Practical Machine Learning Project - Weight Lifting Exercise Data

Jennifer Williams

January 31, 2016

Credits

Credits to http://groupware.les.inf.puc-rio.br/har#weight_lifting_exercises (http://groupware.les.inf.puc-rio.br/har#weight_lifting_exercises) and this paper about using parallel processing - https://github.com/lgreski/datasciencectacontent/blob/master/markdown/pml-randomForestPerformance.md (https://github.com/lgreski/datasciencectacontent/blob/master/markdown/pml-randomForestPerformance.md), both were very helpful.

Preliminary Discussion

I of course was procrastinating so if I had a lot of time, I would have run some different models using cross validation and maybe even stacked some models. I probably would have run a regular linear regression model first. After reviewing discussions though and the instructions, it appears as though a random forest is going to get you to the right results. And the random forest with k-fold cross validation is the way to go. I followed a similar procedure to greski's paper.

I opened up the training and testing data in Excel first to see what the data looked like. In the testing data set, there were several columns with no data or NA's in the data. I just removed all of these columns in both the training and testing data sets.

Load Data & Libraries & Set Seeds and Prerequisites

```
trainingimport <- read.csv("pml-training.csv")
traininguse <- trainingimport[,c(8:11,37:49,60:68,84:86,102,113:124,140,151:160)]

testingimport <- read.csv("pml-testing.csv")
testinguse <- testingimport[,c(8:11,37:49,60:68,84:86,102,113:124,140,151:159)]
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.2.3
```

```
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 3.2.2
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.2.3
```

```
## randomForest 4.6-12
## Type rfNews() to see new features/changes/bug fixes.
```

```
set.seed(1000)
y <- traininguse[,53]
x <- traininguse[,-53]
```

Step 1 Configure Parallel Processing

```
library(parallel)
library(doParallel)
```

```
## Warning: package 'doParallel' was built under R version 3.2.3
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.2.2
```

```
## Loading required package: iterators
```

```
## Warning: package 'iterators' was built under R version 3.2.2
```

```
cluster <- makeCluster(detectCores() - 1) # convention to leave 1 core for OS
registerDoParallel(cluster)
```

Step 2 Configure trainControl object

```
fitControl <- trainControl(method = "cv", number = 10, allowParallel = TRUE)
```

Step 3 Develop Training Model & de-register processing cluster

```
fit <- train(x,y, method="rf",data=traininguse,trControl = fitControl)
stopCluster(cluster)
predictions <- predict(fit,testinguse)
fit$finalModel
```

```
## 
## Call:
##  randomForest(x = x, y = y, mtry = param$mtry, data = ..1) 
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
## 
##         OOB estimate of  error rate: 0.44%
## Confusion matrix:
##      A    B    C    D    E  class.error
## A 5576    3    0    0    1 0.0007168459
## B   11 3784    2    0    0 0.0034237556
## C    0   19 3401    2    0 0.0061367621
## D    0    0   41 3173    2 0.0133706468
## E    0    0    0    5 3602 0.0013861935
```

```
predictions
```

```
##  [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
```

Out of Sample Error Estimate

The out of sample error estimate is 0.44% from the final model fit.

Prediction Quiz

I got 20 out of 20 correct on the quiz so I guess the model is working correctly!