PML Course Project

KOALA Valentin

26 septembre 2017

Background

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]

The data for this project come from this source: [http://groupware.les.inf.puc-rio.br/har]. If you use the document you create for this class for any purpose please cite them as they have been very generous in allowing their data to be used for this kind of assignment.

The classe variable contains 5 different ways barbell lifts were performed correctly and incorrectly:

- Class A: exactly according to the specification
- Class B: throwing the elbows to the front
- Class C: lifting the dumbbell only halfway
- Class D: lowering the dumbbell only halfway
- Class E: throwing the hips to the front

Objective

The goal my project is to predict the manner in which people performed barbell lifts. This is the classe variable in the training set. You may use any of the other variables to predict with. You should create a report describing how you built your model, how you used cross validation, what you think the expected out of sample error is, and why you made the choices you did. You will also use your prediction model to predict 20 different test cases.

Loading the data

Packages used for analysis. This assumes the packages are already installed. Use the install.packages("") command if a package not installed yet.

```r
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```r
library(randomForest)
```

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(rpart)
library(rpart.plot)
library(RColorBrewer)
```

Load the data into R

```r
# The location where the training data is to be downloaded from
trainUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
# The location where the testing data is to be downloaded from
testUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

# Reading/loading the training data
train_data <- read.csv(url(trainUrl), na.strings=c("NA","#DIV/0!",""))
# Reading/loading the testing data in your working directory
test_data <- read.csv(url(testUrl), na.strings=c("NA","#DIV/0!",""))

# Take a look at the Training data classe variable
summary(train_data$classe)
```

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

Partitioning the data for Cross-validation

The training data is split into two data sets, one for training the model and one for testing the performance of our model. The data is partitioned by the classe variable, which is the varible we will be predicting. The data is split into 60% for training and 40% for testing.

```r
inTrain <- createDataPartition(y=train_data$classe, p = 0.60, list=FALSE)
training <- train_data[inTrain,]
testing <- train_data[-inTrain,]

dim(training)
```

```
## [1] 11776   160
```

```r
dim(testing)
```

```
## [1] 7846  160
```

Data Processing

Drop the first 7 variables because these are made up of metadata that would cause the model to perform poorly.

```
training <- training[,-c(1:7)]
```

Remove NearZeroVariance variables

```
nzv <- nearZeroVar(training, saveMetrics=TRUE)
training <- training[, nzv$nzv==FALSE]
```

There are a lot of variables where most of the values are 'NA'. Drop variables that have 60% or more of the values as 'NA'.

```
training_clean <- training
for(i in 1:length(training)) {
  if( sum( is.na( training[, i] ) ) /nrow(training) >= .6) {
    for(j in 1:length(training_clean)) {
      if( length( grep(names(training[i]), names(training_clean)[j]) ) == 1)  {
        training_clean <- training_clean[ , -j]
      }
    }
  }
}

# Set the new cleaned up dataset back to the old dataset name
training <- training_clean
```

Transform the test_data dataset

```
# Get the column names in the training dataset
columns <- colnames(training)
# Drop the class variable
columns2 <- colnames(training[, -53])
# Subset the test data on the variables that are in the training data set
test_data <- test_data[columns2]
dim(test_data)
```

