Coursera Practical Machine Learning: Prediction Assignment Writeup

Michael Gao

April 22, 2016

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, the goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants to answer these questions.

The training data for this project are available here: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv)

The test data are available here: https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv)

The data for this project come from this source: http://groupware.les.inf.puc-rio.br/har (http://groupware.les.inf.puc-rio.br/har).

Cleaning Data

Columns containing NA data were removed, along with the first seven columns.

```
library(caret)
```

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

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

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.2.5
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
data_set_1 = read.csv('pml-training.csv',na.strings=c('','NA'))

data_set_2 = data_set_1[,!apply(data_set_1,2,function(x) any(is.na(x)) )]

data_set_3 = data_set_2[,-c(1:7)]
```

Cross validation and predictive model

For cross validation, We split the testing data into subgroups at a 60:40 ratio.

```
subGroups = createDataPartition(y=data_set_3$classe, p=0.6, list=FALSE)

subTraining = data_set_3[subGroups,]

subTesting = data_set_3[-subGroups,]
```

Random forest paradigm was used to make a predictive model with the subTraining group. We then predict the outcome with the subTesting group, and examine the confusion matrix to verify the predictive model performance.

```
model = randomForest(classe~., data=subTraining, method='class')

pred  = predict(model, subTesting, type='class')

confusionMatrix(pred, subTesting$classe)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    A    B    C    D    E
##          A 2229   14    0    0    0
##          B    3 1501    4    0    0
##          C    0    3 1363   18    0
##          D    0    0    1 1267   13
##          E    0    0    0    1 1429
## 
## Overall Statistics
## 
##                Accuracy : 0.9927
##                  95% CI : (0.9906, 0.9945)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
## 
##                   Kappa : 0.9908
##  Mcnemar's Test P-Value : NA
## 
## Statistics by Class:
## 
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9987   0.9888   0.9963   0.9852   0.9910
## Specificity            0.9975   0.9989   0.9968   0.9979   0.9998
## Pos Pred Value         0.9938   0.9954   0.9848   0.9891   0.9993
## Neg Pred Value         0.9995   0.9973   0.9992   0.9971   0.9980
## Prevalence             0.2845   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2841   0.1913   0.1737   0.1615   0.1821
## Detection Prevalence   0.2859   0.1922   0.1764   0.1633   0.1823
## Balanced Accuracy      0.9981   0.9938   0.9966   0.9915   0.9954
```

Testing Set Analysis and Predictions

Moving on to the testing data set.

```
data_set_4 = read.csv('pml-testing.csv', na.strings=c('','NA'))

data_set_5 = data_set_4[,!apply(data_set_4,2,function(x) any(is.na(x)) )]

data_set_6 = data_set_5[,-c(1:7)]

predicted=predict(model, data_set_6, type='class')

predicted
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  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
```

These are the final results for the Course Project Prediction Quiz.