HUMAN ACTIVITY RECOGNITION USING SMARTPHONES DATA SET

CODE BOOK: This codebook includes information about the source data, the transformations performed after collecting the data and some information about the variables of the resulting data sets.

Raw data collection

Raw data are obtained from UCI Machine Learning repository. In particular we used the Human Activity Recognition Using Smartphones Data Set [1], that was used by the original collectors to conduct experiments exploiting Support Vector Machine (SVM) [2]. Activity Recognition (AR) aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions [3]. The collectors used a sensor based approach employing smartphones as sensing tools. Smartphones are an effective solution for AR, because they come with embedded built-in sensors such as microphones, dual cameras, accelerometers, gyroscopes, etc. The data set was built from experiments 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, 3-axial linear acceleration and 3-axial angular velocity were captured at a constant rate of 50Hz.

The experiments have been video-recorded to label the data manually [4]. The obtained data set 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.

A video of the experiment including an example of the 6 recorded activities with one of the participants can be seen in the following link: [Web Link]

An updated version of this dataset can be found at [Web Link]. It includes labels of postural transitions between activities and also the full raw inertial signals instead of the ones pre-processed into windows.

Attribute Information:

For each record in the dataset it is provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.

Signals
The 3-axial time domain [5] signals from accelerometer and gyroscope were captured at a constant rate of 50 Hz [6]. Then they were filtered to remove noise. Similarly, the acceleration signal was then separated into body and gravity acceleration signals using another filter. Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals [7]. Also the magnitude [8] of these three-dimensional signals were calculated using the Euclidean norm [9]. Finally a Fast Fourier Transform (FFT) [10] was applied to some of these time domain signals to obtain frequency domain [11] signals. The signals were sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window at 50 Hz). From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

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(): Autoregression 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 some vectors.

No unit of measures is reported as all features were normalized and bounded within [-1,1].

Study Design
The source data was collected from the UCI Machine Learning Repository to complete an assignment for a Coursera course named Getting and Cleaning Data instructed by Jeff Leek. The assignment involved working with the data set and producing tidy data representation of the source data. Below is a list of the operations done to achieve the outputs.

- Downloaded the data set
- Loaded the libraries : plyr, dplyr & data.table
- Unzipped the data set into the chosen working directory

- Loaded test and training data sets into data frames
- Loaded source variable names for test and training data sets
- Loaded activity labels
- Combined test and training data frames using rbind
- Paired down the data frames to only include the mean and standard deviation variables
- Replaced activity IDs with the activity labels for better readability
- Combined the data frames to produce one data frame containing the subjects, measurements and activities
- Produced "maindata" with the combined data frame as the first expected output
- Created another data set using the data.table library to easily group the tidy data by subject and activity
- Then applied the mean and standard deviation calculations across the groups
- Produced "Tidy.txt" as the second expected output

Variables

Subject Id: 1 to 30 each representing a participant in the study
Activity: the activity that the subject was doing at the time of the measurement
Feature Fields

- tBodyAcc-mean()-X (column 1)
- tBodyAcc-mean()-Y (column 2)
- tBodyAcc-mean()-Z (column 3)
- tBodyAcc-std()-X (column 4)
- tBodyAcc-std()-Y (column 5)
- tBodyAcc-std()-Z (column 6)
- tGravityAcc-mean()-X (column 41)
- tGravityAcc-mean()-Y (column 42)
- tGravityAcc-mean()-Z (column 43)
- tGravityAcc-std()-X (column 44)
- tGravityAcc-std()-Y (column 45)
- tGravityAcc-std()-Z (column 46)
- tBodyAccJerk-mean()-X (column 81)
- tBodyAccJerk-mean()-Y (column 82)
- tBodyAccJerk-mean()-Z (column 83)
- tBodyAccJerk-std()-X (column 84)
- tBodyAccJerk-std()-Y (column 85)
- tBodyAccJerk-std()-Z (column 86)
- tBodyGyro-mean()-X (column 121)
- tBodyGyro-mean()-Y (column 122)
- tBodyGyro-mean()-Z (column 123)
- tBodyGyro-std()-X (column 124)
- tBodyGyro-std()-Y (column 125)
- tBodyGyro-std()-Z (column 126)
- tBodyGyroJerk-mean()-X (column 161)
- tBodyGyroJerk-mean()-Y (column 162)
- tBodyGyroJerk-mean()-Z (column 163)
- tBodyGyroJerk-std()-X (column 164)

