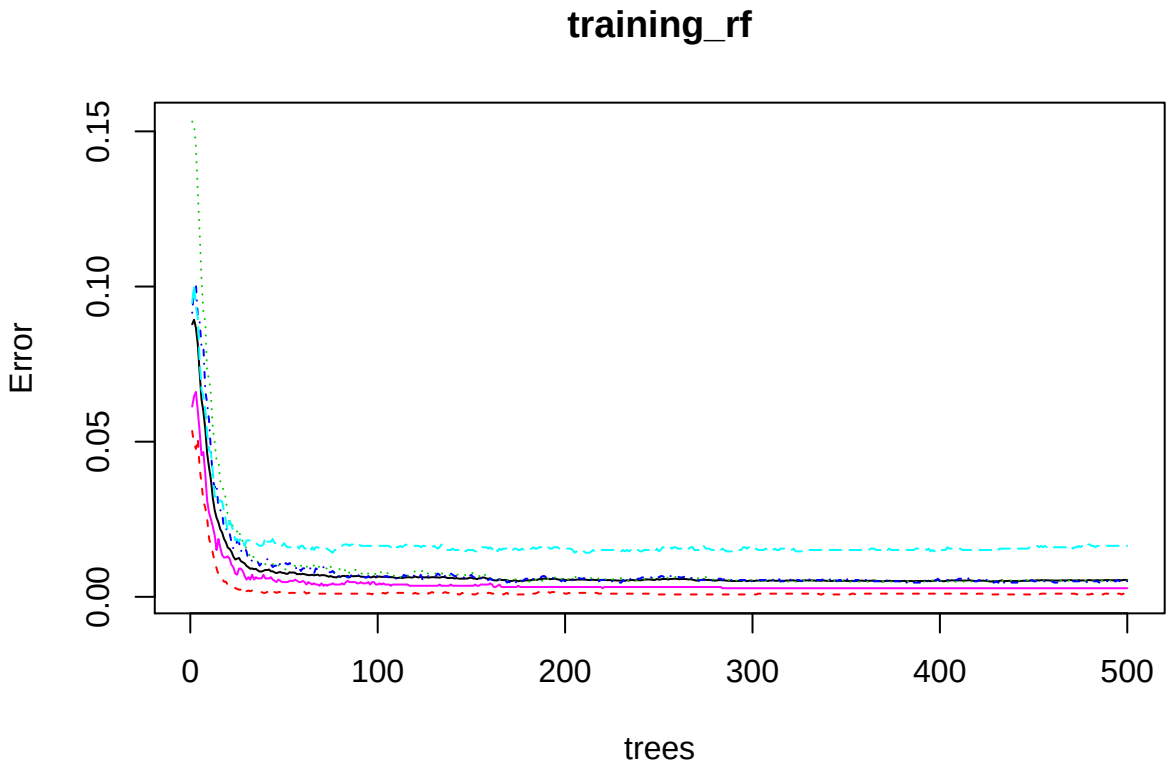
A random forest model is built:

```
set.seed(7)
training_rf <- randomForest(x=trainset[,1:(ncol(trainset)-1)], y=trainset["classe"], importance=TRUE, c
```

```
## ntree      OOB      1      2      3      4      5
##   100:   0.65%  0.10%  0.83%  0.67%  1.64%  0.40%
##   200:   0.54%  0.10%  0.56%  0.50%  1.55%  0.32%
##   300:   0.52%  0.08%  0.53%  0.50%  1.55%  0.28%
##   400:   0.52%  0.10%  0.53%  0.50%  1.55%  0.28%
##   500:   0.54%  0.10%  0.53%  0.50%  1.64%  0.28%
```

OOB = 0.50% Accuracy=1-OOB=99.5%

```
plot(training_rf)
```



The model is tested on the validation set:

```
pred<-predict(training_rf,validationset);validationset$predRight<-pred==validationset$classe
confusionMatrix(pred, validationset$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1669   10    0    0    0
##          B    5 1125    7    0    0
##          C    0    4 1018    7    1
##          D    0    0    1  956    3
##          E    0    0    0    1 1078
##
## Overall Statistics
##
##                Accuracy : 0.9934
##                  95% CI : (0.991, 0.9953)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9916
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9970   0.9877   0.9922   0.9917   0.9963
## Specificity            0.9976   0.9975   0.9975   0.9992   0.9998
## Pos Pred Value         0.9940   0.9894   0.9883   0.9958   0.9991
## Neg Pred Value         0.9988   0.9971   0.9984   0.9984   0.9992
## Prevalence             0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate         0.2836   0.1912   0.1730   0.1624   0.1832
## Detection Prevalence   0.2853   0.1932   0.1750   0.1631   0.1833
## Balanced Accuracy      0.9973   0.9926   0.9949   0.9954   0.9980
```

Accuracy : 0.9968 oob=1-0.9968=0.0032

Use prediction model to predict 20 testcases

```
pred2<-predict(training_rf,testnew)
pred2
```

```
##  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
```