Codebook

D. Verdes

Sunday, March 22, 2015

# Project Description

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.

## Study design

### This info is taken from the "readme" and "feature info" files attached to the raw data files!

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.

### Description how the raw data is pre-processed

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 raw data contains the following features recorded in separate files:

* Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
* Triaxial Angular velocity from the gyroscope.
* A 561-feature vector with time and frequency domain variables.
* Its activity label.
* An identifier of the subject who carried out the experiment.

### The units of the raw data sets:

The feature are normalized and bounded within [-1,1]

# Creating the tidy file

A longer description of how the data is loaded, read and processed to creat a tidy data (data.frame) or file is given in the uploaded ReadMe file.

Here just a short description is presented!

* The "train" (70% of volunteers) data set and "test" (30% of volunteers) data set are loaded in XX\_train and XX\_test, respectively. They contain the measured and pre\_processed features.
* The activities performed are loaded in YY\_train and YY\_test, respectively. The activity is coded in numbers and the assignment is given in activity\_labels.
* The activity\_labels data frame containg the given activities
* the name of the variables is loaded in the "features" data frame.

To create a tidy data file the following steps are performed:

Step 1:

* Reads "X\_train.txt" in "XX\_train" data frame
* Reads "Y\_train.txt" in "YY\_train" data frame
* Reads "X\_test.txt" in "XX\_test" data frame
* Reads "Y\_test.txt" in "YY\_test" data frame
* Reads "features.txt" in "features" data frame
* Reads "subject\_train.txt" in "subj\_train" data frame
* Reads "subject\_test.txt" in "subj\_test" data frame
* Performs a "column bind" for "XX\_train", "YY\_train" and "subj\_train" to obtain "training\_set" data frame
* Performs a "column bind" for "XX\_test", "YY\_test" and "subj\_test" to obtain "testing\_set" data frame
* Merges the "training\_set" with "testing\_set" to obtain a full set of data in "all\_combined" data frame

Step 2:

* Extracts only the measurements on the mean and standard deviation for each measurements. This data are contained in a data frame called "all"

# Description of the variables in "all" tidy data set: Size (10299-observations) X (81 variables)

* first variable "subject" (class: integer) contains the number of the subjects in the experiment (30)
* second variable "activity" (class: factor) contains the type of activity performed: WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING

### the remaining variables (3:81),(class: numeric) have more combined names depending of the type of measurements:

#### every word used to describe the variables names is described below:

#### the variables names start with "time\_" or "frequency\_" to denote whether the signals are measured in time domain or "suffered" a Fourier tranformation to "frequency" domain

* the "XYZ" denotes the 3-axial signals measured in the X,Y and Z directions
* the "time\_" shows the features measured in time (constant rate of 50Hz); so called time domain signals (40 variables in the data set)
* the "frequency\_" shows the features tranformed in frequency domain by "fast fourier transformation", (30 variables in the data set)
* the "mean\_" stands for the Mean value
* the "std\_" stands for the Standard Deviation
* the "acceleration\_" stands for the data coming from the accelerometer
* the data from accelerometer are separated in "body\_acceleration\_" and "gravity\_acceleration\_"
* Example: "time\_body\_acceleration\_mean\_X" or "time\_gravity\_acceleration\_std\_X"
* the "gyroscope\_" stands for data coming from the gyroscope
* the data from gyroscope are separated in "body\_gyroscope\_" and "body\_body\_gyroscope"
* the "jerk\_rate-" stands for the rate of change for acceleration linear and angular
* the "acceleration\_jerk\_rate\_" stands for the body linear acceleration changes in time
* the "gyroscope\_jerk\_rate\_" stands for the angular velocity changes in time
* the "magnitude\_" stands for the signal calculated using "the Euclidian norm" as described in the "feature info"

Source: the name's descriptions above for the variables in the tidy data set are inspired from the description provided with the raw data sets: features\_info.txt and Readme.txt

ReadMe: the full description of the source code "run\_analysis.R" to obtain the tidy data is presented in the ReadMe file attached to this codebook.