# Wearable Movement Analysis

William Chan

Saturday, November 22, 2014

## 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.

The main task for this prediction project is to develop a statistical model to predict the following 5 classes of action:

### exactly according to the specification (A)

### throwing elbows to the front (B)

### lifting the dumbbell only halfway (C)

### lowering the dumbbell only halfway (D)

### throwing the hips to the front (E)

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).

## Data

The training data can be found at https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv While the test data is located at https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

```
training <- read.csv("pml-training.csv", na.strings=c("", "NA"))
testing <- read.csv("pml-testing.csv", na.strings=c("", "NA"))
```

The first column is index within the dataset, so we will remove it since it is not a predictor.

```
training$X <- NULL
```

In the process of anonymize the data with irrelevant dimension like the user name and the timestamp.

```
remove_cols <- c("user_name", "raw_timestamp_part_1", "raw_timestamp_part_2",
"cvtd_timestamp")
for(col in remove_cols) {
                training[, col] <- NULL
}
```

Furthermore, there is a lot of missing data within the variables. We can remove the dimension from the train and test dataset with too many missing values.

```
missing <- apply(training, 2, function(x) {sum(is.na(x))})
training <- training[, which(missing == 0)]
```

Upon further investigation of the activity data, the data appear to have value close to zero which will be remove using the near zero variance with the caret package.

```
library(caret)

## Warning: package 'caret' was built under R version 3.1.2

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

nsv <- nearZeroVar(training)
training <- training[-nsv]
testing <- testing[-nsv]
```

After cleaning the dataset, the final set of predictors used to build the classification algorithm are:

```
names(training)

##  [1] "num_window"           "roll_belt"            "pitch_belt"
##  [4] "yaw_belt"             "total_accel_belt"     "gyros_belt_x"
##  [7] "gyros_belt_y"         "gyros_belt_z"         "accel_belt_x"
## [10] "accel_belt_y"         "accel_belt_z"         "magnet_belt_x"
## [13] "magnet_belt_y"        "magnet_belt_z"        "roll_arm"
## [16] "pitch_arm"            "yaw_arm"              "total_accel_arm"
## [19] "gyros_arm_x"          "gyros_arm_y"          "gyros_arm_z"
## [22] "accel_arm_x"          "accel_arm_y"          "accel_arm_z"
## [25] "magnet_arm_x"         "magnet_arm_y"         "magnet_arm_z"
## [28] "roll_dumbbell"        "pitch_dumbbell"       "yaw_dumbbell"
## [31] "total_accel_dumbbell" "gyros_dumbbell_x"     "gyros_dumbbell_y"
## [34] "gyros_dumbbell_z"     "accel_dumbbell_x"     "accel_dumbbell_y"
## [37] "accel_dumbbell_z"     "magnet_dumbbell_x"    "magnet_dumbbell_y"
## [40] "magnet_dumbbell_z"    "roll_forearm"         "pitch_forearm"
## [43] "yaw_forearm"          "total_accel_forearm"  "gyros_forearm_x"
## [46] "gyros_forearm_y"      "gyros_forearm_z"      "accel_forearm_x"
## [49] "accel_forearm_y"      "accel_forearm_z"      "magnet_forearm_x"
## [52] "magnet_forearm_y"     "magnet_forearm_z"     "classe"
```

## Model Building

In building the predictive model, the classification method of random forest will be used to predict the activity class. In order to measure the model accuracy, a 10-fold K cross validation with 80:20 split will be implemented. The 80% of the data will be used for training while the remaining 20% will be used to test the model.

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.1.2
```

```
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

```
set.seed(1234)
obs <- c()
preds <- c()
for(i in 1:10) {
        inTrain = sample(1:dim(training)[1], size=dim(training)[1] * 0.8,
replace=F)
        train_cross = training[inTrain,]
        test_cross = training[-inTrain,]
        rf <- randomForest(classe ~., data=train_cross)
        obs <- c(obs, test_cross$classe)
        preds <- c(preds, predict(rf, test_cross))
}
```

To measure the accuracy of the model, a confusion matrix was created based on the random forest classification results.

```
conf_mat <- confusionMatrix(table(preds, obs))
conf_mat$table
```

```
##      obs
## preds     1     2     3     4     5
##     1 11092    13     0     0     0
##     2     0  7564    19     0     0
##     3     0     1  6799    44     0
##     4     0     0     0  6511     7
##     5     1     0     0     1  7198
```

The random forest model appears to provide a good classification for the 5 classes of actions. the accuracy is 99.78% and it only missed classified 86 cases. Then we can use the training model to predict the whole dataset given the activity measurements.

```
final_model <- randomForest(classe~., data=training)
```

## References

Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013. http://groupware.les.inf.puc-rio.br/har#ixzz3Jr74bA7g