Human Activity Recognition

Franklin X. Dono

11/9/2020

Executive summary

There is a growing interest to quantify human activity but it is important to also evaluate how well the activities are performed. In this study, each of six(6) participants performed dumbbell lifts in five (5) different ways; (A - perfect, B - elbow.front, C - lifting.halfway, D - lowering.halfway, E - hips.front). The goal of this project is to predict which way an activity was performed. The data was captured with the aid of sensors on the belt, forearm, arm, and dumbbell.

The generalized booting model is utilized for the purpose of predicting how an activity was performed. The model is fit to the data set. The out-sample prediction accuracy observed is 0.9692. The model is considered as suitable for predicting data from the same distribution.

Libraries and sources of data

The data for this project is obtained from http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har. The caret, dplyr, gridExtra and doParallel libraries in R were used in this project.

Uploading libraries

```
suppressPackageStartupMessages({
        library(caret)
        library(dplyr)
        library(gridExtra)
        library(doParallel)
})
```

Uploading data

```
url1 <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
url2 <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

if(!file.exists("pml-training.csv") | !file.exists("pml-testing.csv")){
        download.file(url1, "./pml-training.csv")
        download.file(url2, "./pml-testing.csv")
}

train.data <- read.csv("./pml-training.csv")
test.data <- read.csv("./pml-testing.csv")
```

Exploratory data analysis

Two sets of data were provided for the project. One set was intended for training the model and the other for testing the model. The data contained measurements of the activities of six participants who performed dumbbell lifts in five different ways. The training data set contained 19622 observations and 160 variables;

```
dim(train.data)
```

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

and the testing data set contained 20 observations and 160 variables

```
dim(test.data)
```

```
## [1]  20 160
```

Data preparation and transformations

During the exploratory data analysis it was discovered that most of variables (columns) either had no values available or the values were less than 50% of the observations. Only complete or otherwise, columns with more than 50% of observations were selected.

```
train.data <- select(train.data, roll_belt:total_accel_belt, gyros_belt_x:magnet_belt_z,
                     roll_arm:total_accel_arm, gyros_arm_x:magnet_arm_z, roll_dumbbell:yaw_dumbbell,
                     total_accel_dumbbell, gyros_dumbbell_x:magnet_dumbbell_z, roll_forearm:yaw_forearm,
                     total_accel_forearm, gyros_forearm_x:magnet_forearm_z, classe)

test.data <- select(test.data, roll_belt:total_accel_belt, gyros_belt_x:magnet_belt_z,
                     roll_arm:total_accel_arm, gyros_arm_x:magnet_arm_z, roll_dumbbell:yaw_dumbbell,
                     total_accel_dumbbell, gyros_dumbbell_x:magnet_dumbbell_z, roll_forearm:yaw_forearm,
                     total_accel_forearm, gyros_forearm_x:magnet_forearm_z)
```

The column names were inspected for spaces and all underscores "_" replaced with the period “.”

```
names(test.data) <- gsub("_", ".", names(test.data))
names(train.data) <- gsub("_", ".", names(train.data))
names(test.data) <- gsub(" ", ".", names(test.data))
names(train.data) <- gsub(" ", ".", names(train.data))
train.data$classe <- as.factor(train.data$classe)
```

