Quantitative Model to Classify Barbell Lift Movements

Machine Learning - Prediction Assignment

Bryan Xu

January 29, 2017

Summary

We used the Weight Lifting Exercises Dataset[1] to train and compare several statistical models, which aim at classifying barbell lift movements based on quantitative data provided by sensors worn by exercisers. The most accurate model is Random Forest, which provide 100% identification rate on the training data with an estimated error of 0.38%.

[1. Reference http://groupware.les.inf.puc-rio.br/har#ixzz4XCjGGacm (http://groupware.les.inf.puc-rio.br/har#ixzz4XCjGGacm)]

1. Introdution

The advent of wearable personal fitness monitoring devices, such as *Fitbit, Jawbone Up, and Nike FuelBand*, allow a large amount of data to be collected. Using the data to analyse how much a wearer exercises has become a popular subject. The Weight Lifting Exercises Dataset is designed to answer a more sophisticated question: “how (well)” an activity was performed by the wearer.

Six participants were asked to perform the Unilateral Dumbbell Biceps Curl in five different fashions.

- Class A: exactly according to specification
- Class B: throwing the elbows to the front
- Class C: lifting the dumbbell only halfway
- Class D: lowering the dumbbell only halfway
- Class E: throwing the hips to the front

Class A corresponds to the “correct” execution of the exercise, while the other 4 classes correspond to common mistakes.

We used the dataset to build a model to identify the class of a movement based on the quantitative data collected by the wearable monitoring devices.

2. Description of the Weight Lifting Exercises Dataset

The dataset consists of two parts.

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)

```
    training <- read.csv("./pml-training.csv")
    testing  <- read.csv("./pml-testing.csv")
```

The training dataset has 160 columns and 19622 rows, i.e. observations. The test dataset has 160 columns and 20 observations, which will be used as the final test of the model.

3. Tidying Data

Columns 1 to 7 contain exercisers, time stamps and other information not relevant to movements. We'll omit these columns.

Many columns of the test dataset have only NAs. The corresponding columns of the training dataset have no useful data at all, or only a small amount of useful data. We'll omit these columns.

For example, Columns 11 to 14. We'll only keep column 11.

Training Dataset:

```
total_accel_belt kurtosis_roll_belt kurtosis_picth_belt kurtosis_yaw_belt
Min.   : 0.00             :19216             :19216            :19216
1st Qu.: 3.00    #DIV/0!  :   10    #DIV/0!  :   32     #DIV/0!:  406
Median :17.00    -1.908453:    2    47.000000:    4
Mean   :11.31    -0.016850:    1    -0.150950:    3
3rd Qu.:18.00    -0.021024:    1    -0.684748:    3
Max.   :29.00    -0.025513:    1    -1.750749:    3
                 (Other)  :  391    (Other)  :  361
```

Testing Dataset:

```
total_accel_belt kurtosis_roll_belt kurtosis_picth_belt kurtosis_yaw_belt
Min.   : 2.00    Mode:logical       Mode:logical        Mode:logical
1st Qu.: 3.00    NA's:20            NA's:20             NA's:20
Median : 4.00
Mean   : 7.55
3rd Qu.: 8.00
Max.   :21.00
```

After examine all columns, we kept Columns: 8:11, 37:49, 60:68, 84:86, 102, 113:124, 140, 151:160, and created new training and testing datasets.

```
keep <- c(8:11, 37:49, 60:68, 84:86, 102, 113:124, 140, 151:160)
new_training <- training[, keep]
new_testing <- testing[, keep]
```

4. Build and Compare Models.

We built several models using:

- R caret package
- K-fold Cross Validation with k = 5
- Classification methods: rpart, gbm, lda, svm, rf

The table below summarized the cross validation accuracy and corresponding 95% confidence intervals.

Method	Accuracy	95% Confidence Interval	Kappa
rpart	0.496	(0.489, 0.503)	0.341
gbm	0.975	(0.972, 0.977)	0.968
lda	0.705	(0.698, 0.711)	0.626
svm	0.945	(0.941, 0.948)	0.930
rf	1.000	(0.9998, 1.00)	1.000

The best model is Random Forest.

