Getting and Cleaning Data course Assignment

Tidied Data Code Book for accelerometers from Samsung Galaxy S smartphone

Written by: Wang Rui Xian

Introduction:

The data is obtained from https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip and the full description of the data can be found at http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones .

This code book describes the general file characteristics of the tidied data, description of location of the variables.

The Data

The tidied data is stored in the file “tidy.txt”.

Upon reading the file with read.table in R, a dataframe containing the data can be obtained, with the variables as the column headers and the respective data in the rows.

Variables:

The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ.

These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise. Similarly, the acceleration signal was then separated into body and gravity acceleration signals (tBodyAcc-XYZ and tGravityAcc-XYZ) using another low pass Butterworth filter with a corner frequency of 0.3 Hz.

Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ). Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag).

Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).

'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

-mean’ is used to denote the mean value.

'std' is used to denote the standard deviation.

In the original data set,

File name	Representation	Observations	Variables
X_train.txt	Training set	7352	561
X_test.txt	Test set	2947	561
Y_train.txt	Activity labels for training	7352	1
Y_test.txt	Activity labels for test	2947	1
Subject_train.txt	Train subject labels	7352	1
Subject_test.txt	Test subject labels	2947	1

Processing:

In the compilation of this dataset, no missing values were found in the original dataset. Hence, there was no requirement to fill up on any missing values.

The data in the tidied dataset represents the mean values for the variables for every individual test subject, for every individual activity.

For example:

For the activity of "laying" there will be 30 separate rows, each representing the mean values of variables such as "tBodyAcc.mean…X, tBodyAcc.mean…Y", for each of the 30 test subjects.

As such, the tidied dataset has a total of 180 observations of 81 variables.