ProjectAssignmentML

G. Hol

December 4, 2017

Summary

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. The goal is to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants and predict the manner in which they did the exercise. The random forest model had an accuracy of 0.993.

Data The training data: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv The test data: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv Data source: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har.

Approach: 0. install libraries; load data; set seed 1. explore data to decide on preprocessing steps 2. select only explaning variables that occurred in the test data set; exclude time measurements. 3. create own testing data set to enable cross validation 4. run random forest and cross validate 5. apply the 'best' model to predict classe for the testing data set

```
#step 0

library(caret); library(klaR); library(randomForest)
```

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

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

```
## Warning: package 'klaR' was built under R version 3.4.3
```

```
## Loading required package: MASS
```

```
## 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
```

```
set.seed(32343)

fileurl1<-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
training<-read.csv(fileurl1,stringsAsFactors = FALSE)
fileurl2<-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
testing<-read.csv(fileurl2,stringsAsFactors = FALSE)
training$classe<-as.factor(training$classe)
```

During data exploration I noticed a high numbers of NAs, and decided to only use variables that occurred in the test dataset. I also looked at the correlation between predictors; those were high and thus I decided to use pca for preprocessing. Here a data partition is made to create an own testing set to test how well the model is at prediction the correct classes.

```r
#step 4
modrf<-train(classe ~., method="rf", data=training2)

prf<-predict(modrf,testing2)

#out of sample error
sum(prf==testing2$classe)/nrow(testing2)
```

```
## [1] 0.9930669
```

```r
#step 5 Course Project Prediction Quiz Portion
predict(modrf, testing)
```

```
##  [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
```

Random forest gave the highest accuracy, with more than 99% cases correctly predicted.