Practical Machine Learning: Week 4 Peer Graded Assignment

Nora Myerson

December 11, 2018

Overview

We will build a model that uses movement data from activity devices to predict the type of Bicep Curl movement being done (correctly or incorrectly). The different ways of completing the curl are classified the following way: Class A: exactly accoring to the specification Class B: throwing the elbows out to the front Class C: lifitng the dumbbell only halfway Class D: lowering the dumbbell only halfway Class E: throwing the hips to the front

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

Load Data and necessary packages

```
library(xlsx); library(caret); library(e1071)
test <- read.csv(file = "C:/Users/preisn1/Documents/pml-testing.csv",
na.strings = c("NA","","#DIV/0!"))
train <- read.csv(file = "C:/Users/preisn1/Documents/pml-
training.csv",na.strings = c("NA","","#DIV/0!"))
```

Review & Explore Data

```
dim(train) ; dim(test)
##We see that the data sets have the same number of variables. The test set
can be used for cross validation of the model built on the training set.
##Next we look at the variables in more detail
summary(train)
##We saw evidence of NA's above, Find the percentage of NA's in each column
colMeans(is.na(train))
```

Clean the Data to address NA’s

```
##Remove Columns with over 95% NA's or blank ("") values
cleanIndex <- colSums(is.na(train)/nrow(train)) <.95
trainClean <- train[,cleanIndex]
##validate clean and review data
colMeans(is.na(trainClean))
head(trainClean,3)
##timestamps, names, and windows are not relevant to our movement
```

```
classification prediction so we can remove columns 1-7 them from the set.
trainClean<-trainClean[,-c(1:7)]
```

Clean the test data so we can proceede with modeling

```
testClean <- test[,names(trainClean[1:52])]
##test should now have 1 less column than train (class column) and 20
observations
dim(testClean);dim(trainClean)

## [1] 20 52

## [1] 19622    53
```

Split training data into training and test sets for validation

We will use the traning set to train and the testing set to test before predicting on the Test data imported originally, which does not have Classe identified. The test set we create will suffice for Cross Validation.

```
set.seed(519)
inTrain <- createDataPartition(trainClean$classe,p=0.70)[[1]]
inTrainClean <- trainClean[inTrain,]
inTestClean <- trainClean[-inTrain,]
```

Build models for classification

We will build two different models and pick the one with the highest accuracy and lowest error.

```
## Since this is a classification problem, we will use a random forest (rf)
and Stochastic Gradient Boosing (gbm) to classify the types of Bicep Curls
being performed.

set.seed(2719)
mod_GBM <- train(classe~., data = inTrainClean, method = "gbm",verbose =
FALSE )
mod_RF <- train(classe~.,data = inTrainClean, method = "rf")
```

Compare model accuracy and error on testing data and choose optimal model

```
#Predict on test data for CV
pred_RF <- predict(mod_RF,inTestClean)
pred_GBM <- predict(mod_GBM, inTestClean)

#Random Forest Accuracy
postResample(pred = pred_RF, obs =inTestClean$classe)

##  Accuracy     Kappa
## 0.9932031 0.9914013
```

```
#GBM Accuracy
postResample(pred = pred_GBM, obs =inTestClean$classe)

##  Accuracy     Kappa 
## 0.9600680 0.9494838

#Random Forest Error
1-as.numeric(confusionMatrix(inTestClean$classe, pred_RF)$overall[1])

## [1] 0.006796941

#GBM Error
1-as.numeric(confusionMatrix(inTestClean$classe, pred_GBM)$overall[1])

## [1] 0.03993203
```

Because accuracy was highest (99%) in our random forest model and therefore error was the lowest at 0.68%. We will choose this as our final model to predict on the test set.

Predict classes for the original set of Test data

```
pred_Test <- predict(mod_RF, testClean)
pred_Test

##  [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
```

Source:

http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har
This site was used for details pertaining to the data used to build this model.