Weight Lifting Exercise Analysis

Christoph Fabianek

Thursday, August 20, 2015

Overview

This project investigates data collected during weight lifting exercises and applys a machine learning algorithm from the CARET Package of the R programming language to predict the manner in which exercises were performed. This report was written for the course *Practical Machine Learning* of the *Coursera Data Science Specialization*.

Introduction

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. In this project, the goal is to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://groupware.les.inf.puc-rio.br/har.

Data Processing

First the underlying training and test data are downloaded from the web and read.

```r
# load libraries
library(caret)

# load & read data
setwd("~/Documents/coursera/dataScience/MachineLearning")
if(!file.exists("data")) {
        dir.create("data")
}
if(!file.exists('./data/pml-training.csv')) {
        fileUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
        download.file(fileUrl,
                      destfile="./data/pml-training.csv",
                      method="curl")
        dateDownloaded_training <- date()
}
training <- read.csv("./data/pml-training.csv", header = TRUE)

if(!file.exists('./data/pml-testing.csv')) {
        fileUrl <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
        download.file(fileUrl,
                      destfile="./data/pml-testing.csv",
                      method="curl")
        dateDownloaded_test20 <- date()
}
test20 <- read.csv("./data/pml-testing.csv", header = TRUE)
```

For the scope of this analysis the data is cleaned in the following way:

- remove the first 7 columns (`X`, `user_name`, `time_stamps`, `*_window`) since they are not relevant for classification
- remove columns with over 60% NAs
- remove neear zero variance predictors
- convert `classe` into a factor variable

```r
# remove first 7 columns
training <- training[, 8:ncol(training)]

# remove columns with >60% NAs
NAs <- apply(training, 2, function(x) {sum(is.na(x))})
training <- training[, which(NAs < nrow(training)*0.6)]

# remove near zero variance predictors
NZVs <- nearZeroVar(training, saveMetrics = TRUE)
training <- training[, NZVs$nzv == FALSE]

# convert classe into factor
training$classe <- factor(training$classe)
```

Afterwards the dataset is split into a 60% training and a 40% testing set.

```r
set.seed(210777)
trainset <- createDataPartition(training$classe, p = 0.6, list = FALSE)
data_training <- training[trainset, ]
data_testing <- training[-trainset, ]
```