CodeBook

Jesus Felix

3/14/2020

Project Description

The objective of this project is to create a tidy data set suitable for future analysis from a untidy dataset. The script "run_analysis.R" creates a tidy subset of data from the data set provided for the final project of the Getting and Cleaning Data.

Study design and data processing

Collection of the raw data

The data used in this exercise was collected from the accelerometers from the Samsung Galaxy S smartphone.

Creating the tidy datafile

Guide to create the tidy data file

The following steps are required to create the tidy data set:

1. Download the data here. Unzip the file and that will create a folder with some txt files and two more folders (test and train). Do not modify the names nor the structure.
2. Set your R working directory to the unzipped folder
3. Run the run_analysis.R script
4. The script output (tidy_dataset) is a tidy dataset with the average of each variable for each subject and each activity

Cleaning of the data

The cleaning process of the data includes the following steps:

1. Load the data
2. Combine the train and test sets into a single object
3. Extract those features that contain mean or std in the feature name
4. Use the activity_labels.txt file to replace the activity IDs with descriptive activity names
5. Label the data set variables with descriptive names
6. Create a tidy data set with the average of each variable for each subject and each activity

Description of the tidy dataset

The tidy dataset can be found in the repository as tidy_dataset.txt. The dataset has the following characteristics:

- Dimensions of the dataset: 180 81
- Summary of the data:
- Variables present in the dataset: columns 1 and 2 contain the Subject and Activity, respectively. Columns 3 to 81 contain the average of each of the following variables in that order: Subject, Activity,

tBodyAccmeanX, tBodyAccmeanY, tBodyAccmeanZ, tGravityAccmeanX, tGravityAccmeanY, tGravityAccmeanZ, tBodyAccJerkmeanX, tBodyAccJerkmeanY, tBodyAccJerkmeanZ, tBodyGyromeanX, tBodyGyromeanY, tBodyGyromeanZ, tBodyGyroJerkmeanX, tBodyGyroJerkmeanY, tBodyGyroJerkmeanZ, tBodyAccMagmean, tGravityAccMagmean, tBodyAccJerkMagmean, tBodyGyroMagmean, tBodyGyroJerkMagmean, fBodyAccmeanX, fBodyAccmeanY, fBodyAccmeanZ, fBodyAccmeanFreqX, fBodyAccmeanFreqY, fBodyAccmeanFreqZ, fBodyAccJerkmeanX, fBodyAccJerkmeanY, fBodyAccJerkmeanZ, fBodyAccJerkmeanFreqX, fBodyAccJerkmeanFreqY, fBodyAccJerkmeanFreqZ, fBodyGyromeanX, fBodyGyromeanY, fBodyGyromeanZ, fBodyGyromeanFreqX, fBodyGyromeanFreqY, fBodyGyromeanFreqZ, fBodyAccMagmean, fBodyAccMagmeanFreq, fBodyBodyAccJerkMagmean, fBodyBodyAccJerkMagmeanFreq, fBodyBodyGyroMagmean, fBodyBodyGyroMagmeanFreq, fBodyBodyGyroJerkMagmean, fBodyBodyGyroJerkMagmeanFreq, tBodyAccstdX, tBodyAccstdY, tBodyAccstdZ, tGravityAccstdX, tGravityAccstdY, tGravityAccstdZ, tBodyAccJerkstdX, tBodyAccJerkstdY, tBodyAccJerkstdZ, tBodyGyrostdX, tBodyGyrostdY, tBodyGyrostdZ, tBodyGyroJerkstdX, tBodyGyroJerkstdY, tBodyGyroJerkstdZ, tBodyAccMagstd, tGravityAccMagstd, tBodyAccJerkMagstd, tBodyGyroMagstd, tBodyGyroJerkMagstd, fBodyAccstdX, fBodyAccstdY, fBodyAccstdZ, fBodyAccJerkstdX, fBodyAccJerkstdY, fBodyAccJerkstdZ, fBodyGyrostdX, fBodyGyrostdY, fBodyGyrostdZ, fBodyAccMagstd, fBodyBodyAccJerkMagstd, fBodyBodyGyroMagstd, fBodyBodyGyroJerkMagstd

- Description of the variables: The features selected for this 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).

  Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).

  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
  - fBodyAcc-XYZ
  - fBodyAccJerk-XYZ
  - fBodyGyro-XYZ
  - fBodyAccMag
  - fBodyAccJerkMag
  - fBodyGyroMag
  - fBodyGyroJerkMag

  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