```
## [1] 20 52
```

Cross-Validation: Prediction with Random Forest

A Random Forest model is built on the training set. Then the results are evaluated on the test set

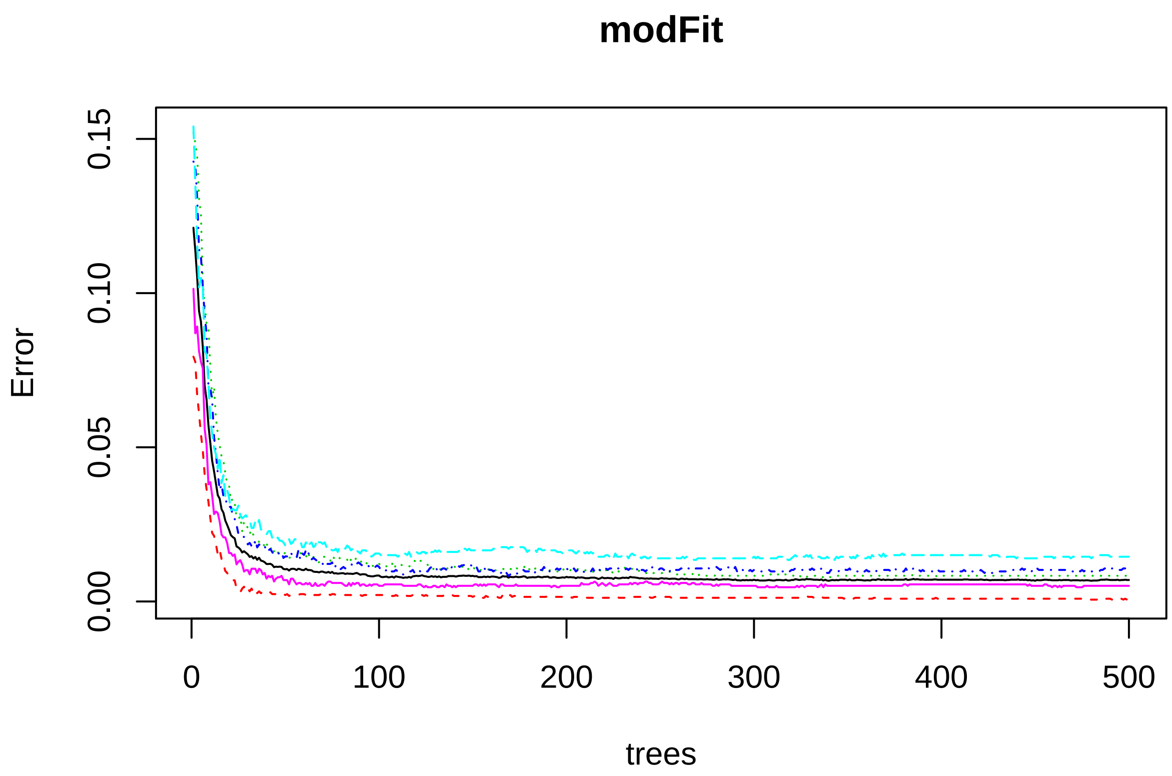
```
set.seed(54321)
modFit <- randomForest(classe ~ ., data=training)
prediction <- predict(modFit, testing)
cm <- confusionMatrix(prediction, testing$classe)
print(cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2230    6    0    0    0
##          B    1 1507    7    0    0
##          C    0    4 1360   25    5
##          D    0    0    1 1260    4
##          E    1    1    0    1 1433
##
## Overall Statistics
##
##                Accuracy : 0.9929
##                  95% CI : (0.9907, 0.9946)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.991
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9991   0.9928   0.9942   0.9798   0.9938
## Specificity            0.9989   0.9987   0.9948   0.9992   0.9995
## Pos Pred Value         0.9973   0.9947   0.9756   0.9960   0.9979
## Neg Pred Value         0.9996   0.9983   0.9988   0.9960   0.9986
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2842   0.1921   0.1733   0.1606   0.1826
## Detection Prevalence   0.2850   0.1931   0.1777   0.1612   0.1830
## Balanced Accuracy      0.9990   0.9957   0.9945   0.9895   0.9966
```

```
overall.accuracy <- round(cm$overall['Accuracy'] * 100, 2)
sam.err <- round(1 - cm$overall['Accuracy'],2)
```

The model is 99.29% accurate on the testing data partitioned from the training data. The expected out of sample error is roughly 0.01.

```
plot(modFit)
```



In the above figure, error rates of the model are plotted over 500 trees. The error rate is less than 0.04 for all 5 classe.

Cross-Validation: Prediction with a Decision Tree

