CodeBook

Brett Taylor

February 13, 2015

## Study Design

### The steps to producing the data set.

*Processing Overview* The script was developed to properly load the data, and process it in a simplified method.

**1 Clean Feature names**

The process includes first, reading in the feature\_names, and cleaning them up. The feature names that were included in the data from the file features.txt had several issues, including duplicate names, embedding of characters that would not work well as column names, and abbreviations that are difficult to understand. A function, normalize.features() is created that replaces these issues with more detail. I also created a prefix *fid[0-9,0-9,0-9]* that includes the feature identifier from the source features.txt file. This allows for simplified tracing of the data back to the source.

**2 Read Source Data**  
The next step is to read in all of the source data files. The files that were required include the following:

|  |  |  |
| --- | --- | --- |
| File | Obs # | Description |
| subject\_train.txt | 7352 | This associates the subject identifier with the x and y training observations. |
| subject\_test.txt | 2947 | This associates the subject identifier with the x and y testing observations. |
| X\_train.txt | 7352 | 561 feature measurement columns for each observation. This is the data was sampled for ML training. |
| X\_test.txt | 2947 | 561 feature measurement columns for each observation. This is the data was sampled for testing the outcome of the ML training. |
| y\_train.txt | 7352 | This includes a single attribute that represents the activity code. This is the data was sampled for ML training. |
| y\_test.txt | 2947 | This includes a single attribute that represents the activity code. This is the data was sampled for testing the outcome of the ML training. |
| features.txt | 561 | This includes columns for feature identifier, and the feature description. |
| activity.txt | 6 | Columns include activity id, and activity description. |
| **activity\_summary.txt** | 180 | The final summarized output file that includes the average of the mean, and standard deviation features aggregated by Activity and Subject. |

**3 Combine & Merge data frames**

First combine the columns together from each of the data sets. This creats a test and training data set. Next combine the the test and training data set.

Output: data.frame all.data

**4 Label all.data**

Add the activity labels to the all.data dataframe.

Output: data.frame all.data.labeled

*5* **Select the correct variables for all.data.labeled** Only keep the columns that are the subject, activity, and those that include the mean or standard deviation. The decision was made to not include angle measurements because they are synthesis of other vectors that happen to have the term mean in their name. Use regular expressions to eliminate the other columns.

Output: data.frame all.data.labeled  
**6** Summarize all.data.labeled\_\_

Create a summary of the aggregates grouped by Activity and Subject. Apply to function mean() to all of the numeric measurement columns. Sort the data by Activity, and Subject

Output: data.frame tidy.data

**7 Write tidy.data to file**  
 This file is the final output that contains the summary data from the tidy.data dataframe. The activity\_summary file meets the requirements of the Course project 5 step (see above).

Output: file "activity\_summary.txt"

### Reference:

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

## Code Book

### Includes a list of the variables

## Functions

Function Description  
 mean Mean value  
 std Standard deviation  
 mad Median absolute value  
 max Largest values in array  
 min Smallest value in array  
 sma Signal magnitude area  
 energy Average sum of the squares   
 iqr Interquartile range  
 entropy Signal Entropy  
 arCoeff Autorregresion coefficients   
 correlation Correlation coefficient  
 maxFreqInd Largest frequency component  
 meanFreq Frequency signal weighted average  
 skewness Frequency signal Skewness   
 kurtosis Frequency signal Kurtosis  
 energyBand Energy of a frequency interval   
 angle Angle between two vectors

## Variables

The variables included in this data set follow a naming pattern that can help you understand its nature.

|  |  |
| --- | --- |
| Column | Description |
| variable\_name | This the name of the column displayed in the summary data set. |
| feature\_id | This is the identifier associated with the feature |
| domain: | na |
| measure | na |
| Function | na |
| dimension | na |

