Summarized Human Activity Recognition Using Smartphones Dataset

Original Datasource:
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip
Courtesy of:
Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto.
Smartlab Non Linear Complex Systems Laboratory
DITEN Università degli Studi di Genova.
Via Opera Pia 11A, I-16145, Genoa, Italy.

Experiment Background:

The experiments were 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, the researchers captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments were video-recorded to label the data manually. The researchers randomly partitioned the obtained dataset 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 was assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, the researchers obtained a vector of features calculating variables from the time and frequency domains.

Dataset

The dataset provided here (tidydata.txt,220 KB)is comprised of 180 observations of 68 variables extracted from the larger dataset provided by Reyes-Ortiz, Anguita, Ghio and Oneto. It presents means averaged for each activity and subject for those variables in the original dataset which were represents as means and standard deviations of features of interest. The data provided by the researchers was normalized and bounded within [-1,1], thus the derived data is also. These features (composing a total of 66 variables)are:

For time data (total of 40 variables):
1. Axial signals in the X, Y and Z directions for linear accelation both gravitational and body components (12 variables)
2. Magnitude of linear acceleration for both gravitational and body components (4 variables)
3. Axial signals in the X, Y and Z directions for angular velocity (6 variables)
4. Magnitude of angular velocity (2 variables)

5. Axial signals in the X, Y and Z directions for linear accelation derived in time for body component only (Note that these are the signals coded "Jerk" in the original dataset) (6 variables)
6. Magnitude of linear acceleration derived in time for body component only (coded "Jerk" in the original dataset) (2 variables)
7. Axial signals in the X, Y and Z directions derived in time for angular velocity (coded "Jerk" in the original dataset) (6 variables)
8. Magnitude of angular velocity derived in time (coded "Jerk" in the original dataset) (2 variables)

For frequency data (total of 26 variables):
9. Axial signals in the X, Y and Z directions for linear accelation body component only (6 variables)
10. Magnitude of linear acceleration for body component only (2 variables)
11. Axial signals in the X, Y and Z directions for angular velocity (6 variables)
12. Magnitude of angular velocity (2 variables)
13. Axial signals in the X, Y and Z directions for linear accelation derived in time for body component only (coded "Jerk" in the original dataset) (6 variables)
14. Magnitude of linear acceleration derived in time for body component only (coded "Jerk" in the original dataset) (2 variables)
15. Magnitude of angular velocity derived in time (coded "Jerk" in the original dataset) (2 variable)

This description of the data may be more clearly understood by means of the following table. (The file in this dataset "Table1.pdf" reproduces this table.)

Table 1: Variables provided in the dataset

		Linear Acceleration		Angular Velocity
		Body Component	**Gravity Component**	
filtered only	**axial signals along X,Y and Z axes**	time (6) frequency (6)	time (6)	time (6) frequency (6)
	magnitude of signal	time (2) frequency (2)	time (2)	time (2) frequency (2)
derived in time (coded "Jerk" by researchers)	**axial signals along X,Y and Z axes**	time (6) frequency (6)		time (6)
	magnitude of signal	time (2) frequency (2)		time (2) frequency (2)

- "time" and "frequency" in the body of the table indicate whether data was provided for that domain
- numbers in parentheses indicate the number of variable for condition
- for each case listed in the body of the table, a mean and a standard deviation have been provided

The dataset includes the following files:

- “README.md”: a text version of this document
- “README.pdf”: this document
- “codebook.txt”: provides a complete list of variable names and their descriptions
- “codebook.pdf”: a pdf version of "codebook.txt" displaying the contents in a table. Provided as an easily-readable alternative for the convenience of project graders.
- “Table1.pdf”: a graphic providing a visual summary of the features described in the document
- “tidydata.txt”: the dataset itself, provided as a comma delimited text file
- “run_analysis.R”: an R script that will generate "tidydata.txt", including downloading and unzipping the original dataset (assuming that it remains available at the specified URL). See The R Script below for more detail.

Notes:

1. To most easy inspect tidydata.txt in Excel, open Excel and File Open tidydata.txt, which should launch the Text Import Wizard. On the first screen of the wizard, select "Delimited." On the second page, be sure that only "Comma" is selected. You can leave "Treat consecutive delimiters as one" unchecked and set "Text qualifier" to ". On the third screen, select "General."
2. To easily inspect tidydata.txt in R, use read.table with header = TRUE, sep = "," and quote = "\""" to load file into R. Inspect using View in RStudio or the display commands of your choice if you are working at the command line.

The R Script

The included R script will generate a tidy dataset, "tidydata.txt", including downloading and unzipping the original dataset (assuming that it remains available at the specified URL).

The RScript "run_analysis.R" should run on any system with base R installed and internet access. It will download and unzip the original data set and create a tidy dataset as described in the README.txt and codebox.txt found with this script. To customize it for your own use, modify the data_path to the location in which you would like the data stored. It will create a /data subdirectory in that location if it does not already exist and download and extract the research data to the /data subdirectory. If you prefer to download and unzip manually, you may either comment out the relevent lines of code or ensure that your /data directory includes both a zip file named "UCI_HAR_Dataset.zip" and a subdirectory /UCI HAR Dataset containing the unzipped files in the file structure in which they were archived.

After downloading and unzipping data , the script creates a vector which identifies which variables from the original data set will be kept, a vector which contains key-value pairs for recoding activity data, based on the file activity_labels.txt in the original data file, and a vector of tidy variable names . The tidy variable names were created external to the R program. They are included in the R script to ensure that they are available to any user without the need to download an additional file that was not part of the original dataset, although it might be neater to include them as a separate file.

The function “cleanset” reads in observed data from the an original dataset (test or train), adds the appropriate subject and activity data, recodes activities, drops unneeded columns and adds tidy variable names. It is run twice by the script, once for the test data and once for the training data.

After clean datasets have been created for the test and training data, they are bound into one large table. “Aggregate” is used to average each variable for each activity and each subject. The aggregate variable names are tidied up and the final tidy dataset is written out to a comma delimited text file.

License: (reproduced from original dataset)

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

This dataset is distributed AS-IS and no responsibility implied or explicit can be addressed to the authors or their institutions for its use or misuse. Any commercial use is prohibited.