Codebook – Data Cleaning Project

Ashok Natesan

In this project, we have derived 2 tidy data sets from the base data

- A consolidated tidied data set (X_consolidated.txt) from the training and test data sets, that retains 79 (mean and std deviation) metrics of the 561 metrics compiled by the UCI UAR project.
- A second derivied tidy data set (meanBySubjectActivity.txt) shows observation data that is grouped by Subject, Activity and summarizes (using mean) the 79 metrics from the X_consolidated.txt.

Both these files are generated by running the Run_Analysis.R script in the directory, "HARtidyDS".

Since much of this data is derived from the UCI HAR Datasets, which are well-documented at http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones, we refer the reader to that site for details on the specifics of the numeric UAR metrics. We provide an overview, primarily of additional fields or modifications we provided

X_consolidated.txt

This is the combined observation data across the Training and Test data set. The following are the main fields in this dataset

Field Name	Description	Value range and type
Index	Serial number – unique for each observation	1:10299 (integer)
Subject	Identifier for subject in UCI HAR test	1:30 (integer)
ActivityCode	ID associated with activity label	1:6 (integer) 1 WALKING 2 WALKING_UPSTAIRS 3 WALKING_DOWNSTAIRS 4 SITTING 5 STANDING

		6 LAYING
Activity	Label for activity, associated with Activity Code	6 values (character)
DatasetType	Indicates whether observation came from Training or Test data set	Character. Values are either “Train” or “Test”
tBodyAcc-mean()-X … fBodyBodyGyroJerkMag-std()	79 metrics which are exactly identical to the UCI UAR dataset values – in terms of types, semantics and values	Features are normalized and bounded within [-1,1]. Please refer to UCI HAR Dataset Features_info.txt and Readme.txt

meanBySubjectActivity.txt

This is a dataset derived for X_consolidated.txt. In this dataset, the UAR metrics are summarized, grouped by subject and activity and then the mean of each feature within the group is taken

Field Name	Description	Value range and type
Index	Serial number – unique for each observation	1:180 (integer)
Subject	Identifier for subject in UCI HAR test	1:30 (integer)
ActivityCode	ID associated with activity label	1:6 (integer) 1 WALKING 2 WALKING_UPSTAIRS 3 WALKING_DOWNSTAIRS 4 SITTING 5 STANDING 6 LAYING
Activity	Label for activity, associated with Activity Code	6 values (character)
tBodyAcc-mean()-X …	79 metrics which are exactly identical to the UCI UAR dataset values – in terms of types, semantics and values.	Features are normalized and bounded within [-1,1]. Please refer to UCI HAR Dataset

fBodyBodyGyroJerkMag-std()	In this dataset, the MEAN within the group (defined by subject and activity) is computed	Features_info.txt and Readme.txt