5. Final Model

Based on accuracy, we chose Random Forest as the final model. The estimated error rate is 0.38%.

```
model_rf_k$finalModel
```

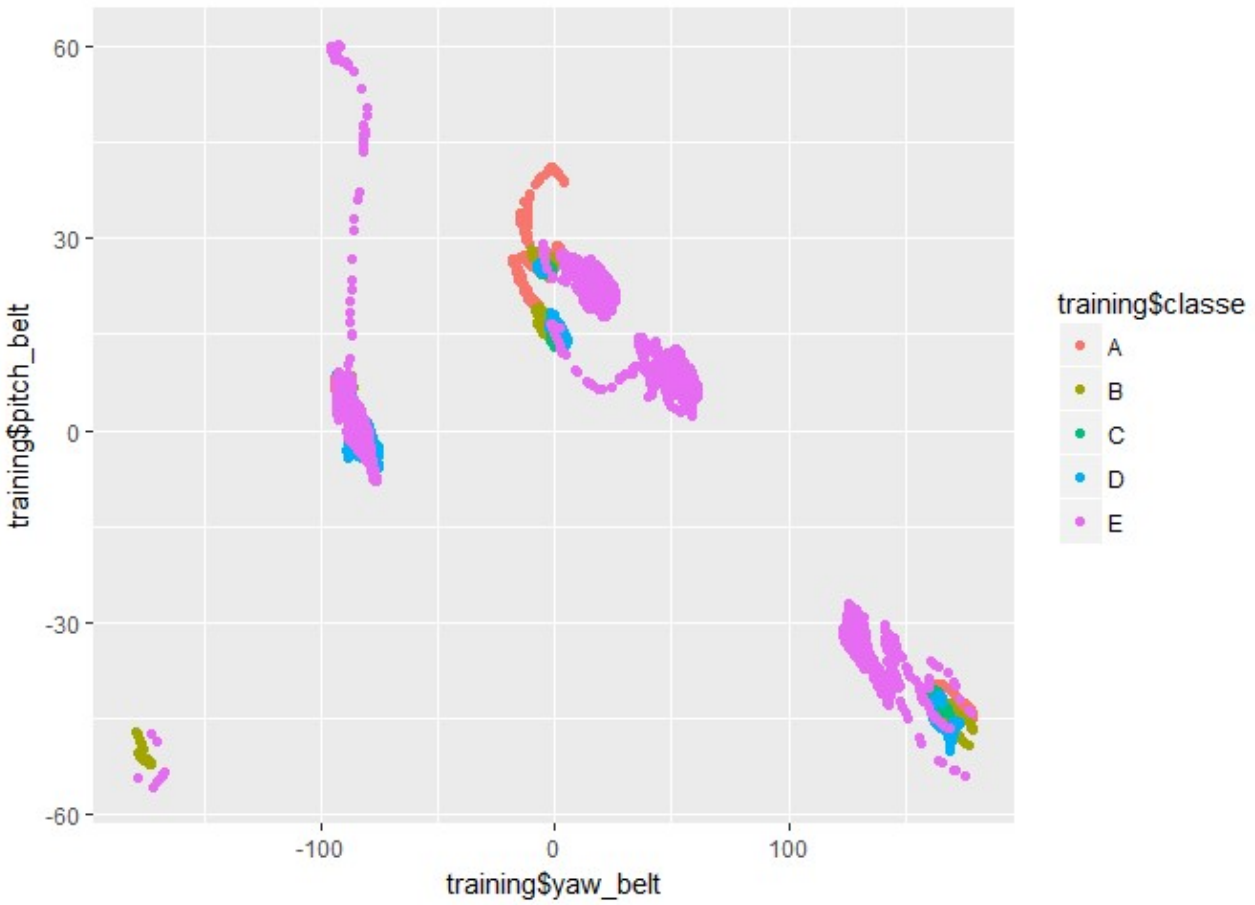
```
Call:
 randomForest(x = x, y = y, mtry = param$mtry, verbose = FALSE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0.38%
Confusion matrix:
     A    B    C    D    E  class.error
A 5578    2    0    0    0 0.0003584229
B   10 3784    3    0    0 0.0034237556
C    0   16 3405    1    0 0.0049678551
D    0    0   37 3178    1 0.0118159204
E    0    0    0    5 3602 0.0013861935
```