```
set.seed(54321)
modFit2 <- rpart(classe ~ ., data=training, method="class")
prediction2 <- predict(modFit2, testing, type="class")
cm2 <- confusionMatrix(prediction2, testing$classe)
print(cm2)
```

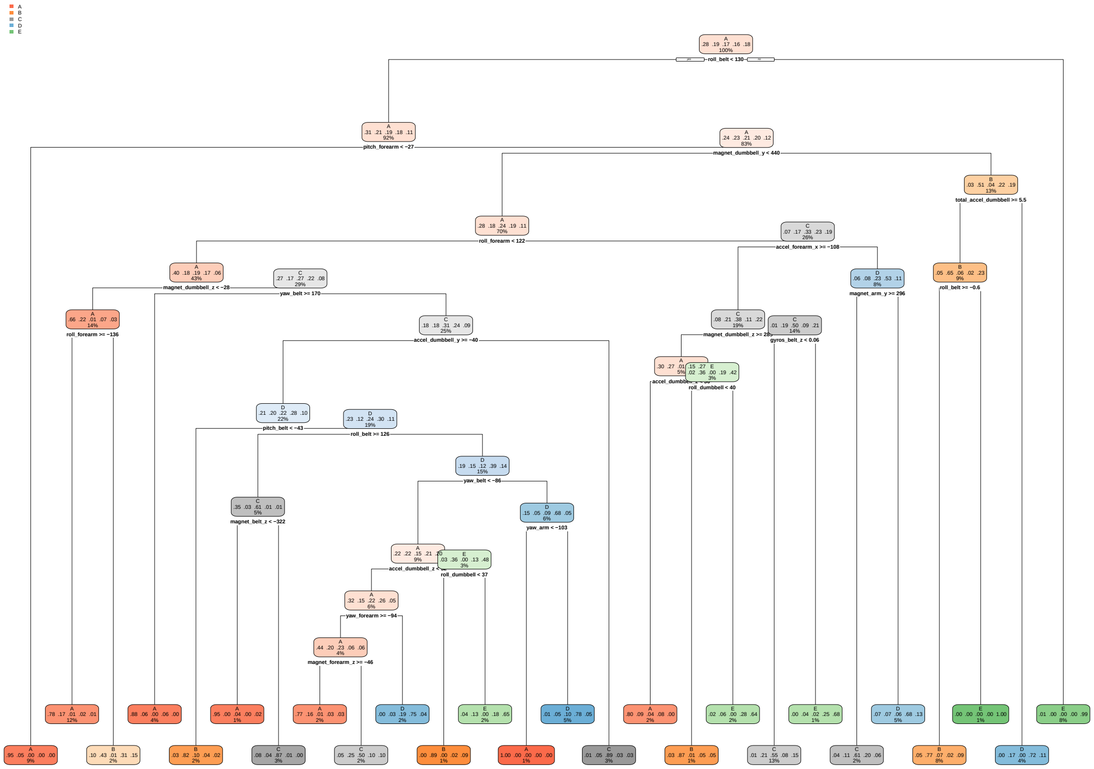
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2060  246   25   47    7
##          B   68  860   72   99  106
##          C   54  287 1163  151  200
##          D   34   90  108  873   95
##          E   16   35    0  116 1034
##
## Overall Statistics
##
##                Accuracy : 0.7634
##                  95% CI : (0.7539, 0.7728)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7003
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9229   0.5665   0.8501   0.6788   0.7171
## Specificity            0.9421   0.9455   0.8932   0.9502   0.9739
## Pos Pred Value         0.8637   0.7137   0.6270   0.7275   0.8609
## Neg Pred Value         0.9685   0.9009   0.9658   0.9379   0.9386
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2626   0.1096   0.1482   0.1113   0.1318
## Detection Prevalence   0.3040   0.1536   0.2364   0.1529   0.1531
## Balanced Accuracy      0.9325   0.7560   0.8717   0.8145   0.8455
```

```r
overall.accuracy2 <- round(cm2$overall['Accuracy'] * 100, 2)
sam.err2 <- round(1 - cm2$overall['Accuracy'],2)
```

The model is 76.34% accurate on the testing data partitioned from the training data. The expected out of sample error is roughly 0.24.

Plot the decision tree model

```r
rpart.plot(modFit2)
```



Prediction on the Test Data

The Random Forest model gave an accuracy of 99.29, which is much higher than the 76.34% accuracy from the Decision Tree. So we will use the Random Forest model to make the predictions on the test data to predict the way 20 participates performed the exercise.

```
final_prediction <- predict(modFit, test_data, type="class")
print(final_prediction)
```

```
##  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
```

Conclusions

I chose to compare a Random Forest and Decision Tree model. For this data, the Random Forest proved to be a more accurate way to predict the manner in which the exercise was done.