Codebook for Data Cleaning Project for Coursera Course

Toby Popenfoose

This is the codebook to describe the input files and resulting variables.

Introduction

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. See 'features_info.txt' for more details.

For each record it is provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.
- A 561-feature vector with time and frequency domain variables.
- Its activity label.
- An identifier of the subject who carried out the experiment.

The datasets were obtained from the following link: Project Data which has the original datasets archived at the University of California in Irvine (UCI).

Subjects

The 30 volunteer subjects were given ID's of 1 to 30.

Activities

There were six different activities for each of the 30 subjects:

1. WALKING

2. WALKING_UPSTAIRS
3. WALKING_DOWNSTAIRS

4. SITTING
5. STANDING
6. LAYING

Data text files used

Eight text files were used for my input:

- `X_train.txt` and `X_test.txt`: contain the actual processed averages and standard deviations for each subject and each activity for each observation.
- `subject_train.txt` and `subject_test.txt`: These contain the IDs of the subject for each observation.
- `y_train.txt` and `y_test.txt`: contain the Activity ID of each observation.
- `features.txt`: contain variables of each of the 562 features.
- `activity_labels.txt`: mapping of the ID of each activity.

Variables included with tidy dataset

I ended up with 66 variables in the tidy data set (besides the subject ID and Activity). They are unitless mean and standard deviation with a range from -1.0 to 1.0

These are:

- "TimeBodyAccMeanX"
- "TimeBodyAccMeanY"
- "TimeBodyAccMeanZ"
- "TimeBodyAccStdX"
- "TimeBodyAccStdY"
- "TimeBodyAccStdZ"
- "TimeGravityAccMeanX"
- "TimeGravityAccMeanY"
- "TimeGravityAccMeanZ"
- "TimeGravityAccStdX"
- "TimeGravityAccStdY"
- "TimeGravityAccStdZ"
- "TimeBodyAccJerkMeanX"
- "TimeBodyAccJerkMeanY"

- “TimeBodyAccJerkMeanZ”
- “TimeBodyAccJerkStdX”
- “TimeBodyAccJerkStdY”
- “TimeBodyAccJerkStdZ”
- “TimeBodyGyroMeanX”
- “TimeBodyGyroMeanY”
- “TimeBodyGyroMeanZ”
- “TimeBodyGyroStdX”
- “TimeBodyGyroStdY”
- “TimeBodyGyroStdZ”
- “TimeBodyGyroJerkMeanX”
- “TimeBodyGyroJerkMeanY”
- “TimeBodyGyroJerkMeanZ”
- “TimeBodyGyroJerkStdX”
- “TimeBodyGyroJerkStdY”
- “TimeBodyGyroJerkStdZ”
- “TimeBodyAccMagMean”
- “TimeBodyAccMagStd”
- “TimeGravityAccMagMean”
- “TimeGravityAccMagStd”
- “TimeBodyAccJerkMagMean”
- “TimeBodyAccJerkMagStd”
- “TimeBodyGyroMagMean”
- “TimeBodyGyroMagStd”
- “TimeBodyGyroJerkMagMean”
- “TimeBodyGyroJerkMagStd”
- “FreqBodyAccMeanX”

- “FreqBodyAccMeanY”
- “FreqBodyAccMeanZ”
- “FreqBodyAccStdX”
- “FreqBodyAccStdY”
- “FreqBodyAccStdZ”
- “FreqBodyAccJerkMeanX”
- “FreqBodyAccJerkMeanY”
- “FreqBodyAccJerkMeanZ”
- “FreqBodyAccJerkStdX”
- “FreqBodyAccJerkStdY”
- “FreqBodyAccJerkStdZ”
- “FreqBodyGyroMeanX”
- “FreqBodyGyroMeanY”
- “FreqBodyGyroMeanZ”
- “FreqBodyGyroStdX”
- “FreqBodyGyroStdY”
- “FreqBodyGyroStdZ”
- “FreqBodyAccMagMean”
- “FreqBodyAccMagStd”
- “FreqBodyAccJerkMagMean”
- “FreqBodyAccJerkMagStd”
- “FreqBodyGyroMagMean”
- “FreqBodyGyroMagStd”
- “FreqBodyGyroJerkMagMean”
- “FreqBodyGyroJerkMagStd”

Summary of how the data was tidied

See README.md

Where to get the final tidy data set

IndependentTidyData.txt from my git repository (or from my coursera project page).

Thank you for reading this Codebook.