Machine Learning Assignment

Created by Sambit Datta March 14, 2016 Synopsis ——— Given 2 files, we are asked to develop a machine learning paradigm to predict the type of exercise performed based an a variety of measurements. The first file is the training file, used to develop our algorithm, and the second is used to make our final predictions.

The training data is [1]:https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv) The testing data is: [2]:https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv)

Once downloaded to our working directory, we read in the file to perform some basic exploratory data analysis. I continue to examine and remove columns which contain NA's, as well as remove columns which I do not believe have any outcome on the class.

In order to remove the na :

```
a=read.csv('pml-training.csv',na.strings=c('','NA'))
b=a[,!apply(a,2,function(x) any(is.na(x)) )]
c=b[,-c(1:7)]
```

This leave us with 19622 observations and 53 predictors (one of which is the response variable)

To continue with the analysis we download the necessary packages

```
library('randomForest')
```

```
## Warning: package 'randomForest' was built under R version 3.1.3
```

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

```
library('caret')
```

```
## Warning: package 'caret' was built under R version 3.1.3
```

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

```
library('e1071')
```

```
## Warning: package 'e1071' was built under R version 3.1.3
```

For cross validation, We split our testing data into sub groups, 60:40

```
subGrps=createDataPartition(y=c$classe, p=0.6, list=FALSE)
subTraining=c[subGrps,]
subTesting=c[-subGrps, ]
dim(subTraining);dim(subTesting)
```

```
## [1] 11776    53
```

```
## [1] 7846   53
```

We see there are 11776 in the subTraining group, and 7846 in the subTesting group.

I then continue to make a predictive model based on the random forest paradigm, as it is one of the best performing, using the subTraining group. Once the model is made, we predict the outcome of the other group, subTesting, and examine the confusion matrix to see how well the predictive model performed

```
model=randomForest(classe~., data=subTraining, method='class')
pred=predict(model,subTesting, type='class')
z=confusionMatrix(pred,subTesting$classe)
save(z,file='test.RData')
```

```
setwd('C:/Users/user/Documents/Coursera')
load('test.RData')
z$table
```

```
##           Reference
## Prediction    A    B    C    D    E
##          A 2231    8    0    0    0
##          B    0 1503    5    0    0
##          C    0    7 1362   22    1
##          D    1    0    1 1263    9
##          E    0    0    0    1 1432
```

```
z$overall[1]
```

```
##  Accuracy
## 0.9929901
```

The accuracy is 99.31%. The out of sample error, that is the error rate on a new (subTesting) data set, here is going to be 0.69%, with a 95% confidence interval of 0.52% to .9%.

```
z$overall[1]
```

```
##  Accuracy
## 0.9929901
```

Final Data Set Analysis and Predictions

This is very good, so I continue with the final testing data set. I read it in and preporcess it the same way as the training set previously.

```
d=read.csv('pml-testing.csv',na.strings=c('','NA'))
e=d[,!apply(d,2,function(x) any(is.na(x)) )]
f=e[,-c(1:7)]
```

Once the dataset it processed, I continue to analyse it using the model developed above

```
predicted=predict(model,f,type='class')
save(predicted,file='predicted.RData')
```

The final prediction for the 20 ends up as:

```
setwd('C:/Users/user/Documents/Coursera')
load('predicted.RData')
p
```

```
##       (Intercept)        HTGD     RED.H      RED.A     POINTS_H
## Draw 1.998401e-14 2.08698e-05 0.5994287 0.39096550 0.0068282980
## Win  1.865619e-12 0.00000e+00 0.2755665 0.07201193 0.0001336174
##          POINTS_A  TOTAL_H_P    TOTAL_A_P        FGS.0        FGS.1
## Draw 2.455325e-02 0.95984215 0.0074179603 0.000000e+00 3.679546e-11
## Win  5.822547e-05 0.00718175 0.0002046679 8.881784e-16 8.841182e-09
```