Course Project

Matthew Dunne

October 18, 2017

Executive Summary

This analysis uses data from human activity recognition research. This data was collected to predict "which" activity was performed at a specific point in time. We will assess various predictive models for accuracy.

Choosing Variables and Pre-Processing

We will use **cross-validation** by using both a training set of data and a testing set of data. All models will be formed on the training data and tested for accuracy on the testing data.

We will not use all variables in the training data. This is for two reasons: (1) some variables are not in the test data and hence have no predictive value for that data, and (2) more variables in the model does not always increase accuracy.

We will pre-process the data with Principal Component Analysis. This is because there are a large number of variables, not all of which add predictive value. Based on a comparison of the accuracy of various models, the best threshold appears to be 80%.

Establishing Accuracy

In order to calculate out of sample error, we must first establish what the relevant classe variables are for the test data. We cannot say how accurate the prediction is if there are no true values against which to compare it. In its raw form, the testing data does not contain values for class. For reasons discussed above, I did not use username and time stamp data to predict class. However, for purposes of establishing out of sample error (i.e. the classe values for the test data) these are the best predictors.

The values we will regard as true values for purposes of calculating **out of sample error** are:

```
##  [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
```

Choice of Model

We will combine three different models: CART (rpart), random forest, and linear discriminant analysis. A generalized boosting model, while possibly useful, would take too much time for purposes of this analysis.

The predictions for rpart are:

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

with an accuracy of:

```
## [1] 0.5
```

The predictions for random forest are:

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

with an accuracy of:

```
## [1] 0.5
```

The predictions for linear discriminant analysis:

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

with an accuracy of:

```
## [1] 0.6
```

The predictions for the combined model are:

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

and the accuracy for this model is:

```
## [1] 0.8
```

Summary

A combined model analysis using rpart, random forest, and linear discriminant analysis is more accurate than each model individually. Based on our assumptions of what the true values of classe are in the test set, we see an accuracy (out of sample error) of 80%.