CodeBook Data cleaning assignement

Omar Benjelloun

02/01/2021

Code book describing the variables, the data, and any transformations or work performed to clean up the data

Original Data

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.

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

And other variables that were computed in the original set but we selected only the mean and standard deviation for this work

Note: Features are normalized and bounded within [-1,1].

Link to original data: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

Data tidying

Transformation Pipeline

To come up with our tidy data set we:

- First dowloaded and loaded 4 data sets X_train,Y_train,X_test,Y_test
- Merged all theses data sets, Xs by rows and Ys by column
- Filtered out the variables that were not mean or standard deviation, since there was no specifications we selected all the variables that contained "mean" or "std" ending up with 79 variables or columns
- Changed the activity labels from 1 to 6 to WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING
- Added the subject IDs to the data frame
- Computed the mean of each variables for each subject and activity and stored it in tidy data set
- Finally the data set was saved in a text file: "WideTidyData.txt" with a header.

Final data set

- Each row corresponds to a combination of subject and activity: subject 1 Walking, subject 1 Laying,... subject 30 Laying
- Each value represents the mean by subject and activity for each variable
- Each column corresponds to a variable (listed below)

Variables

"1" "subjectID"
"2" "activity_labels"

“3” “tBodyAcc.mean. . . X”
“4” “tBodyAcc.mean. . . Y”
“5” “tBodyAcc.mean. . . Z”
“6” “tBodyAcc.std. . . X”
“7” “tBodyAcc.std. . . Y”
“8” “tBodyAcc.std. . . Z”
“9” “tGravityAcc.mean. . . X”
“10” “tGravityAcc.mean. . . Y”
“11” “tGravityAcc.mean. . . Z”
“12” “tGravityAcc.std. . . X”
“13” “tGravityAcc.std. . . Y”
“14” “tGravityAcc.std. . . Z”
“15” “tBodyAccJerk.mean. . . X”
“16” “tBodyAccJerk.mean. . . Y”
“17” “tBodyAccJerk.mean. . . Z”
“18” “tBodyAccJerk.std. . . X”
“19” “tBodyAccJerk.std. . . Y”
“20” “tBodyAccJerk.std. . . Z”
“21” “tBodyGyro.mean. . . X”
“22” “tBodyGyro.mean. . . Y”
“23” “tBodyGyro.mean. . . Z”
“24” “tBodyGyro.std. . . X”
“25” “tBodyGyro.std. . . Y”
“26” “tBodyGyro.std. . . Z”
“27” “tBodyGyroJerk.mean. . . X”
“28” “tBodyGyroJerk.mean. . . Y”
“29” “tBodyGyroJerk.mean. . . Z”
“30” “tBodyGyroJerk.std. . . X”
“31” “tBodyGyroJerk.std. . . Y”
“32” “tBodyGyroJerk.std. . . Z”
“33” “tBodyAccMag.mean..”
“34” “tBodyAccMag.std..”
“35” “tGravityAccMag.mean..”
“36” “tGravityAccMag.std..”
“37” “tBodyAccJerkMag.mean..”
“38” “tBodyAccJerkMag.std..”
“39” “tBodyGyroMag.mean..”
“40” “tBodyGyroMag.std..”
“41” “tBodyGyroJerkMag.mean..”
“42” “tBodyGyroJerkMag.std..”
“43” “fBodyAcc.mean. . . X”
“44” “fBodyAcc.mean. . . Y”
“45” “fBodyAcc.mean. . . Z”
“46” “fBodyAcc.std. . . X”
“47” “fBodyAcc.std. . . Y”
“48” “fBodyAcc.std. . . Z”
“49” “fBodyAcc.meanFreq. . . X”
“50” “fBodyAcc.meanFreq. . . Y”
“51” “fBodyAcc.meanFreq. . . Z”
“52” “fBodyAccJerk.mean. . . X”
“53” “fBodyAccJerk.mean. . . Y”
“54” “fBodyAccJerk.mean. . . Z”
“55” “fBodyAccJerk.std. . . X”
“56” “fBodyAccJerk.std. . . Y”

“57” “fBodyAccJerk.std. . . Z”
“58” “fBodyAccJerk.meanFreq. . . X”
“59” “fBodyAccJerk.meanFreq. . . Y”
“60” “fBodyAccJerk.meanFreq. . . Z”
“61” “fBodyGyro.mean. . . X”
“62” “fBodyGyro.mean. . . Y”
“63” “fBodyGyro.mean. . . Z”
“64” “fBodyGyro.std. . . X”
“65” “fBodyGyro.std. . . Y”
“66” “fBodyGyro.std. . . Z”
“67” “fBodyGyro.meanFreq. . . X”
“68” “fBodyGyro.meanFreq. . . Y”
“69” “fBodyGyro.meanFreq. . . Z”
“70” “fBodyAccMag.mean..”
“71” “fBodyAccMag.std..”
“72” “fBodyAccMag.meanFreq..”
“73” “fBodyBodyAccJerkMag.mean..”
“74” “fBodyBodyAccJerkMag.std..”
“75” “fBodyBodyAccJerkMag.meanFreq..”
“76” “fBodyBodyGyroMag.mean..”
“77” “fBodyBodyGyroMag.std..”
“78” “fBodyBodyGyroMag.meanFreq..”
“79” “fBodyBodyGyroJerkMag.mean..”
“80” “fBodyBodyGyroJerkMag.std..”
“81” “fBodyBodyGyroJerkMag.meanFreq..”