codebook

Surya

04/09/2020

Code Book

INTRODUCTION

This code book is used for

- outlining the analyses to clean up the data
- interpreting the variables in the tidyData.txt document

The present version is modified from the original code book provided by Coursera Data Science Specialization Peer-graded Assignment: Getting and Cleaning Data Course Project that can be downloaded from here

ANALYSIS PIPELINE

1. Read in the files and merge data

- label files: activity_labels.txt and features.txt
- training data: subject_test.txt, X_test.txt (assign column names as feature names), and y_test.txt
- test data: subject_train.txt, X_train.txt (assign column names as feature names), and y_train.txt
- merge all training files horizontally into one training data
- merge all test files horizontally into one test data
- merge the training data and test data vertically into one data

2. Extract only the mean and standard deviation of the measurments

- keep the subject Id and actvitiy columns
- identify column names with 'mean' or 'std'

3. Use descriptive activity names to name the activities in the data set

- change the activity column from numberse 1-6 to descriptive activities: WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING

4. Label the data set with descriptive variable names

- remove the () in the columns to make the variable names cleaner and easier to read
- for a description of the variables, please refer to 'FEATURE VARIABLES' below

5. Create a second, independent tidy data set with the average of each variable for each activity and each subject

utilize the dplyr package to

- group the data set by subject ID and activity (group_by)
- summarize the grouped data set and calculate the average of each variable (summarise_all(funs_mean))

For the detailed script, please refer to the run_analysis.R file.

FEATURE SELECTION

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.

FEATURE VARIABLES

Time signals

- tBodyAcc-XYZ
- tGravityAcc-XYZ
- tBodyAccJerk-XYZ
- tBodyGyro-XYZ
- tBodyGyroJerk-XYZ
- tBodyAccMag
- tGravityAccMag
- tBodyAccJerkMag
- tBodyGyroMag
- tBodyGyroJerkMag

Frequency domain signals

- fBodyAcc-XYZ
- fBodyAccJerk-XYZ
- fBodyGyro-XYZ
- fBodyAccMag
- fBodyAccJerkMag

- fBodyGyroMag
- fBodyGyroJerkMag

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

- mean: Mean value
- std: Standard deviation

OTHER VARIABLES

- subjectId: Identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.
- activity: Six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) that the subjects performed wearing a smartphone (Samsung Galaxy S II) on the waist.

REFERENCE

This dataset is downloaded from the following publication:

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