Codebook

Sanchit Sharma

Sunday, January 25, 2015

Course Project Code Book

Source of the original data:
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip
(https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip)

Original description: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
(http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones)

The attached R script (run_analysis.R) performs the following to clean up the data:

- Merges the training and test sets to create one data set, namely train/X_train.txt with test/X_test.txt, the result of which is a 10299x561 data frame, as in the original description (“Number of Instances: 10299” and “Number of Attributes: 561”), train/subject_train.txt with test/subject_test.txt, the result of which is a 10299x1 data frame with subject IDs, and train/y_train.txt with test/y_test.txt, the result of which is also a 10299x1 data frame with activity IDs.
- Reads features.txt and extracts only the measurements on the mean and standard deviation for each measurement. The result is a 10299x66 data frame, because only 66 out of 561 attributes are measurements on the mean and standard deviation. All measurements appear to be floating point numbers in the range (-1, 1).
- Reads activity_labels.txt and applies descriptive activity names to name the activities in the data set:

```
walking

walkingupstairs

walkingdownstairs

sitting

standing

laying
```

- The script also appropriately labels the data set with descriptive names: all feature names (attributes) and activity names are converted to lower case, underscores and brackets () are removed. Then it merges the 10299x66 data frame containing features with 10299x1 data frames containing activity labels and subject IDs. The result is saved as merged_clean_data.txt, a 10299x68 data frame such that the first column contains subject IDs, the second column activity names, and the last 66 columns are measurements. Subject IDs are integers between 1 and 30 inclusive. The names of the attributes are similar to the following:

```
tbodyacc-mean-x

tbodyacc-mean-y

tbodyacc-mean-z

tbodyacc-std-x

tbodyacc-std-y

tbodyacc-std-z

tgravityacc-mean-x

tgravityacc-mean-y
```

- Finally, the script creates a 2nd, independent tidy data set with the average of each measurement for each activity and each subject. The result is saved as data_set_with_the_averages.txt, a 180x68 data frame, where as before, the first column contains subject IDs, the second column contains activity names (see below), and then the averages for each of the 66 attributes are in columns 3…68. There are 30 subjects and 6 activities, thus 180 rows in this data set with averages.