CodeBook

César Aguirre Rivadeneira

26/1/2020

CodeBook

This code book describes the variables, the data, and any transformations or work performed to clean up the data.

Source Data

The source data is provided by the course, is available in the folder "UCI HAR Dataset".

The original data is available here: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Data Set Information(Extracted from the web page)

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.

Data Set Files

The data set include the following files:

- README.txt
- features_info.txt
- features.txt
- activity_labels.txt
- train/X_train.txt
- train/y_train.tx
- test/X_test.txt
- test/y_test.txt

Data Getting/Cleaning

The script run_analysis.R perform the Getting/Cleaning actions with the main purpouse of following the instructions given in the course project:

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.

3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names.
5. 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.

Read the TRAIN dataset files

Fist step is to read the TRAIN datasets located in "/UCI HAR Dataset/train". The file with the data is called X_train.txt, this file has a total of 7352 rows and 561 columns. The function read.csv is used to load the file, giving a dataframe with 7352 rows and 561 columns called "data_train".

Asign column names

To asign column names, the first step is load the file with the measurements name called "features.txt". And the set this vector as the columns name.

Select the mean and standard deviation columns

In the process of selecting the right columns, we detected duplicated names in the columns. First, the script deletes this problematic columns and then select those who contains "-mean()" or "-std()" as part of the name of the column. Leaving 66 measurements columns thar are mean or standard deviation.

Add the columns subject and activity

The next step is to add the columns subject and activity that are stored in the files "subject_train.txt" and "y_train.txt" respectively. Then, the script add those files as a columns in the end of the main dataframe "data_train". To load the activities as a column, first we need to change the number 1 to 6, to the description value, this transformation is detailed in the file "activity_labels.txt"

Load and Clean the TEST dataset files

For the test files the script repeat the same steps as it did with the train dataset. This dataset is called "data_test"

Merge TRAIN and TEST datasets

Having clean datasets for TEST and TRAIN, the script merge them together resulting in a 10299x68 dataframe. The dataframe is called DataAll

Create the tidy dataset

Then the script aggregate and order the informacion by Activity and Subject as it is required by the project, resulting in a dataframe of 180 rows and the same 68 variables that had the previous dataframe, this dataframe is called AggData. Having in the first column the Activity, in the second column the Subject and in the rest 66 columns the average of each metric.

Ejecution

To run the code you have to load the function run_analysis.R and execute:

AggData <- run_analysis()