Practical Machine Learning

Julio Bolivar

10 May 2017

Background

The goal of this project is to predict the manner in some persons performed some exercises. This is the “classe” variable in the training set. We will consider 5 activity classes, gathered from 4 subjects wearing accelerometers mounted on their waist, left thigh, right arm, and right ankle.

Model Building

The data was collected per user and for some period of time. There are several features available. We should remove some features which are irrelevant for prediction. Some features are only relevant for aggregated values calculated for observations with new window = yes. They should be removed, since they are not present in the test set.

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

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

Cross Validation

We will try a random forest and perform cross validation by splicing our data into train and test sets.

```
## Loading required package: randomForest
```

```
## 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
## Random Forest
##
## 13453 samples
##    52 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 13453, 13453, 13453, 13453, 13453, 13453, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.9881585  0.9850221
##   27    0.9891320  0.9862528
##   52    0.9802073  0.9749625
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 27.
```

Expected Out of Sample Error

The trained model has high accuracy (98%). We can test its generalization by using the test set.

```
##
## pred    A    B    C    D    E
##    A 1637   10    0    0    0
##    B    1 1104    5    0    0
##    C    2    1  993   10    1
##    D    0    0    7  932    1
##    E    1    0    0    2 1056
```

We can indeed confirm that the model does not make many errors. The highest error number occur for predictions of classe D, where some predictions are incorrectly classifed as C.