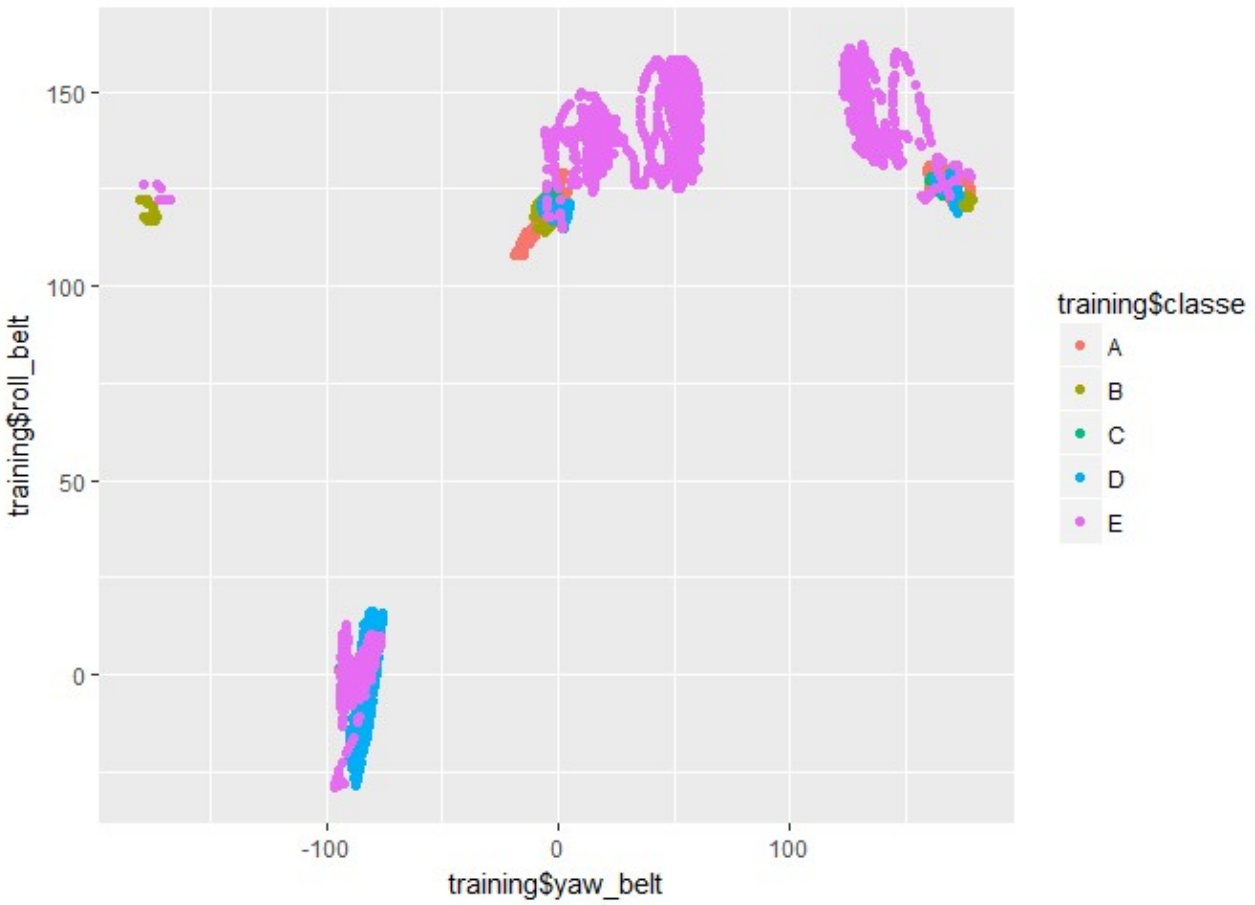
Using this model to classify the movement recorded in the test dataset,