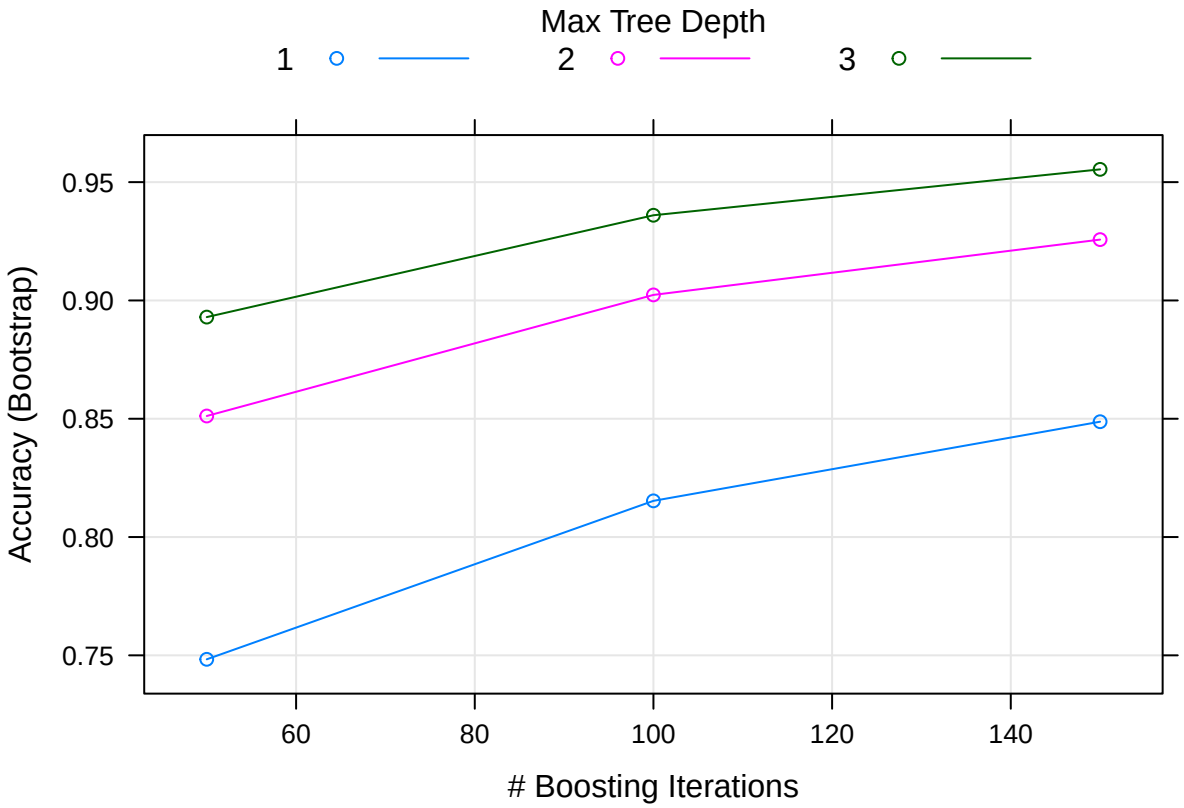
Modeling methodology

The training data provided was further partitioned into two as new training and validation data sets. The validation set was one-third of the original training data set.

```
set.seed(86479)
inTrain <- createDataPartition(train.data$classe, p = 3/4, list = FALSE )

training <- train.data[inTrain, ]
validate <- train.data[-inTrain, ]
```

The model was built out of the training data set using the generalized boosting model.

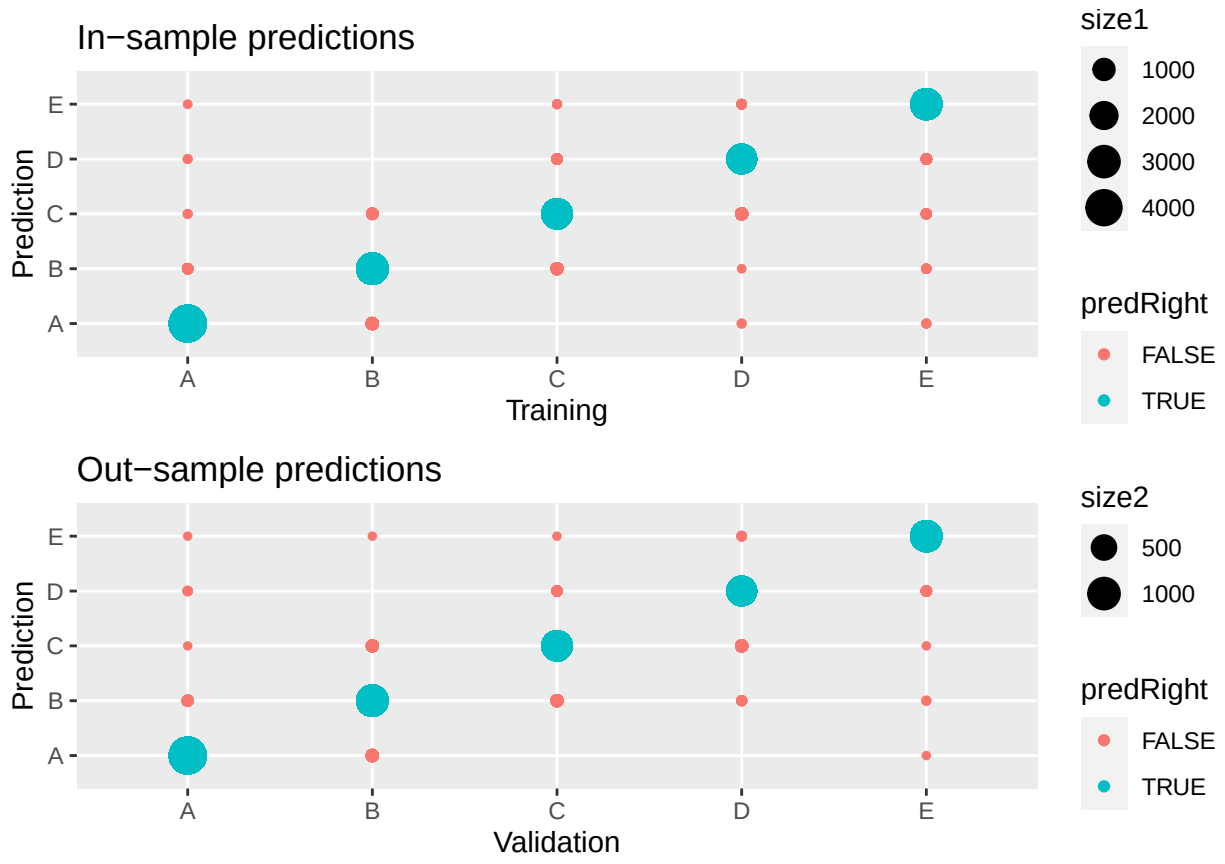
```
cl <- makePSOCKcluster(15)

registerDoParallel(cl)
mod2 <- train(classe ~., data = training, method = "gbm", verbose = FALSE)
stopCluster(cl)

plot(mod2)
```