- tBodyGyroJerk-std()-Y (column 165)
- tBodyGyroJerk-std()-Z (column 166)
- tBodyAccMag-mean() (column 201)
- tBodyAccMag-std() (column 202)
- tGravityAccMag-mean() (column 214)
- tGravityAccMag-std() (column 215)
- tBodyAccJerkMag-mean() (column 227)
- tBodyAccJerkMag-std() (column 228)
- tBodyGyroMag-mean() (column 240)
- tBodyGyroMag-std() (column 241)
- tBodyGyroJerkMag-mean() (column 253)
- tBodyGyroJerkMag-std() (column 254)
- fBodyAcc-mean()-X (column 266)
- fBodyAcc-mean()-Y (column 267)
- fBodyAcc-mean()-Z (column 268)
- fBodyAcc-std()-X (column 269)
- fBodyAcc-std()-Y (column 270)
- fBodyAcc-std()-Z (column 271)
- fBodyAccJerk-mean()-X (column 345)
- fBodyAccJerk-mean()-Y (column 346)
- fBodyAccJerk-mean()-Z (column 347)
- fBodyAccJerk-std()-X (column 348)
- fBodyAccJerk-std()-Y (column 349)
- fBodyAccJerk-std()-Z (column 350)
- fBodyGyro-mean()-X (column 424)
- fBodyGyro-mean()-Y (column 425)
- fBodyGyro-mean()-Z (column 426)
- fBodyGyro-std()-X (column 427)
- fBodyGyro-std()-Y (column 428)
- fBodyGyro-std()-Z (column 429)
- fBodyAccMag-mean() (column 503)
- fBodyAccMag-std() (column 504)
- fBodyBodyAccJerkMag-mean() (column 516)
- fBodyBodyAccJerkMag-std() (column 517)
- fBodyBodyGyroMag-mean() (column 529)
- fBodyBodyGyroMag-std() (column 530)
- fBodyBodyGyroJerkMag-mean() (column 542)
- fBodyBodyGyroJerkMag-std() (column 543)

Activity Labels

12 WALKING (value 1)
13 WALKING_UPSTAIRS (value 2)
14 WALKING_DOWNSTAIRS (value 3)
15 SITTING (value 4)
16 STANDING (value 5)
17 LAYING (value 6)

Extracted Features Vector
c(1, 2, 3, 4, 5, 6, 41, 42, 43, 44, 45, 46, 81, 82, 83, 84, 85, 86, 121, 122, 123, 124, 125, 126, 161, 162, 163, 164, 165, 166, 201, 202, 214, 215, 227, 228, 240, 241, 253, 254, 266, 267, 268, 269, 270, 271, 345, 346, 347, 348, 349, 350, 424, 425, 426, 427, 428, 429, 503, 504, 516, 517, 529, 530, 542, 543)

