CODEBOOK - SIGNALS FROM SAMSUNG GALAXY S II (2012)

<u>Raw data taken as input</u>

- X_test.txt

  2947 estimations of 561 variables
- X_train.txt

  7352 estimations of 561 variables
- features.txt

  Names of the features observed (561)
- subject_test.txt

  Each row (2947) corresponds to the ID of the subject that was carrying the cell phone and the correspondent estimations are reported in X_test.txt
- subject_train.txt

  Each row (7352) corresponds to the ID of the subject that was carrying the cell phone and the correspondent estimations are reported in X_train.txt

Introduction from the README.txt and features info.txt provided by the original experiments [1]

The experiments have been carried out with a group of 30 volunteers within an age bracket of 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 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).

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

The set of variables that were estimated from these signals are:

mean(): Mean value

std(): Standard deviation

mad(): Median absolute deviation

max(): Largest value in array

min(): Smallest value in array

sma(): Signal magnitude area

energy(): Energy measure. Sum of the squares divided by the number of values.

iqr(): Interquartile range

entropy(): Signal entropy

arCoeff(): Autorregresion coefficients with Burg order equal to 4

correlation(): correlation coefficient between two signals

maxInds(): index of the frequency component with largest magnitude

meanFreq(): Weighted average of the frequency components to obtain a mean frequency

skewness(): skewness of the frequency domain signal

kurtosis(): kurtosis of the frequency domain signal

bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.

angle(): Angle between to vectors.

Additional vectors obtained by averaging the signals in a signal window sample.

These are used on the angle() variable:

gravityMean

tBodyAccMean

tBodyAccJerkMean

tBodyGyroMean

tBodyGyroJerkMean

The complete list of variables of each feature vector is available in 'features.

txt'

Processing the data

All the estimations of the variables provided by the raw data were normalized and bounded within [-1,1], and so must be the values reported in the TidyData.txt as they are an average (see the info below).

1. The X_test.txt and X_train.txt data sets were merged in one data set, and every column was labeled with the name provided by the features.txt file.
2. As mentioned above, each of the 561 columns in the X_test.txt and X_train.txt data sets corresponds to an estimation. Two of these kind of estimations are the mean [mean()] and the standard deviation [std()], which were extracted from the data set obtained from the last step, getting a total of 66 columns (33 for the mean and 33 for the standard deviation). The names of the columns were changed for a more readable fashion:
    - Truncated words were completed
    - f substituted by “Fourier”
    - t substituted by “Time”

- “Mean” or “SD” added at the start of the name, depending on the case

3. Finally, each row was labeled with the corresponding subject who performed the activity, and the columns were averaged according to the subject. This output contains 30 rows (plus 1 with the names of the columns) each one corresponding to exactly 1 subject, and 66 columns (plus 1 indicating the subject ID). This is the TidyData.txt reported in the repository.