Code Book for the Project of Getting and Cleaning Data

Xiao Lu

May 18, 2020

1 Background

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.he 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.

2 Data

2.1 The Raw 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).These signals can be found in [the path for test data](./data/UCI HAR Dataset/test/Inertial Signals) and [the path for training data](./data/UCI HAR Dataset/training/Inertial Signals) embeeded in the [UCI HAR Dataset](./data/UCI HAR Dataset). These signals are pre-processed for generating the raw data set used for this project as follows: 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.

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: (1) train/subject_train.txt: Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30. (2) '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. (3) train/Inertial Signals/body_acc_x _train.txt: The body acceleration signal obtained by subtracting the gravity from the total acceleration. (4) 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.

2.2 Tidying the Data

Firstly, I created the directory for data. Afterwards, the compressed file was downloaded by the download.file() function and then unzipped via the unzip().

Secondly, the variables' names were extracted by reading the features.txt. Several adaptations were carried out, such as transforming them to lower and transposing it to a row vector. The reason is when we are trying to assign these values to the column names in the dataset, the column names are in a row vector. The meanings of many other ordinary manipulations are not elucidated here, since they are self-explained in the comment of run_analysis_xl.R.

Another noteworthy point is in Step 4, when relabeling the column names, I invoked the escape pattern "\\" to remove the "()".

3 The Final Outcome

I attached a screenshot of the final output, displayed in the R Studio instead of the .txt file.

	activity_labels	person_labels	time-bodyacceleration-mean-x	time-bodyacceleration-mean-y	time-bodyacceleration-mean-z	time-bodyacceleration-std-x
1	lying	1	0.2215982	-0.040513953	-0.11320355	-0.9280565
2	lying	2	0.2813734	-0.018158740	-0.10724561	-0.9740595
3	lying	3	0.2755169	-0.018955679	-0.10130048	-0.9827766
4	lying	4	0.2635592	-0.015003184	-0.11068815	-0.9541937
5	lying	5	0.2783343	-0.018304212	-0.10793760	-0.9659345
6	lying	6	0.2486565	-0.010252917	-0.13311957	-0.9340494
7	lying	7	0.2501767	-0.020441152	-0.10136104	-0.9365136
8	lying	8	0.2612543	-0.021228173	-0.10224537	-0.9430412
9	lying	9	0.2591955	-0.020526822	-0.10754972	-0.9423331

Figure 1: The Final Outcome

In a nutshell, the data in this file is first grouped by the 'activity_label', such as lying or standing. The second layer of grouping variable is 'person_labels', referring to each specific participant in this experiment. The variables in this final dataset include:

* time-bodyacceleration-mean-x/y/z: the mean time of body acceleration in axial x, y, or z.
* time-bodyacceleration-std-x/y/z: the standard deviation of body acceleration in axial x, y or z.
* time-gravityacceleration-mean-x/y/z: the mean time of gravity acceleration in axial x, y, or z.
* time-gravityacceleration-std-x/y/z: the standard deviation of body acceleration in axial x, y, or z.
* time-bodyacceleration-jerk-signal-mean/std-x/y/z: by the same token, these are either mean or standard deviation of the jerk signal detected by the equipment's sensor.
* I skip typing the similar variables' names here as this is so time-consuming and repetitive. The keynote is gyro is replaced by gyroscope, which is more meaningful.