Predictions using the Weight Lifting Exercises Dataset

Mohammed

Hosni

August-2018

Executive Summary

Based on a dataset provide by HAR http://groupware.les.inf.puc-rio.br/har (http://groupware.les.inf.puc-rio.br/har) we will try to train a predictive model to predict what exercise was performed using a dataset with 159 features

We'll take the following steps:

- Process the data, for use of this project
- Explore the data, especially focussing on the two paramaters we are interested in
- Model selection, where we try different models to help us answer our questions
- Model examination, to see wether our best model holds up to our standards
- A Conclusion where we answer the questions based on the data
- Predicting the classification of the model on test set

Processing

First change 'am' to factor (0 = automatic, 1 = manual) And make cylinders a factor as well (since it is not continious)

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

Exploratory data analyses

Look at the dimensions & head of the dataset to get an idea

```
# Res 1
dim(training.raw)
```

```
## [1] 19622 160
```

```
# Res 2 - excluded because excessivness
# head(training.raw)

# Res 3 - excluded because excessivness
# str(training.raw)

# Res 4 - excluded because excessivness
# summary(training.raw)
```

What we see is a lot of data with NA / empty values. Let's remove those

```
maxNAPerc = 20
maxNACount <- nrow(training.raw)/100 * maxNAPerc
removeColumns <- which(colSums(is.na(training.raw) | training.raw=="")>maxNACount)
training.cleaned01 <- training.raw[,-removeColumns]
testing.cleaned01 <- testing.raw[,-removeColumns]
```

Also remove all time related data, since we won't use those

```
removeColumns <- grep("timestamp", names(training.cleaned01))
training.cleaned02 <- training.cleaned01[,-c(1, removeColumns)]
testing.cleaned02 <- testing.cleaned01[,-c(1, removeColumns)]
```

Then convert all factors to integers