Coursera_JH_ML Assignment

Oscar BA

4/6/2020

Coursera-JH Practical Machine Learning Final Project

The aim of this project is to predict how well did subjects do exercises using data from acceloremeters put on belts, arms, forearms and dumbells for six participants. The exercise under scrutiny was dumbbell biceps curl, done in five different ways: correctly (A), throwing the elbows to the front (B), lifting only halfway (C), lowering only halfway (D) and throwing the hips to the front (E). The data can be found in http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har).

The necessary libraries for this project are:

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

For the construction of the model used, I first subset the control variables and kept only those with positive variability on their observations.

```
har <- read.csv("pml-training.csv", na.strings = '#DIV/0!')
validation <- read.csv('pml-testing.csv', na.strings = '#DIV/0!')
names(validation) <- names(har)
set.seed(5321)

x <- vector(length = length(names(har)))

for (i in 1:length(names(har))) {
  if (i!=2& i!=5& i!=6 & i!=160) {
    har[,i] <- as.numeric(as.character(har[,i]))
    validation[,i] <- as.numeric(as.character(validation[,i]))
  }
  if (class(har[,i])!='factor') {
    x[i] <- isTRUE(sum(is.na(har[,i]))==length(har[,i]) | var(har[,i], na.rm = TRUE)==0
                   | is.na(var(har[,i], na.rm = TRUE)) | sum(is.na(validation[,i]))==length(vali
dation[,i]))
  }
}

x[c(1,3,4,5)] <- TRUE
# Subset:
har <- har[,!x]
validation <- validation[,!x]
```

Then, the training set was partitioned to be able to test the in-sample accuracy; 70% into the train set and 30% into the testing set.

```
intrain <- createDataPartition(har$classe, p=.7, list=TRUE)[[1]]
training <- har[intrain,]
testing <- har[-intrain,]
tests <- testing[complete.cases(testing),]
```

With the training set, a gradient boosting model was constructed.

```
fitgbm <- train(classe~., data = training, method='gbm', na.action = na.omit)
```

The in sample accuracy was of this model is high: 99% were correctly predicted in the testing set.

```
predgbm <- predict(fitgbm, tests)
cmgbm <- confusionMatrix(tests$classe, predgbm)
cmgbm$overall[1]
```

```
## Accuracy
## 0.988785
```

Additionally, predicted correctly all the cases in the validation set.

```
pdct <- predict(fitgbm, validation, na.action = na.omit)
```

This suggests that the out of sample accuracy of the model can be very high.

A random forest model was also constructed. However, the run time was too long, which is why the GBM model was preferred.