README

Sam Vennell

27 Aug 2016

Course Project: Getting and Cleaning Data

CourseRA course taught by John Hopkins University README Sam Vennell 27/08/2016

Project Aim:

To produce a tidy dataset containing the average of each measure of mean and standard deviation from the combined testing and training data in the UCI HAR Dataset, for each subject and activity type.

Method:

All aspects of the test and train data, namely the X component of the data (eg "X_test.txt") which contains the variables and their values, the y component (eg "y_test.txt") which contains the activity numbers for these and the subject IDs (eg "subject_test.txt") are all loaded into R. These aspects of the testing and training data are combined into the "wide format" data frames "testdata" and "traindata" respectively, adding a column "datatype" which indicates whether the data is testing or training data.

These two data frames are subsequently combined into the dataframe "combdata". The data containing the means and standard deviations of the variables are then extracted into the "wide format" data frame "UCIHAR_meansd_wide" using regex expressions.

These are then "gathered" into a tidyer, "narrow" format data frame called "UCIHAR_meansd".

Finally, the average (mean) of each variable for each activity and subject is calculated using the {dplyr} package and stored in the data frame "UCIHAR_meansd_avg". Note that the training and test data are combined in this calculation.

Notes:

- I have opted to have the final data frame in "narrow" format, and to keep the original variable names. This qualifies as tidy data because it is easy to manipulate, plot and perform further summarisation with. This is confirmed in Hadley Wickham's paper "Tidy Data" http://vita.had.co.nz/papers/tidy-data.pdf.
- I have chosen to include the "angle" means, eg "angle(Z,gravityMean)", as well as the mean and standard devation variables in the standard format (eg "fBodyGyro-std()-Y"), as the initial assignment brief was vague with regard to this matter.
- I could have chosen to decompose the variable names into multiple fields, eg "tBodyAcc-std()-X" into three fields "tBodyAcc", "std()" and "X" but I opted not to since the formatting of the variable names is not uniform (as with the "angle(Z,gravityMean)" example given above).