Practical Machine Learning Course Project

Christian Milbank

November 30, 2018

Synopsis

The purpose of this project is to build a machine learning algorithm that will predict the type of barbell lift based on data from belt, forearm, arm, and dumbell accelerometers. Three different models were built and cross-validated, then the predictions were applied to the test data set.

Data Processing

The data was downloaded as separate training and testing data sets. Further, the training data set was split 50%/50% into training and cross-validation data sets, which were used to test the different models. Finally, we remove any redundant columns (columns with a large number of missing/null values).

We will also take this opportunity to load the "xgboost" and "caret" libraries which will be used in the analysis.

```
library(caret, lib.loc = "H:/Decision Support/Stats Jam/R files/Libraries")
```

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

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

```
library(xgboost, lib.loc = "H:/Decision Support/Stats Jam/R files/Libraries")

URLtrain <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-
training.csv"
URLtest <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-
testing.csv"

download.file(URLtrain, destfile = "pmltrain.csv")
download.file(URLtest, destfile = "pmltest.csv")

pmltrainraw <- read.csv("pmltrain.csv")
pmltest <- read.csv("pmltest.csv")

inTrain <- createDataPartition(y = pmltrainraw$classe, p = .5, list = FALSE)
pmltrain <- pmltrainraw[inTrain,]
pmlCV <- pmltrainraw[-inTrain,]
```

```
pmltrain <- pmltrain[,-c(1:5, 12:36, 50:59, 69:83, 87:101, 103:112, 125:139,
141:150)]
pmlCV <- pmlCV[,-c(1:5, 12:36, 50:59, 69:83, 87:101, 103:112, 125:139,
141:150)]
pmltest <- pmltest[,-c(1:5, 12:36, 50:59, 69:83, 87:101, 103:112, 125:139,
141:150)]
```

Exploratory Data Analysis

The next step is to perform some exploratory analysis on the data. Specifically, we will look at the target variable, "classe", to understand this outcome better.

```
summary(pmltrain$classe)
```

```
##    A    B    C    D    E
## 2790 1899 1711 1608 1804
```

Model Building

Now we will build three different models and see how they perform on the cross-validation data. The types of models we will examine are the following:

- Gradient Boosing (GBM)
- Random Forest
- xgboost

GBM

The first model to build is the GBM. We build the model on 50% of the data and use the remaining 50% to cross-validate. From the output below, we see that the expected out-of-sample error rate is **1.49%**.

```
set.seed(5)
GBM <- train(classe ~ ., method = "gbm", data = pmltrain, verbose = FALSE)
pmlCV$GBMpredict <- predict(GBM, newdata = pmlCV)
mean(pmlCV$GBMpredict == pmlCV$classe)
```

```
## [1] 0.9851172
```

Random Forest

The second model to examine is the random forest. We train and cross-validate the data in the same manner as the GBM above. From this output, we see that the expected out-of-sample error rate is **0.40%**.

```
set.seed(5)
RF <- train(classe ~ ., method = "rf", data = pmltrain)
```

```
pmlCV$RFpredict <- predict(RF, newdata = pmlCV)
mean(pmlCV$RFpredict == pmlCV$classe)

## [1] 0.9960245
```

xgboost

Finally, we also build and test the xgboost model in the same manner as the prior models. From the output below, we see that the expected out-of-sample error rate is **0.06%**.

```
set.seed(5)
XGB <- train(classe ~ ., method = "xgbTree", data = pmltrain)
pmlCV$XGBpredict <- predict(XGB, newdata = pmlCV)
mean(pmlCV$XGBpredict == pmlCV$classe)

## [1] 0.9993884
```

Results

From the analysis above we see that xgboost is the best performing model. We then apply this model to the test data to determine our predictions.

```
pmltest$XGBpredict <- predict(XGB, newdata = pmltest)
pmltest$XGBpredict

##  [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
```

In fact, if we apply our predictions from the GBM and random forest models to the test data we would notice that the predictions are exactly the same as the predictions under the xgboost model. This serves as a sanity check for our model, since all three agree on the predictions.