Codebook for Getting & Cleaning Data Project 2

Author: W. J. Raynor

22 March 2015

Background

This document summarizes the contents of the tidy_data.txt file created for the Coursera course "Getting and Cleaning Data" (012-March 2015). It includes a summary of the variables, the data sources, and the programming steps used to transform the raw data files into the final data step.

The data may be imported into R using the read.table() command. It contains 180 observations and 69 variables. The 66 measurement values are the means of the mean() and std() variables reported in the original files. See the README.txt file that accompanies the original data for a summary of the data pre-processing.

Variables

Variable	**Contents and *values***
Group	Which data group that line belong in
	a labelled factor, with values of *test* and *train*
Subject	unique Subject number (integer)
	1 - 30
Activity	A factor labeling the subject's activity during data collection
	(each subject engaged in multiple activities)
	Walking
	Walking Upstairs
	Walking Downstairs
	Sitting
	Standing
	Laying
tBodyAcc-mean-X	numeric mean of the *tBodyAcc-mean()-X* feature for the subject x activity
tBodyAcc-mean-Y	numeric mean of the *tBodyAcc-mean()-Y* feature for the subject x activity
tBodyAcc-mean-Z	numeric mean of the *tBodyAcc-mean()-Z* feature for the subject x activity
…*(62 further feature means)*	*see the README.txt file included with the data*
fBodyBodyGyroJerkMag-std	numeric mean of the *fBodyBodyGyroJerkMag-std()* feature for the subject x activity

Data Source

The source data were collected from the accelerometers of Samsung Galaxy S smartphones. There were 30 subjects in this study, whose data were collected during each of 6 activities. The raw data were preprocessed by the experimenters as outlined in the file *README.txt* and summarized by a 561 element vector of processed data. These included the mean and standard deviation of the measurements, as well as other statistics not included in this analysis. The organization of the data files is summarized in the *README.txt* file in the data directory.

The data were manually downloaded as a zip file and unzipped to produce a directory called *UCI HAR Dataset.* **The run_analysis.R script assumes that is in the current working directory** The script does test for the directory's existence and will print an error message if it is missing.

Processing to produce this dataset

The main directory contains four text files and two subdirectories. The data are provided in multiple files across two subdirectories, labelled *test* and *train*. Each of the subdirectories contains text files containing the preprocessed sensor data (e.g. X_test.txt) as well as two files (e.g. *subject_test.txt* and *y_test.txt*) containing the subject ids and activity ids for each line in the sensor data file.

The process:

- Read and combine the three test data files, along with a fourth which specifies the source of the data
- Repeat for the three training data files
- Merge the training and test data files into one
- Select the features (variables) that are either means or std.
- Convert the Activity variable to a factor with labels
- Rename the measurement (feature) variables
- Collapse that into a summary dataset containing the mean of the measurements for each subject and activity combination
- the data are dumped using a *write.file()* function.