Ryan Yan
UCI HAR Dataset Codebook

I. Introduction

The UCI HAR data set contains data on 561 different features which were normalized and bounded within [-1, 1]. Data was recorded from 30 individual participants performing 6 different activities, then arbitrarily sectioned into 70% as the training set data and 30% for the test data set.

II. Files of significance:

Data and labels: "train/X_train.txt" – the numeric feature data of the training set, "train/y_train.txt" – the activity codes for the training set data, "test/X_test.txt" – the numeric feature data of the test set, "test/y_test.txt" – the activity codes for the test set data. "activity_labels.txt" – contains information used to translate numerical activity codes to readable labels (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING). "/train/subject_train.txt" and "/test/subject_test.txt" contain the subject numbers aligned properly for each respective data set.

Feature info: "features_info.txt" – information about the variables including names and descriptions. Refer to this file for information on the variable names. I will not be describing all of them here. "features.txt" – List of all features

III. Raw to tidy data steps:

The analysis uses the RScript 'run_analysis.R' which must be run with the 'UCI HAR Dataset' folder set as the working directory. The read.table() functions assume the original locations of each folder and subfolder has not been changed.

This script will read the following files into R: "./test/X_test.txt", "./train/X_train.txt", "./test/y_test.txt", "./train/y_train.txt", "./train/subject_train.txt", "./test/subject_test.txt", "./activity_labels.txt", and "./features.txt".

At first, the test and training set data is separated. The first modification to the raw data is made by using cbind() to merge the x_test, y_test, and subject_test tables together (and then repeated with the training files). This will create a table for both test and training sets, which will have the subject number (1-30), the activity code (1-6), and the data for each of the 561 features, in each of the two data tables. The function rbind() is then used to attach the test data table on the bottom of the training data table.

The activity_labels file is read into R as a factor variable and used to change the second column of the merged data set ("Activity_Label") from numeric to readable labels (eg. WALKING, SITTING, etc.)

The features file contains the names of each of the features (ie. Each column of data in the data table). From here, we extract the indices of variables with "mean" or "std" in their name, and use this to create a 'tidy_table' variable which contains only those columns out of the original 561 feature variables. This 'tidy_table' is not output anywhere but is held in the R environment.

The last section of the script puts together the data table 'means' of the dimensions 79 x 36. There are 79 variables related to "mean" or "std", and 30 subjects + 6 activities. This data table contains the average of each of the 79 variables for each subject as well as each of the 6 activities. This can be thought of as a summary of the data for each subject and for each activity. This 'mean' data table is then written into a txt file in your working directory using write.table() and rownames = FALSE.