# Detailed Description of run\_analysis.R script:

This project utilizes source data from the

Human Activity Recognition Using Smartphones Dataset Version 1.0.

(See detailed description of study from the authors below).

The data consists of several files of training and test data for 30 people (Subject) performing each of six activities (Activity). The data consists of 562 observations as outlined in features.txt. The following diagram describes the relationship of each of the data files.

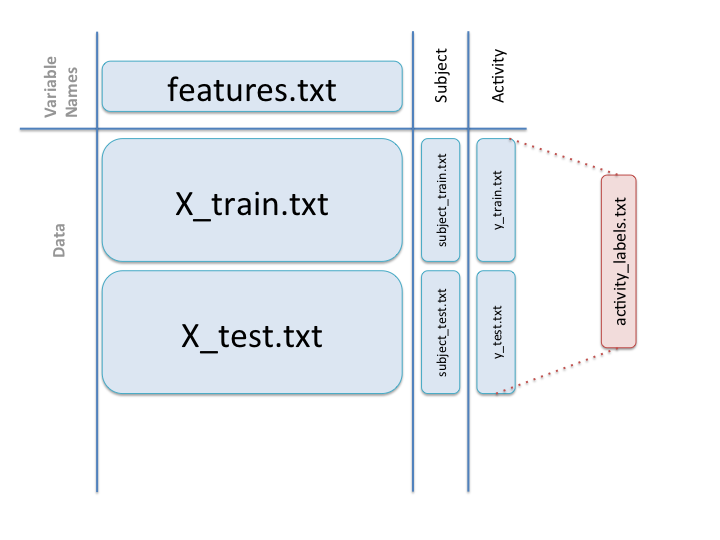


Figure 1: Relationship of data in source data files. Courtesty of David Hood

This project developed a run\_analysis.R script which accomplishes the following tasks:

* Merges the training and the test sets to create one data set.
* Extracts only the measurements on the mean and standard deviation for each measurement.
* Uses descriptive activity names to name the activities in the data set
* Appropriately labels the data set with descriptive variable names.
* From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

The run\_analysis.R script utilizes the following packages:

* plyr (used for mapvalues() function)
* dplyr (used for Selecting colums easily)
* reshape2 (used for the melt and dcast commands)

To start, you must set your working directory to the directory that contains the run\_analysis.R script. The script will create a ./data subdirectory of one does not already exist, and download the source data zip archive to this directory.

The script will then extract the files from that zip archive, into the “UCI HAR Dataset” subdirectory.

Several files are loaded that are used to provide column names for the detail data:

* features\_names (from features.txt)
* activity\_labels (from activity\_labels.txt)
* names for the test and train Subjects (from subject\_test.txt and subject\_train.txt)

Next, the script will load the Activity Vector files that tie activity codes to the observations:

* test\_activity\_vector (from y\_test.txt)
* train\_activity\_vector (from y\_train.txt)

Finally, the script reads in the actual data vectors for the Train and Test datasets, using the appropriate column names from the files above:

* X\_train <- read.csv("./train/X\_train.txt", header = FALSE, sep = "", col.names = features\_names[ ,2])
* X\_test <- read.csv("./test/X\_test.txt", header = FALSE, sep = "", col.names = features\_names[ ,2])

The script will then combine the subject #’s into the data frames:

* X\_train <- cbind(train\_subject\_names, X\_train)
* X\_test <- cbind(test\_subject\_names, X\_test)

And then add the activity #'s to the two datasets

* X\_train <- cbind(train\_activity\_vector, X\_train)
* X\_test <- cbind(test\_activity\_vector, X\_test)

NOW the script can easily combine the test and train data sets into a single data frame:

* X\_data <- rbind(X\_train, X\_test)

For the purposes of this project, we want only the mean and standard deviation variables, for the time series only. Thus we welect only the columns that contain "mean" or "std" and that don't start with an "f" (we do not need the Fast Fourrier Transforms):

* X\_data\_final <- cbind(select(X\_data, contains("Subject"), contains("Activity"), contains("mean"), contains("std"), -starts\_with("f")))

For the purpose of creating a Tidy data set, we want to replace the numeric values for Activity with the descriptive values from the activity\_labels.txt file:

* X\_data\_final$Activity <- mapvalues(as.numeric(X\_data\_final$Activity), activity\_labels[, 1], as.character(activity\_labels[,2]))

We now create a data set containing the Average of each variable per Step 5 of the instructions, with Subject and Activity being the Key values.

First, create a vector of column names that indicate the Key columns

* id.vars <- c("Subject", "Activity")

Next, MELT the data to just key values and each variable on a separate row

* myMelt <- melt(X\_data\_final, id.vars = id.vars)

Now, we will summarize the melted data by calculating the mean of each variable, grouped by Subject, Activity, and variable

* id.vars2 <- c("Subject", "Activity", "variable")
* mySummary <- ddply(myMelt, id.vars2, summarise, mean = mean(value))

Finally, return a tidy data set by dcast'ing the melted data so the each variable is its own column:

* tidyData <- dcast(mySummary, Subject + Activity ~ variable, mean)

Order the data by Subject and Activity

* tidyData <- arrange(tidyData, Subject, Activity)

The final task is to tidy up the variable names so they are descriptive and well formatted. We get a vector of the column names to work with:

* vNames <- names(tidyData)

Change abbreviation "t" to "timeseries"

* vNames <- sub("^t", "TimeSeries", vNames)

