**CODE BOOK**

[**Getting and Cleaning Data**](https://class.coursera.org/getdata-014/) **Course project**

**STUDY DESIGN**

* **Introduction:**

This study is based on the “Human Activity Recognition Using Smartphones” Dataset Version 1.0, conducted by Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto inside of the Smartlab - Non Linear Complex Systems Laboratory from the DITEN university “ Università degli Studi di Genova” (contact address: [activityrecognition@smartlab.ws](mailto:activityrecognition@smartlab.ws), website: www.smartlab.ws)

* **Experimental conditions:**

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.

* **Data collection:**

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.

* **Included measures:**

From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See 'features\_info.txt' for more details.

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

|  |  |
| --- | --- |
| * 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 |

* **Included features:**

Sets of variables were estimated from these signals:

* 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 are used on the angle() variable:

|  |  |
| --- | --- |
| * gravityMean * tBodyAccMean * tBodyAccJerkMean | * tBodyGyroMean * tBodyGyroJerkMean |

* **Provided data:**
* - 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.
* **Included files:**

- 'README.txt'

- 'features\_info.txt': Shows information about the variables used on the feature vector.

- 'features.txt': List of all features.

- 'activity\_labels.txt': Links the class labels with their activity name.

- 'train/X\_train.txt': Training set.

- 'train/y\_train.txt': Training labels.

- 'test/X\_test.txt': Test set.

- 'test/y\_test.txt': Test labels.

The following files are available for the train and test data. Their descriptions are equivalent.

- 'train/subject\_train.txt': Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.

- 'train/Inertial Signals/total\_acc\_x\_train.txt': The acceleration signal from the smartphone accelerometer X axis in standard gravity units 'g'. Every row shows a 128 element vector. The same description applies for the 'total\_acc\_x\_train.txt' and 'total\_acc\_z\_train.txt' files for the Y and Z axis.

- 'train/Inertial Signals/body\_acc\_x\_train.txt': The body acceleration signal obtained by subtracting the gravity from the total acceleration.

- 'train/Inertial Signals/body\_gyro\_x\_train.txt': The angular velocity vector measured by the gyroscope for each window sample. The units are radians/second.

**DATA PROCESSING**

* **Data sets loading:**

Data were all loaded previous to processing. Path to the main directory was loaded in a dirpath character variable. Then all table are loaded using the read.table() function, and the path to the different file names included in the pattern= parameter, is indicated using the dir() function including the dirpath variable and a recursive= TRUE parameter. All data table are loaded in a variable corresponding to file name.

>dirpath <- "C:/Users/soso/Documents/Coursera/Cours 3/UCI HAR Dataset"

>features <- read.table(dir(dirpath, recursive=T, pattern="features.txt", full.name=T))

>activity\_labels <- (dir(dirpath, recursive=T, pattern="activity\_labels.txt", full.name=T))

>X\_train <- read.table (dir(dirpath, recursive=T, pattern="X\_train.txt", full.names=T))

>X\_test <- read.table (dir(dirpath, recursive=T, pattern="X\_test.txt", full.names=T))

>y\_train <- read.table (dir(dirpath, recursive=T, pattern="^y\_train.txt", full.names=T))

>subject\_train <- read.table (dir(dirpath, recursive=T, pattern="subject\_train.txt", full.names=T))

>y\_test <- read.table (dir(dirpath, recursive=T, pattern="^y\_test.txt", full.names=T))

>subject\_test <- read.table (dir(dirpath, recursive=T, pattern="subject\_test.txt", full.names=T))

>activity\_labels<- read.table (dir(dirpath, recursive=T, pattern="activity\_labels.txt", full.names=T))

* **Merge training and test sets and name the different columns**

Subject files subject\_train.txt and subject\_test.txt were merged by row binding rbind() in one dataset named subject, the same thing was done for the y\_train.txt and y\_test.txt files in a data set called Activity, and for the X\_train.txt and X\_test.txt files in a feature one.

>subject= rbind (subject\_train, subject\_test)

>activity= rbind(y\_train, y\_test)

>featureX= rbind(X\_train, X\_test)

Those three data sets are named using the colnames() function. The subject and activity one column datasets are named respectively “SUBJECT” and “ACTIVITY”. The featuresX columns are named using the features data sets second column “V2” still using the colnames() function. Then those three data sets are merged into one named mergedall by column binding using the cbind() function.

>colnames(subject) <- "SUBJECT"

>colnames(activity) <- "ACTIVITY"

>colnames(featureX)<- features$V2

>mergedall= cbind(subject, activity, featuresX)

* **Extracts only mean and SD columns**

The column containing “men” or “std” sting are extracted from the featureX data table using the grepl() function, and stocked in a variable called extracteddata. A final data set called finaldata, including those extracted data is created using the cbind() function.

>extracteddata<-featureX[,grepl("mean|std",names(featureX))]

>finaldata<- cbind(subject, activity, extracteddata)

* **Name the activities in the data set**

The activities included in the y\_test and y\_train data tables are coded according to a numeric correspondence which is included in the activity\_labels data table. In order to replace the numbers in the activity by their corresponding names we matched the numbers in the first column of activity dataset with the second of the activity\_labels one.

>activity[,1]= activity\_labels[activity[,1],2]

* **Label appropriately the data set**

Features in the finaldata table are named not properly. We Used the gsub() function to replace them. As a consequence Acc is replaced by Acceleration, Gyro by Gyroscope, Mag by Magnitude etc…

>names(finaldata)<-gsub("Acc", "Acceleration", names(finaldata))

>names(finaldata)<-gsub("Gyro", "Gyroscope", names(finaldata))

>names(finaldata)<-gsub("BodyBody", "Body", names(finaldata))

>names(finaldata)<-gsub("Mag", "Magnitude", names(finaldata))

>names(finaldata)<-gsub("^t", "Time", names(finaldata))

>names(finaldata)<-gsub("^f", "Frequency", names(finaldata))

>names(finaldata)<-gsub("tBody", "TimeBody", names(finaldata))

>names(finaldata)<-gsub("-mean()", "Mean", names(finaldata), ignore.case = TRUE)

>names(finaldata)<-gsub("-std()", "Standard deviation", names(finaldata), ignore.case = TRUE)

>names(finaldata)<-gsub("-freq()", "Frequency", names(finaldata), ignore.case = TRUE)

>names(finaldata)<-gsub("angle", "Angular", names(finaldata))

>names(finaldata)<-gsub("gravity", "Gravity", names(finaldata))

* **Create an independent tidy data set with the average of each variable for each activity and each subject**

To finish we used the subject id as a factor to calculate the mean of the measures for each activity and used the aggregate() function to create a data set named tidy\_data. Then we order this data set according to the activity and subject respectively. Using the write.table() function including an argument row.names=FALSE we recorded a “.txt” file names “tidy\_data.txt”

>finaldata[,1]<- as.factor(finaldata[,1])

>tidy\_data = aggregate(. ~SUBJECT+ ACTIVITY, finaldata, mean)

>tidy\_data<- order(finaldata$ACTIVITY, finaldata$SUBJECT)

>write.table(tidy\_data, file="tidy\_data.txt", row.names=FALSE)

**REFERENCES AND COPYRIGHTS**

* 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
* <http://www.insideactivitytracking.com/data-science-activity-tracking-and-the-battle-for-the-worlds-top-sports-brand/>
* http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

This dataset is distributed AS-IS and no responsibility implied or explicit can be addressed to the authors (Jorge L. Reyes-Ortiz, Alessandro Ghio, Luca Oneto, Davide Anguita. November 2012) or their institutions for its use or misuse.

Any commercial use is prohibited.