The model was subsequently used to predict on the validation set and the in-sample and out-sample errors examined.

```
pred_mod2.In <- predict(mod2, training)
pred_mod2.Out <- predict(mod2, validate)

training$predRight <- pred_mod2.In == training$classe

size1 <- sapply(1:length(training$classe), function(i) { sum(training$classe == training$classe[i] & pr

plot1 <- qplot(classe, pred_mod2.In, colour = predRight, data = training,
      main = "In-sample predictions", xlab = "Training", ylab = "Prediction", cex = size1)

validate$predRight <- pred_mod2.Out == validate$classe

size2 <- sapply(1:length(validate$classe), function(j) { sum(validate$classe == validate$classe[j] & pr
```

```
plot2 <- qplot(classe, pred_mod2.Out, colour = predRight, data = validate,
      main = "Out-sample predictions", xlab = "Validation", ylab = "Prediction", cex = size2)

grid.arrange(plot1, plot2, ncol = 1)
```



See confusion matrix and summary of statistics below;

```
confusionMatrix(pred_mod2.Out, validate$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1376   26    0    0    1
##          B   14  899   25    6    2
##          C    1   23  820   24    1
##          D    3    0    9  770    9
##          E    1    1    1    4  888
##
## Overall Statistics
##
##                Accuracy : 0.9692
##                  95% CI : (0.964, 0.9739)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2e-16
```

```
##
##                   Kappa : 0.961
##
##  Mcnemar's Test P-Value : 0.01168
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9864   0.9473   0.9591   0.9577   0.9856
## Specificity            0.9923   0.9881   0.9879   0.9949   0.9983
## Pos Pred Value         0.9808   0.9503   0.9436   0.9735   0.9922
## Neg Pred Value         0.9946   0.9874   0.9913   0.9917   0.9968
## Prevalence             0.2845   0.1935   0.1743   0.1639   0.1837
## Detection Rate         0.2806   0.1833   0.1672   0.1570   0.1811
## Detection Prevalence   0.2861   0.1929   0.1772   0.1613   0.1825
## Balanced Accuracy      0.9893   0.9677   0.9735   0.9763   0.9919
```

Prediction

The model was applied on the 20 observations in the original test data and the result is as follows;

```
predict(mod2, test.data)
```

```
##  [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
```

Conclusion

The implemented and acceptable model for the data is the generalized booting model. The out-sample prediction accuracy was observed at 0.9692 at 95% confidence interval. The model is considered to be suitable for predicting. data from the same distribution.