Machine Learning Course Project

Michael C. Boudreau

April 8, 2017

Assignment Background

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. These type of devices are part of the quantified self movement - a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

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

Retrieving and Cleaning Data

For this project, I will load the data directly into the environment using the fread function and clean it up in preparation for performing our modeling and prediction sets.

```r
library(data.table)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.3.3
```

```r
training <- fread("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv",
                  na.strings=c("NA","#DIV/0!",""))
testing <- fread("https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv",
                 na.strings=c("NA","#DIV/0!",""))
```

To further prep the data, I take out the first 7 columns which represent the identifications of the test subjects as well as some timestamp data which would not affect the result. I also noticed there were many columns that either did not contain data or had very little data. This may lead to overfitting or overweighted outliers, so we filter out these columns as well.

```r
training <- data.frame(training[,-c(1:7)])
training <- training[,colSums(is.na(training))==0]
testing <- data.frame(testing[,-c(1:7)])
testing <- testing[,colSums(is.na(testing))==0]
```

Cross Validation

In order to make sure we can validate our model, we will fit our model on 70% of the training data and use the remaining 30% for cross validation.

```
inTrain <- createDataPartition(training$classe,p=0.7,list=F)
train <- training[inTrain,]
cv <- training[-inTrain,]
```

Models and Prediction

I will attempt to fit the data based on three common models; Random Forest, Boosting Trees, and Linear Discrimination Analysis. The three models will be used to predict the cross validation data set and will compare the accuracy of each model to choose the best option for predicting our test set results. Keep in mind that the random forest model, in particular, can take a while to run.

```
fitRF <- train(classe~.,data=train,method="rf")
```

```
## Warning: package 'randomForest' was built under R version 3.3.3
```

```
fitGBM <- train(classe~.,data=train,method="gbm")
```

```
## Warning: package 'gbm' was built under R version 3.3.3
```

```
fitLDA <- train(classe~.,data=train,method="lda")

predRF <- predict(fitRF,cv)
predGBM <- predict(fitGBM,cv)
predLDA <- predict(fitLDA,cv)
```

Here is a table of the accuracy for each model. As you can see, using a random forest method looks like our best choice to move forward with.

```
results <- data.frame("Accuracy" = c(confusionMatrix(predRF,cv$classe)$overall[1],
                                     confusionMatrix(predGBM,cv$classe)$overall[1],
                                     confusionMatrix(predLDA,cv$classe)$overall[1]),
                      row.names = c("Random Forest","Boosted Trees","Linear Discriminant Analysis"))
results
```

```
##                               Accuracy
## Random Forest                0.9932031
## Boosted Trees                0.9619371
## Linear Discriminant Analysis 0.6958369
```

Conclusion

As we learned in this course, random forests and boosting are two of the most reliable forms of predictive modelling. From the results here, we see that we achieved a high level of accuracy from these models with random forest perform at 99.3% accuracy. I will use this model to predict the outcomes of the test data. I have suppressed the outcome as it includes the answers to the project quiz.

```
predictions <- predict(fitRF,testing)
```

```
## Warning: package 'randomForest' was built under R version 3.3.3
```

```
predictions
```