Getting and Cleaning Data: Course Project

Project completed by Catherine White for the Coursera class, Getting and Cleaning Data (https://www.coursera.org/course/getdata).

About the data

Origin

This repository contains code to generate a summary of a subset of the data from a Samsung accelerometer study on identifying human activity through a smart phone's accelerometer readings. The description of the experiment from the data's README file:

> The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

Full information about the study can be found here: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. The complete data set can be downloaded here: http://archive.ics.uci.edu/ml/machine-learning-databases/00240/.

If you use this data set, please cite the original source:

> Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.

Contents of the original set

The original data set contained 10299 observations of 30 test subjects wearing a smart phone at their waist. The observations in the data set are derived from the readings from an accelerometer and a gyroscope while a subject was performing one of 6 activities. The raw readings were processed in a number of ways before being released, as described in their README file:

> The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

The features obtained through this analysis are normalized to fall between -1 and 1 and appear to be unitless.

Organization of the original data set

The original data set stores the information needed to parse the data set across several files. The actual observations of the features are in the files `train/X_train.txt` and `test/X_test.txt`. Those files lack column names. The column names for the `X_` files are in `features.txt`.

The activities that the rows of the `X_` correspond to are stored in `train/y_train.txt` and `test/y_test.txt`. The activities are coded as numbers in the `y_` files. The activities that the numbers correspond to are listed in `activity_labels.txt`.

The subjects performing the actions for each observation are stored in `train/subject_train.txt` and `test/subject_test.txt`.

Contents of the processed data set

This repository contains an R script that reads in the feature information from the original data set and combines and reduces the data in such a way as to provide the average value of each feature's mean and standard deviation for each subject for each activity. Units are the same as in the original data set. The text file the script produces is named means_and_stds_for_all_subjects.txt.

Columns in means_and_stds_for_all_subjects.txt The original data set provides means and standard deviations of particular features (combinations of readings from the accelerometers and gyroscopes). The code in this repository takes those means and standard deviations and presents the average mean and average standard deviation for each feature for each subject and each activity.

- **subject_number**: The subject ID of the participant performing the actions. Participant IDs are integers from 1 to 30.
- **activity**: Which activity the subject was performing. The six activities are standing, sitting, laying down, walking, walking up stairs, and walking down stairs (in the file as [STANDING, SITTING, LAYING, WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS]).
- **average_t<feature>_mean**: The average value of the mean of the time domain feature named <feature>.
- **average_t<feature>_std**: The average value of the standard deviation about the mean value of the time domain feature, <feature>.
- **average_t<feature>_mean_<axis>**: The average value of the mean of the time domain feature named <feature> in the <axis> direction. Axis options are [X, Y, Z].
- **average_t<feature>_std_<axis>**: The average value of the standard deviation about the mean of the time domain feature named <feature> in the <axis> direction.
- **average_f<feature>_mean**: The average value of the mean of the frequency domain feature named <feature>.
- **average_f<feature>_std**: The average value of the standard deviation about the mean value of the frequency domain feature, <feature>.
- **average_f<feature>_mean_<axis>**: The average value of the mean of the frequency domain feature named <feature> in the <axis> direction.
- **average_f<feature>_std_<axis>**: The average value of the standard deviation about the mean of the frequency domain feature named <feature> in the <axis> direction.

Processing

The R script in this repository reads in several text files, aggregates the data, selects only the means and standard deviations for each feature, and finds the means for each subject for each activity of the means and standard deviations. This script assumes that the working directory is the directory of the unzipped data set from the UCI site linked above.

Sections of the code

1. *Get the names of the columns*: The data is stored in unlabeled text files. The first section of the script reads in the file that contains the column names for the files containing the observations of the features. Not all of columns are relevant, so we also extract only the column names containing "mean()" or "std()". The names are not formatted well for output with write.table, so the parentheses are removed and dashes are replaced with underscores in the column names.

2. *Pull in the data sets*: This section reads in the files containing the actual data as data.tables. The fread function allows the reading of only certain columns, so it only reads in the column numbers corresponding to means and standard deviations. It also brings in the files with the labels for the activities.

3. *Combine the tables*: This section converts the activites in the `y_` files to strings instead of activity codes and adds the subject numbers and activities to the main bodies of the data sets. It also combines the test and training data sets together and arranges by subject number.

4. *Make a separate data table with summary statistics*: This section takes the combined data.table and uses it to compute the mean for each subject and activity of each column (that isn't the subject number or activity). The combined data.table is grouped by the subject_number and activity. This means that using mutate() to add a column to the data set will perform the functions by group, i.e. unique combinations of subject_number and activity. It uses the pipeline operator to generate the mean of each column in turn, ungroup it (so the information can be separated from the subject number and ID), and add the means to the summary data set. The unique function at the end reduces the summary data set to only unique rows, one for each combination of subject ID and activity.