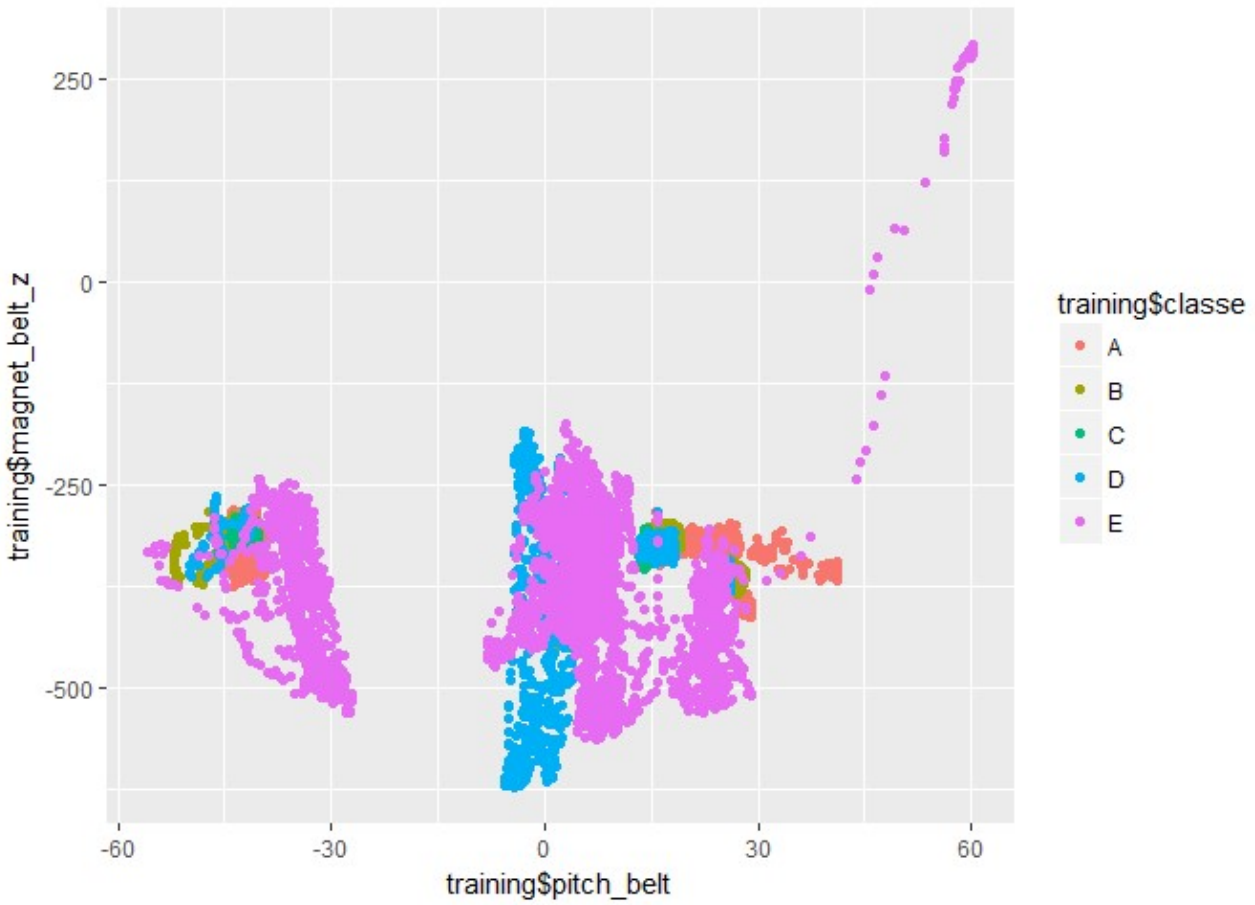
```
      predict_rf_k <- predict(model_rf_k, new_testing)
      predict_rf_k
```

