Practical Machine Learning Course Project

Youri Matiounine

April 7, 2018

Summary

In this project we will construct a prediction model to recognize the manner in which people perform the Unilateral Dumbbell Biceps Curl. We will then apply our model to the test data set, and use predicted results to answer questions in the final quiz.

Background

Human Activity Recognition - HAR - has emerged as a key research area in the last years and is gaining increasing attention by the pervasive computing research community. This human activity recognition research has traditionally focused on discriminating between different activities, i.e. to predict "which" activity was performed at a specific point in time. But in this project we will try to predict how the activity is performed.

The data for this project comes from this source: http://groupware.les.inf.puc-rio.br/har. Full source:

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013.

Data Pre-processing

First, we load the data.

```
library(caret)
d0<-read.csv("pml-training.csv",header=TRUE,na.strings ="NA")
dim(d0)
```

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

If we look at the summary of the data (not shown here due to its large size), we see that many variables are not present for all the data samples. So we will exclude all the variables not present for all data samples. We will also exclude variables not directly related to the outcome (such as name, sample identifier, time, etc).

```
ii<-c(8,9,10,11,37,38,39,40,41,42,43,44,45,46,47,48,49,60,61,62,63,64,65,66,67,68,84,85,86,102,113,114,
d1<-d0[,ii]
dim(d1)
```

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

This leaves us with 52 variables which we can use in our model.

Selecting and Traing the Prediction Model

For classification a natuaral model choice is a form of a decision tree, such as "random forest" model. We will use "caret" package's "train" function with method set to "random forest" to train our model. We will use all the variables, and resampling with 5-fold cross validation.

```
set.seed(78259)
tc<-trainControl(method="cv",number=5,verboseIter=FALSE)
f<-train(classe~.,data=d1,method="rf",trControl=tc)
print(f$finalModel)
```

```
##
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 27
##
##         OOB estimate of  error rate: 0.41%
## Confusion matrix:
##      A    B    C    D    E class.error
## A 5574    3    2    0    1 0.001075269
## B   18 3775    4    0    0 0.005794048
## C    0   10 3403    9    0 0.005552309
## D    0    0   19 3194    3 0.006840796
## E    0    1    5    5 3596 0.003049626
```

This model produces training acuracy of approximately 99.6%. Even with cross validation this acuracy is overestimated, so on the test data we expect acuracy to be somewhat lower.

Predicting Outcomes of Test Set

Now we will use "predict" function to predict outcomes for the test data set.

```
d0t<-read.csv("pml-testing.csv",header=TRUE,na.strings ="NA")
p1<-predict(f,d0t)
print(p1)
```

```
##  [1] 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
```

These outcomes will be used to answer questions of the final quiz.