Code Book: Analysis of UCI HAR Dataset

Background

Data set: Human Activity Recognition Using Smartphones Data Set

Source: Jorge L. Reyes-Ortiz(1,2), Davide Anguita(1), Alessandro Ghio(1), Luca Oneto(1), Xavier Parra(2)

1 - Smartlab - Non-Linear Complex Systems Laboratory,
DITEN - Università degli Studi di Genova, Genoa (I-16145), Italy.

2 - CETpD - Technical Research Centre for Dependency Care and Autonomous Living,
Universitat Politècnica de Catalunya (BarcelonaTech). Vilanova i la Geltrú (08800),
Spain

(activityrecognition@smartlab.ws)

Subjects:

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.

Original Documentation (included with data set):

'README.txt'	Overview of data set.
'features_info.txt'	Information about variables in features vector.
'features.txt'	List of features.

Measurements

Data Acquisition Parameters

3-axial linear acceleration and 3-axial angular velocity.
Constant sampling rate (50 Hz).
Fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window).
Butterworth low-pass filter (???) with 0.3 Hz cut-off frequency to remove gravitational force.

Signals Acquired

The raw data apparently consisted of 6-channel recordings, supplemented by time derivatives (yielding jerk signals), and bandwidth-based splitting (to separate gravity from dynamics). From these signals, 10 time domain computations and 7 frequency domain computations were obtained. The resulting meta signals were:

#	Name in Source Data	Description
1.	tBodyAcc-XYZ	3-axis body acceleration (time domain)
2.	tGravityAcc-XYZ	3-axis gravity component (time domain)
3.	tBodyAccJerk-XYZ	3-axis body jerk (time domain)
4.	tBodyGyro-XYZ	3-axis body gyroscope (time domain)
5.	tBodyGyroJerk-XYZ	3-axis body gyroscope jerk (time domain)
6.	tBodyAccMag	body acceleration magnitude (time domain)
7.	tGravityAccMag	gravity acceleration magnitude (time domain)
8.	tBodyAccJerkMag	body jerk magnitude (time domain)
9.	tBodyGyroMag	body gyroscope magnitude (time domain)
10.	tBodyGyroJerkMag	body gyroscope jerk magnitude (time domain)
11.	fBodyAcc-XYZ	3-axis body acceleration (frequency domain)
12.	fBodyAccJerk-XYZ	3-axis body jerk (frequency domain)
13.	fBodyGyro-XYZ	3-axis body gyroscope (frequency domain)
14.	fBodyAccMag	body acceleration magnitude (frequency domain)
15.	fBodyAccJerkMag	body jerk magnitude (frequency domain)
16.	fBodyGyroMag	body gyroscope magnitude (frequency domain)
17.	fBodyGyroJerkMag	body gyroscope jerk magnitude (frequency domain)

Gravity was measured in standard gravity units (*g*). Angular velocity ("gyroscope") was measured in radians/second. The original documentation makes no further mention of units; however, given its European origin, it is reasonable to suppose that SI units were used, that is:

$[\text{linear acceleration}] = \text{m}\cdot\text{s}^{-2}$

$[\text{linear jerk}] = \text{m}\cdot\text{s}^{-3}$

$[\text{angular acceleration}] = \text{rad}\cdot\text{s}^{-2}$

$[\text{angular jerk}] = \text{rad}\cdot\text{s}^{-3}$

Columns in the 'Features' Data Set

The data set of interest here is sourced from two files (X_test.txt and X_train.txt), each of which contains 561 features extracted from the above signals, for each record. The names of the resulting columns are given in file 'features.txt'. (Nb. 17 signals $*$ 33 metrics = 561 features). Not all of these metrics were of interest in the present analysis (see next section).

Note that some of the feature names are reused three times (e.g. "fBodyAcc-bandsEnergy()-1,16"), in violation of best practice for tidy data, but the affected features are not of interest in the present analysis. Therefore, no attempt was made to tidy these anomalies.

Analysis

Data Selection

Prior to extracting the columns of interest, the training and test data sets were merged. The training set contributed 7,352 records and the test set 2,947 for a total of 10,299 observations.

The metrics of interest in the present analysis were the mean ("mean") and standard deviation ("sdev"). Column names containing "meanFreq" and ",gravity" were deemed to be derived (rather than measured), and were excluded; none of these columns had corresponding standard deviation columns (names containing "std") in the source data. These steps reduced the number of columns of interest to 66 (33 means and 33 standard deviations).

Entity Renaming

The *activities* were read from the data set file 'activity_labels.txt' and renamed as follows, using string functions:

WALKING	walking
WALKING_UPSTAIRS	walking upstairs
WALKING_DOWNSTAIRS	walking downstairs
SITTING	sitting
STANDING	standing
LAYING	laying

User-friendly Variable names were obtained from the data set file 'features.txt' by breaking out the metrics and dimensions, while preserving the core signal names (and ensuring compatibility with R naming rules). For example:

tBodyAcc-mean()-Y → mean.Y.tBodyAcc

This was implemented in an external function ('friendly.name'), invoked using 'vapply'.

Data Processing

Processing occurred in 5 stages. In the 1st stage, the 'test' and 'training' data sets were merged. The original feature names were read and applied to the data frame.

In the 2nd stage, unwanted columns were removed (see Data Selection, above).

In the 3rd stage, the activity names were read and renamed, the activity codes (one per observation) were read, and the data frame was extended by prepending a new activity names column (by substituting for the corresponding activity codes).

In the 4th stage, the subject ID codes (one per observation) were read and prepended to the data frame as a new column. The resulting data frame ('dat') was then converted to a data frame table,

and was not further modified. A copy was made (‘dat1’), within which user-friendly variable names were then substituted for the original column names.

In the 5th stage, ‘dat’ was used as input to a new data frame, ‘dat2’, in which the data were reduced to ensemble averages (using ‘summarise_each’), by grouping over activity and subject. User-friendly variable names were again substituted, this time with the prefix “ensemble” on the new metrics.

Output

Following execution of the solution script, the three 3 data frame tables will be present:

dat	Original variable naming, with columns 'activity' and 'subject' added.
dat1	Dame as 'dat' but with user-friendly variable names.
dat2	Derived from 'dat' by grouping over 'activity' and 'subject', and averaging over all measures, with user-friendly variable names.

Upon finalisation of each of stages 4 and 5, the 'str' command is executed on 'dat1' and 'dat2', respectively, to display the data structures on the console. In addition, ‘dat2’ is written to file ‘tidydat.txt’ (without row names).

Instruction List

The source data set is not included in the analysis file set, and must be downloaded by the end user: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

The file ‘README.md’, created for this analysis, includes instructions for setting up the environment and running the solution script.

At the final stage of execution of the script, the reduced, tidy data set ('dat2') is written to file 'tidydat.txt', in the working directory.