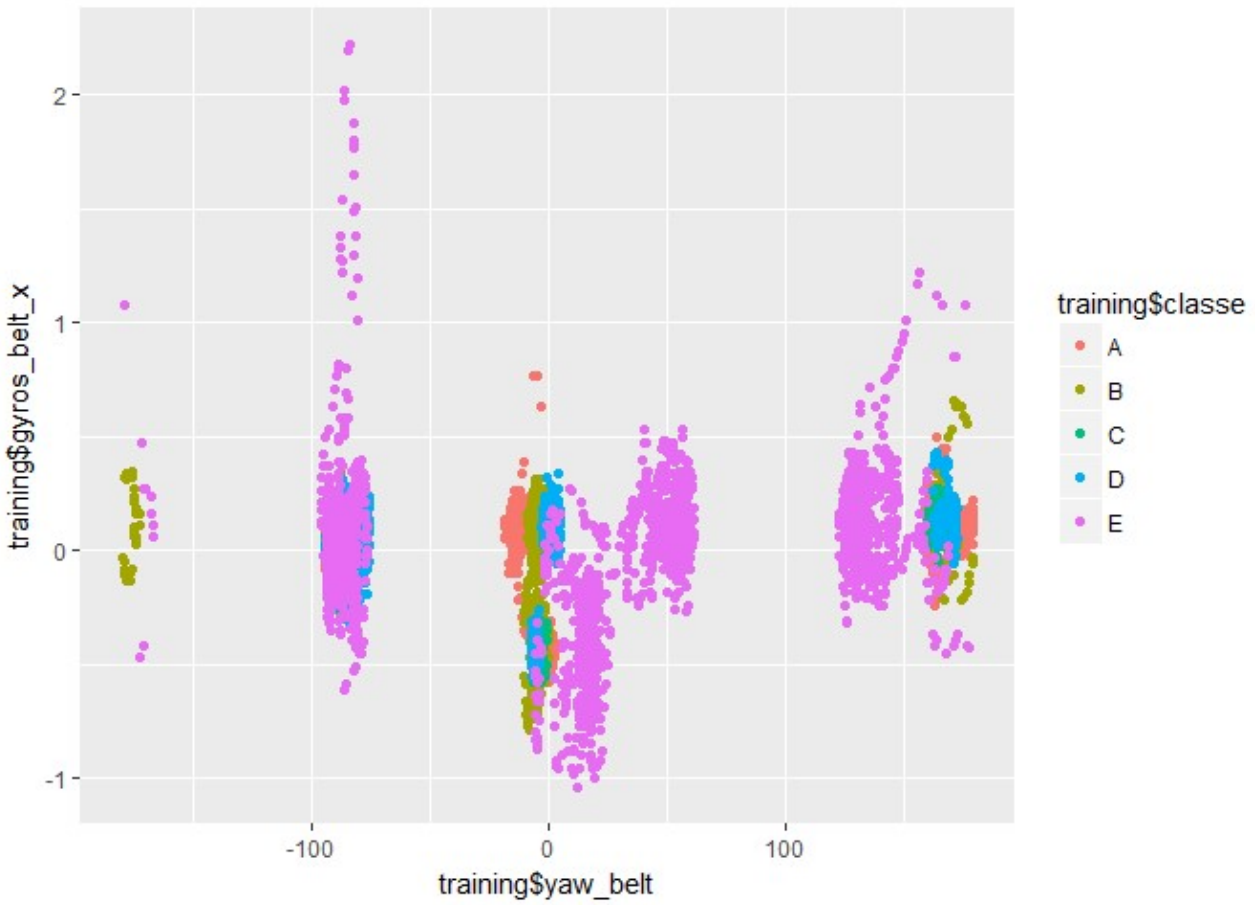
```
 [1] B A B A A E D B A A B C B A E E A B B B
Levels: A B C D E
```

A 100% accuracy was achieved.


60
30
0
-30
-60
training$pitch_belt
-100
0
100
training$yaw_belt
training$classe
A
B
C
D
E


150
100
50
0
training$roll_belt
-100
0
100
training$yaw_belt
training$classe
A
B
C
D
E


250
0
-250
-500
training$magnet_belt_z
-60
-30
0
30
60
training$pitch_belt
training$classe
A
B
C
D
E


training$gyros_belt_x
2
1
0
-1
-100
0
100
training$yaw_belt
training$classe
A
B
C
D
E