Extracted Feature Names Vector
c("tBodyAcc-mean()-X", "tBodyAcc-mean()-Y", "tBodyAcc-mean()-Z", "tBodyAcc-std()-X", "tBodyAcc-std()-Y", "tBodyAcc-std()-Z", "tGravityAcc-mean()-X", "tGravityAcc-mean()-Y", "tGravityAcc-mean()-Z", "tGravityAcc-std()-X", "tGravityAcc-std()-Y", "tGravityAcc-std()-Z", "tBodyAccJerk-mean()-X", "tBodyAccJerk-mean()-Y", "tBodyAccJerk-mean()-Z", "tBodyAccJerk-std()-X", "tBodyAccJerk-std()-Y", "tBodyAccJerk-std()-Z", "tBodyGyro-mean()-X", "tBodyGyro-mean()-Y", "tBodyGyro-mean()-Z", "tBodyGyro-std()-X", "tBodyGyro-std()-Y", "tBodyGyro-std()-Z", "tBodyGyroJerk-mean()-X", "tBodyGyroJerk-mean()-Y", "tBodyGyroJerk-mean()-Z", "tBodyGyroJerk-std()-X", "tBodyGyroJerk-std()-Y", "tBodyGyroJerk-std()-Z", "tBodyAccMag-mean()", "tBodyAccMag-std()", "tGravityAccMag-mean()", "tGravityAccMag-std()", "tBodyAccJerkMag-mean()", "tBodyAccJerkMag-std()", "tBodyGyroMag-mean()", "tBodyGyroMag-std()", "tBodyGyroJerkMag-mean()", "tBodyGyroJerkMag-std()", "fBodyAcc-mean()-X", "fBodyAcc-mean()-Y", "fBodyAcc-mean()-Z", "fBodyAcc-std()-X", "fBodyAcc-std()-Y", "fBodyAcc-std()-Z", "fBodyAccJerk-mean()-X", "fBodyAccJerk-mean()-Y", "fBodyAccJerk-mean()-Z", "fBodyAccJerk-std()-X", "fBodyAccJerk-std()-Y", "fBodyAccJerk-std()-Z", "fBodyGyro-mean()-X", "fBodyGyro-mean()-Y", "fBodyGyro-mean()-Z", "fBodyGyro-std()-X", "fBodyGyro-std()-Y", "fBodyGyro-std()-Z", "fBodyAccMag-mean()", "fBodyAccMag-std()", "fBodyBodyAccJerkMag-mean()", "fBodyBodyAccJerkMag-std()", "fBodyBodyGyroMag-mean()", "fBodyBodyGyroMag-std()", "fBodyBodyGyroJerkMag-mean()", "fBodyBodyGyroJerkMag-std()")

Activities Vector
c(1, 2, 3, 4, 5, 6)

Activity Names Vector
c("WALKING", "WALKING_UPSTAIRS", "WALKING_DOWNSTAIRS", "SITTING", "STANDING", "LAYING")

Data transformation
The raw data sets are processed with run_analisys.R script to create a tidy data set [12].

Merge training and test sets
Test and training data (X_train.txt, X_test.txt), subject ids (subject_train.txt, subject_test.txt) and activity ids (y_train.txt, y_test.txt) are merged to obtain a single data set. Variables are labeled with the names assigned by original collectors (features.txt).

Extract mean and standard deviation variables
From the merged data set is extracted and intermediate data set with only the values of estimated mean (variables with labels that contain "mean") and standard deviation (variables with labels that contain "std").

Use descriptive activity names
A new column is added to intermediate data set with the activity description. Activity id column is used to look up descriptions in activity_labels.txt.

Label variables appropriately
Labels given from the original collectors were changed: to obtain valid R names without parentheses, dashes and commas to obtain more descriptive labels

Create a tidy data set
From the intermediate data set is created a final tidy data set where numeric variables are averaged for each activity and each subject.

The tidy data set contains 180 observations with 479 variables divided in:

- an activity label (Activity): WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING
- an identifier of the subject who carried out the experimnoent (Subject): 1, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30
- a 79-feature vector with time and frequency domain signal variables (numeric)

For variables derived from mean and standard deviation estimation, the previous labels are augmented with the terms "Mean" or "StandardDeviation".

Please refer to run_analysis.R for implementation details.

Relevant Papers:
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

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge L. Reyes-Ortiz. Energy Efficient Smartphone-Based Activity Recognition using Fixed-Point Arithmetic. Journal of Universal Computer Science. Special Issue in Ambient Assisted Living: Home Care. Volume 19, Issue 9. May 2013

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. 4th International Workshop of Ambient Assisted Living, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings. Lecture Notes in Computer Science 2012, pp 216-223.

Jorge Luis Reyes-Ortiz, Alessandro Ghio, Xavier Parra-Llanas, Davide Anguita, Joan Cabestany, Andreu Català. Human Activity and Motion Disorder Recognition: Towards Smarter Interactive Cognitive Environments. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.

Citation Request:
Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.