# Getting and Cleaning Data Project Codebook

1. Data Description
   A data description is provided by the authors of an experiment, which is summarized by an extract from the original paper, below:[1] The entire text describing the experiment is included in the repository containing this codebook.


**Abstract**.
Human-centered computing is an emerging research field that aims to understand human behavior and integrate users and their social context with computer systems. One of the most recent, challenging and appealing applications in this framework consists in sensing human body motion using smartphones to gather context information about people actions. In this context, we describe in this work an Activity Recognition database, built from the recordings of 30 subjects doing Activities of Daily Living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors, which is released to public domain on a well-known on-line repository. Results, obtained on the dataset by exploiting a multiclass Support Vector Machine (SVM), are also acknowledged.


2. Downloading the source data.

The source data, as well as decsriptions of the data set were downloaded from the course website on 20 Aug 2014 at:

https://class.coursera.org/getdata-006/human_grading/view/courses/972584/assessments/3/submissions

From there, a link was found to the following web site:
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

This file was unzipped, and it contained the following files and directories:

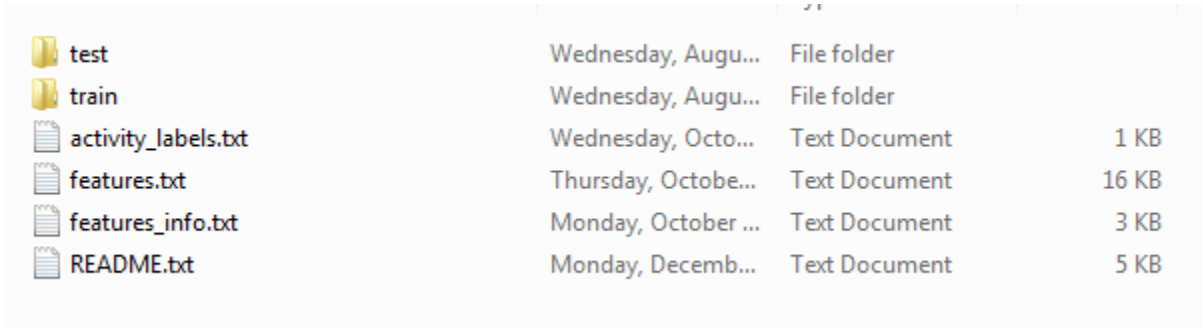


Figure 1 - Downloaded files and drectories

[1] ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.

3. The readme.txt file in FIgure 1 contains a description of the files contained in the download. The curcial part of the file descriptions is contained in Table 1.

**Table 1 - Readme.txt file contents**

activityrecognition@smartlab.ws
www.smartlab.ws
==========

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. See 'features_info.txt' for more details.

For each record it is provided:
==========

- 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 dataset includes the following 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.

Notes:
======
- Features are normalized and bounded within [-1,1].
- Each feature vector is a row on the text file.

For more information about this dataset contact: activityrecognition@smartlab.ws

License:
========
Use of this dataset in publications must be acknowledged by referencing the following publication [1]

[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



Jorge L. Reyes-Ortiz, Alessandro Ghio, Luca Oneto, Davide Anguita. November 2012.

4. As specified in Table 1, the features.txt file contains a list of features. A subset of that list is shown below in

```
1 tBodyAcc-mean()-X
2 tBodyAcc-mean()-Y
3 tBodyAcc-mean()-Z
4 tBodyAcc-std()-X
5 tBodyAcc-std()-Y
6 tBodyAcc-std()-Z
7 tBodyAcc-mad()-X
8 tBodyAcc-mad()-Y
9 tBodyAcc-mad()-Z
10 tBodyAcc-max()-X
11 tBodyAcc-max()-Y
12 tBodyAcc-max()-Z
…
```

**Table 2 – Subset of features.txt**

5. The features_info.txt file contains additional information about the features, and is shown in

```
Feature Selection
=================

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'

**Table 3 - features_info.txt**

6. The activity_labels.txt file contains activity IDs and their associated labels. The activity_labels.txt file is shown in Table 4.

1 WALKING
2 WALKING_UPSTAIRS

| |
|---|
| 3 WALKING_DOWNSTAIRS |
| 4 SITTING |
| 5 STANDING |
| 6 LAYING |

Table 4 - activty_labels.txt

7. We are instructed to create a single data set which merges the training and test data. This is done using R file copy commands, with the merged data file being created in a directory called “merged”.
    a. The merged output files were named subject_merged.txt, X_merged.txt, and y_merged.txt, to be consistent with the raw data.
    b. A sanity check file-length calculations showed that the merged file was the concatenation of the training and test files.
8. The y values in y_merged.txt containing an integer ID which refers to the type of activity being performed. The values in y_merged.txt should be appended as a column in the X_merged.txt data.
9. The python merge command was used to add descriptive activity names from the activity_labels.txt file, to each row in the data set.
10. The python aggregate command was used to take the mean of sub-populations grouped by subject_id and activity_id.
11. The tidy data set was stored in the file “avg_by_subject_activity.txt”.