Change abbreviation ".std" to "standarddeviation"

* vNames <- sub(".std", "StandardDeviation", vNames)

Change abbreviation "acc" to "acceleration"

* vNames <- (sub("Acc", "Acceleration", vNames))

Change .mean to mean

* vNames <- (sub(".mean", "Mean", vNames))

Change ...x to xaxis, similarly for y and z axis

* vNames <- sub("\\.\\.\\.X$", "Xaxis", vNames)
* vNames <- sub("\\.\\.\\.Y$", "Yaxis", vNames)
* vNames <- sub("\\.\\.\\.Z$", "Zaxis", vNames)

Remove any trailing periods

* vNames <- gsub("\\.\\.$", "", vNames)

Finally, we can apply this vector to become the names for the data frame:

* names(tidyData) <- vNames

The last step is to write the file in the ./data directory

* setwd("../")
* write.table(tidyData, file = "tidyData.txt", sep = " ", row.names = FALSE)

# Codebook

The ./data/tidyData.txt file contains the output of the run\_analysis.R script.

This file consists of the following data:

[1] "Subject" : this is an integer from 1 to 30 representing the

individual person who was observed.

[2] "Activity": this is one of the six possible activities the Subject was performing while being measured for this particular observation. Values are:

* Laying
* Sitting
* Standing
* Walking
* Walking Downstairs
* Walking Upstairs

The remaining columns represent the variables that were measured for each

observation. They are each the MEAN of all the observations from the

original SOURCE data set for that particular subject performing that

particular activity.

The values in the original data set are normalized to fall between -1 and 1.

[3] "TimeSeriesBodyAccelerationMeanXaxis"

[4] "TimeSeriesBodyAccelerationMeanYaxis"

[5] "TimeSeriesBodyAccelerationMeanZaxis"

[6] "TimeSeriesGravityAccelerationMeanXaxis"

[7] "TimeSeriesGravityAccelerationMeanYaxis"

[8] "TimeSeriesGravityAccelerationMeanZaxis"

[9] "TimeSeriesBodyAccelerationJerkMeanXaxis"

[10] "TimeSeriesBodyAccelerationJerkMeanYaxis"

[11] "TimeSeriesBodyAccelerationJerkMeanZaxis"

[12] "TimeSeriesBodyGyroMeanXaxis"

[13] "TimeSeriesBodyGyroMeanYaxis"

[14] "TimeSeriesBodyGyroMeanZaxis"

[15] "TimeSeriesBodyGyroJerkMeanXaxis"

[16] "TimeSeriesBodyGyroJerkMeanYaxis"

[17] "TimeSeriesBodyGyroJerkMeanZaxis"

[18] "TimeSeriesBodyAccelerationMagMean"

[19] "TimeSeriesGravityAccelerationMagMean"

[20] "TimeSeriesBodyAccelerationJerkMagMean"

[21] "TimeSeriesBodyGyroMagMean"

[22] "TimeSeriesBodyGyroJerkMagMean"

[23] "angletBodyAccelerationMean.gravity"

[24] "angletBodyAccelerationJerkMeangravityMean"

[25] "angletBodyGyroMean.gravityMean"

[26] "angletBodyGyroJerkMean.gravityMean"

[27] "angleXgravityMean"

[28] "angleYgravityMean"

[29] "angleZgravityMean"

[30] "TimeSeriesBodyAccelerationStandardDeviationXaxis"

[31] "TimeSeriesBodyAccelerationStandardDeviationYaxis"

[32] "TimeSeriesBodyAccelerationStandardDeviationZaxis"

[33] "TimeSeriesGravityAccelerationStandardDeviationXaxis"

[34] "TimeSeriesGravityAccelerationStandardDeviationYaxis"

[35] "TimeSeriesGravityAccelerationStandardDeviationZaxis"

[36] "TimeSeriesBodyAccelerationJerkStandardDeviationXaxis"

[37] "TimeSeriesBodyAccelerationJerkStandardDeviationYaxis"

[38] "TimeSeriesBodyAccelerationJerkStandardDeviationZaxis"

[39] "TimeSeriesBodyGyroStandardDeviationXaxis"

[40] "TimeSeriesBodyGyroStandardDeviationYaxis"

[41] "TimeSeriesBodyGyroStandardDeviationZaxis"

[42] "TimeSeriesBodyGyroJerkStandardDeviationXaxis"

[43] "TimeSeriesBodyGyroJerkStandardDeviationYaxis"

[44] "TimeSeriesBodyGyroJerkStandardDeviationZaxis"

[45] "TimeSeriesBodyAccelerationMagStandardDeviation"

[46] "TimeSeriesGravityAccelerationMagStandardDeviation"

[47] "TimeSeriesBodyAccelerationJerkMagStandardDeviation"

[48] "TimeSeriesBodyGyroMagStandardDeviation"

[49] "TimeSeriesBodyGyroJerkMagStandardDeviation"

Feature Selection

=================

The features above are the **MEANS of the corresponding features from the SOURCE DATA.** The featuresselected for the source 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).

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

The set of variables that were estimated from these signals are:

mean(): Mean value

std(): Standard deviation

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean

tBodyAccMean

tBodyAccJerkMean

tBodyGyroMean

tBodyGyroJerkMean

# Source Data and Study

==================================================================

Human Activity Recognition Using Smartphones Dataset

Version 1.0

==================================================================

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.

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

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.

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