CodeBook for the course project for the course Getting and Cleaning Data (coursera)

Chahid Azarkan

March 6th, 2016

The original data that is used for this assignment is collected by Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio and Luca Oneto of Smartlab [1]. They have carried out experiments with a group of 30 volunteers with ages between 19-48 years. Each person performed six activities (Walking, Walking upstairs, Walking downstairs, Sitting, Standing, Laying) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

The following files were used to create the appropriate dataset (e.g. merged\_dataset.csv and mean\_variable\_subject\_activity.csv): 1. data/train/subject\_train.txt 2. data/train/X\_train.txt 3. data/train/Y\_train.txt 4. data/test/subject\_test.txt 5. data/test/X\_test.txt 6. data/test/Y\_test.txt 7. data/features.txt 8. data/activity\_labels.txt

The files 1., 2. and 3. were merged based on row number for all the subjects in the train dataset. The files 4., 5. and 6. were also merged based on row number for the test dataset. Thereafter the names of the columns of both the train and test datasets were changed by using file 7, any special characters were taken out. The activity labels were attached to the right rows based on file 8, only changed to lowercase letters and without special characters.

Both the train and test datasets were merged resulting in the file merged\_dataset.csv, with all the mean value and standard deviation of all the variables from the original dataset. Based on the file merged\_dataset.csv a mean was calculated for all the variables, per subject and activity resulting in the file mean\_arable\_subject\_activity.csv.

As mentioned before, the above stated variables in the the dataset 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).

These signals were used to estimate variables of the feature vector for each pattern:  
'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

*tBodyAcc-XYZ* tGravityAcc-XYZ *tBodyAccJerk-XYZ* tBodyGyro-XYZ *tBodyGyroJerk-XYZ* tBodyAccMag *tGravityAccMag* tBodyAccJerkMag *tBodyGyroMag* tBodyGyroJerkMag *fBodyAcc-XYZ* fBodyAccJerk-XYZ *fBodyGyro-XYZ* fBodyAccMag *fBodyAccJerkMag* fBodyGyroMag \*fBodyGyroJerkMag

This resulted in the following variables in the files merged\_dataset.csv and mean\_variable\_subject\_activity.csv:

*subjectid (=Unique ID of the subject)* activitylabel (= Name of the activity) *tBodyAccmeanX* tBodyAccmeanY *tBodyAccmeanZ* tBodyAccstdX *tBodyAccstdY* tBodyAccstdZ *tGravityAccmeanX* tGravityAccmeanY *tGravityAccmeanZ* tGravityAccstdX *tGravityAccstdY* tGravityAccstdZ *tBodyAccJerkmeanX* tBodyAccJerkmeanY *tBodyAccJerkmeanZ* tBodyAccJerkstdX *tBodyAccJerkstdY* tBodyAccJerkstdZ *tBodyGyromeanX* tBodyGyromeanY *tBodyGyromeanZ* tBodyGyrostdX *tBodyGyrostdY* tBodyGyrostdZ *tBodyGyroJerkmeanX* tBodyGyroJerkmeanY *tBodyGyroJerkmeanZ* tBodyGyroJerkstdX *tBodyGyroJerkstdY* tBodyGyroJerkstdZ *tBodyAccMagmean* tBodyAccMagstd *tGravityAccMagmean* tGravityAccMagstd *tBodyAccJerkMagmean* tBodyAccJerkMagstd *tBodyGyroMagmean* tBodyGyroMagstd *tBodyGyroJerkMagmean* tBodyGyroJerkMagstd *fBodyAccmeanX* fBodyAccmeanY *fBodyAccmeanZ* fBodyAccstdX *fBodyAccstdY* fBodyAccstdZ *fBodyAccJerkmeanX* fBodyAccJerkmeanY *fBodyAccJerkmeanZ* fBodyAccJerkstdX *fBodyAccJerkstdY* fBodyAccJerkstdZ *fBodyGyromeanX* fBodyGyromeanY *fBodyGyromeanZ* fBodyGyrostdX *fBodyGyrostdY* fBodyGyrostdZ *fBodyAccMagmean* fBodyAccMagstd *fBodyBodyAccJerkMagmean* fBodyBodyAccJerkMagstd *fBodyBodyGyroMagmean* fBodyBodyGyroMagstd *fBodyBodyGyroJerkMagmean* fBodyBodyGyroJerkMagstd \*activitynumber (=Activity number, not present in the file mean\_variable\_subject\_activity.csv)

[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