CODE BOOK

Getting and Cleaning Data Course Project

Liliana Braescu

Problem Description

The unprecedented innovation in data science combined with big amount of data driven from mobile and sensor-driven applications brought new insights into wearable computing applications. To capture the state of the user and its environment, sensors can be attached on the subject body for continuous monitoring of numerous physiological signals.

A novel hardware-friendly approach with reduced costs in terms of energy and computational power was reported by Jorge L. Reyes-Ortiz and his team (Ref. [1]), which used waist-mounted smartphone (Samsung Galaxy S II) with embedded inertial sensors to record data from 30 volunteers with ages between 19-48 years old. The "Human Activity Recognition Using Smartphones Data Set" was built from recordings of those 30 subjects while performing six daily activities: walking, walking-upstairs, walking-downstairs, sitting, standing and laying. The embedded accelerometer and gyroscope captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50HZ. The obtained dataset has been randomly partitioned into two sets: 70% of the volunteers were selected for generating training data, and the remained 30% subjects generated the test data.

For each record, authors provided: (i) tri-axial acceleration from the accelerometer (total acceleration) and the estimated body acceleration; (ii) tri-axial angular velocity from the gyroscope; (iii) a 561-feature vector with time and frequency domain variables; (iv) activities labels; and (v) an identifier of the subject who carried out the experiment.

Project Purpose

Starting from the "Human Activity Recognition Using Smartphones Data Set" (https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip), student should demonstrate ability to collect, work with and clean data set.

For this aim, the R script "run_analysis.R" was created to perform data preparation, as well as the 5 steps required in the project course definition.

Data Preparation

Download the dataset

- Folder "Samsungdata" was created in the current working directory for downloading zip file - called "Samsung.zip".

- Unzip the file “Samsung.zip” into the exit directory “Samsungdata”. The unzipped data are in the folder “UCI HAR Dataset” which contains “test” and “train” folders, together with the files README.txt, features_info.txt, features.txt, and activity_labels.txt.

Read the dataset

Assign each data to variables:

- features <- features.txt with 561 rows and 2 columns
- activities <- activity_labels.txt with 561 rows and 2 columns
- subject_test <- subject_test.txt with 2947 rows and 1 column (test set)
- x_test <- x_test.txt with 2947 rows and 561 columns
- y_test <- y_test.txt with 2947 rows and 1 column
- subject_train <- subject_train.txt with 7352 rows and 1 column (train set)
- x_train <- x_train.txt with 7352 rows and 561 columns
- y_train <- y_train.txt with 7352 rows and 1 column

Steps Required in the Project

1. **Merge the training and the test sets to create one data set**
   - x_train and x_test are merged using **rbind()** function; the obtained “x” has dimension (10299, 561).
   - y_train and y_test are merged using **rbind()** function; the obtained “y” has dimension (10299, 1).
   - subject_train and subject_test are merged using **rbind()** function; the obtained “subject” has dimension (10299, 1).
   - The above “x”, “y”, and “subject” are merged using **cbind()** function; the obtained “Merged_Data” has dimension (10299, 563).
   - Dimensions of the merged objects are checked; structure of the R object “Merged_Data” is printed.
2. **Extract measurements on the mean and standard deviation**
   The “mean_std” is extracted from the “Merged_Data” by subsetting. Columns named “subject” and “code” are selected, and the measurements for the “**mean**” and “**std**” (standard deviation) are generated.
3. **Use descriptive activity names to name the activities in the data set**
   All numbers of the column named “code” from the “mean_std” are replaced with the corresponding activity taken from the 2nd column of the “activities” variable.
4. **Appropriately labels the data set with descriptive variable names**
   For renaming labels from “Merged_Data”, function **gsub()** was used to substitute all old labels with new labels as following:
   - prefix "t" replaced by "time"
   - prefix "f" replaced by "frequency"
   - "Acc" replaced by "accelerometer"
   - "Gyro" replaced by "gyroscope"
   - "Mag" replaced by "magnitude"
   - "BodyBody" replaced by "Body"

- "code" column from Merged_data was replaced by "activity"
- new names of the entire data set “Merged_Data” are printed.

5. **Independent tidy data set with the average of each variable for each activity and each subject**
   The final “Tidy_Data” set was created by summarizing “Merged_Data” with labels from step 4 grouped by “subject” and “activity”; average (mean) of each variable for each activity and each subject was taken using **summarise_all()** function.
   A .txt file was created with **write.table()** using row.name=FALSE according to the instructions received for the course project submission.

References:

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012.