GET DATA COURSE PROJECT

Dalip Sondhi

Sunday, May 24, 2015

Introduction

This project is based on the HAR Dataset archived in the UCI Machine Learning Repository. The website linked above notes that the UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

Process

The project required merger of the means and standard deviations data contained in the *test.txt* and *train.txt* data files. The following steps were followed in the R script file, *run_analysis.R*, to create the *har* data frame in R (see attached code book for the description of this data frame):

1. The data file *activity_labels.txt* was read to create the vector of the six activities stored as *activity*.
2. The data file *features.txt* was read to create the vector of all 561 variable names stored as *varname*.
3. The *grep* command was used to index the locations of all of the variable names in the vector *varname* that contained the substring *mean* and separately, the substring *std*. The vector *meanlist* contains the indexes of 53 means variables. Note that seven of these are related to the *angle* variable that used mean values of parametrs and were regarded as mean angle values. The vector *sdlist* contains the indexes of 33 standard-deviation variables. The *meanlist* and *sdlist* collectively index a total of 86 variable names. The variable names associated with these 86 variables were secured by *varname(meanlist)* and *varname(sdlist)*.
4. The *subject_test.txt* and *subject_train.txt* files were read. The subjects were identied by *subject_id* and although not necessary or required, a variable *subject_type* was added with factor levels of "test" and "train".
5. The *y_test.txt* and *y_train.txt* files were read to create the data frame *subjects.test* containing 2,947 rows for 9 test subjects based on six activity levels. The *subjects.train* data frame contains 7,352 rows for 21 train subjects. Both data frames contain three columns: *subject_id*, *subject_type*, and *activity*. Example: subject_id=2, subject_type=test, activity=STANDING.
6. The *X_test.txt* file was read containing 2,947 rows and 561 columns of measurements for the 9 test sujects. The *X_train.txt* file was read containing 7,352 rows and 561 colums of measurements for the 21 train subjects. The data frame *means.test* contains the 53 columns of means and *sd.test* contains the 33 standard deviation columns for the 9 test subjects. This was accomplished withe index vectors *meanlist* and *sdlist* determined in step 3 above.
7. The data frames *subjects.test*, *means.test*, and *sd.test* were attached with the *cbind* command to create the *merge.test* data frame containing 2,947 rows and 89 columns of data for the 9 test subjects. Likewise the *merge.train* data frame was created containing 7,352 rows and 89 columns of data for the 21 train subjects. These two data frames were combined using the *rbind* command to create the tidy dataset *har*. The final dataset *har* contains 10,299 rows and 89 columns. Each row represents a subject_id. Columns 1-3 contain the subject_id, subject_type, and activity. Columns 4-56 contain the mean values and columns 57-89 contain the standard deviation values.
8. In the final step, the *har* data frame was aggregated by subject_id and activity to write the *summary.txt* file that contains 180 rows (30 subjects x 6 activity levels) and the mean values of the 53 means and 33 standard deviations.