Variable.Name Feature.Id Domain  
 Activity <NA> <NA>  
 Subject <NA> <NA>  
 fid001\_timeBodyAcc\_mean\_X 001 time  
 fid002\_timeBodyAcc\_mean\_Y 002 time  
 fid003\_timeBodyAcc\_mean\_Z 003 time  
 fid004\_timeBodyAcc\_std\_X 004 time  
 fid005\_timeBodyAcc\_std\_Y 005 time  
 fid006\_timeBodyAcc\_std\_Z 006 time  
 fid041\_timeGravityAcc\_mean\_X 041 time  
 fid042\_timeGravityAcc\_mean\_Y 042 time  
 fid043\_timeGravityAcc\_mean\_Z 043 time  
 fid044\_timeGravityAcc\_std\_X 044 time  
 fid045\_timeGravityAcc\_std\_Y 045 time  
 fid046\_timeGravityAcc\_std\_Z 046 time  
 fid081\_timeBodyAccJerk\_mean\_X 081 time  
 fid082\_timeBodyAccJerk\_mean\_Y 082 time  
 fid083\_timeBodyAccJerk\_mean\_Z 083 time  
 fid084\_timeBodyAccJerk\_std\_X 084 time  
 fid085\_timeBodyAccJerk\_std\_Y 085 time  
 fid086\_timeBodyAccJerk\_std\_Z 086 time  
 fid121\_timeBodyGyro\_mean\_X 121 time  
 fid122\_timeBodyGyro\_mean\_Y 122 time  
 fid123\_timeBodyGyro\_mean\_Z 123 time  
 fid124\_timeBodyGyro\_std\_X 124 time  
 fid125\_timeBodyGyro\_std\_Y 125 time  
 fid126\_timeBodyGyro\_std\_Z 126 time  
 fid161\_timeBodyGyroJerk\_mean\_X 161 time  
 fid162\_timeBodyGyroJerk\_mean\_Y 162 time  
 fid163\_timeBodyGyroJerk\_mean\_Z 163 time  
 fid164\_timeBodyGyroJerk\_std\_X 164 time  
 fid165\_timeBodyGyroJerk\_std\_Y 165 time  
 fid166\_timeBodyGyroJerk\_std\_Z 166 time  
 fid201\_timeBodyAccMag\_mean 201 time  
 fid202\_timeBodyAccMag\_std 202 time  
 fid214\_timeGravityAccMag\_mean 214 time  
 fid215\_timeGravityAccMag\_std 215 time  
 fid227\_timeBodyAccJerkMag\_mean 227 time  
 fid228\_timeBodyAccJerkMag\_std 228 time  
 fid240\_timeBodyGyroMag\_mean 240 time  
 fid241\_timeBodyGyroMag\_std 241 time  
 fid253\_timeBodyGyroJerkMag\_mean 253 time  
 fid254\_timeBodyGyroJerkMag\_std 254 time  
 fid266\_frequencyBodyAcc\_mean\_X 266 frequency  
 fid267\_frequencyBodyAcc\_mean\_Y 267 frequency  
 fid268\_frequencyBodyAcc\_mean\_Z 268 frequency  
 fid269\_frequencyBodyAcc\_std\_X 269 frequency  
 fid270\_frequencyBodyAcc\_std\_Y 270 frequency  
 fid271\_frequencyBodyAcc\_std\_Z 271 frequency  
 fid294\_frequencyBodyAcc\_meanFreq\_X 294 frequency  
 fid295\_frequencyBodyAcc\_meanFreq\_Y 295 frequency  
 fid296\_frequencyBodyAcc\_meanFreq\_Z 296 frequency  
 fid345\_frequencyBodyAccJerk\_mean\_X 345 frequency  
 fid346\_frequencyBodyAccJerk\_mean\_Y 346 frequency  
 fid347\_frequencyBodyAccJerk\_mean\_Z 347 frequency  
 fid348\_frequencyBodyAccJerk\_std\_X 348 frequency  
 fid349\_frequencyBodyAccJerk\_std\_Y 349 frequency  
 fid350\_frequencyBodyAccJerk\_std\_Z 350 frequency  
 fid373\_frequencyBodyAccJerk\_meanFreq\_X 373 frequency  
 fid374\_frequencyBodyAccJerk\_meanFreq\_Y 374 frequency  
 fid375\_frequencyBodyAccJerk\_meanFreq\_Z 375 frequency  
 fid424\_frequencyBodyGyro\_mean\_X 424 frequency  
 fid425\_frequencyBodyGyro\_mean\_Y 425 frequency  
 fid426\_frequencyBodyGyro\_mean\_Z 426 frequency  
 fid427\_frequencyBodyGyro\_std\_X 427 frequency  
 fid428\_frequencyBodyGyro\_std\_Y 428 frequency  
 fid429\_frequencyBodyGyro\_std\_Z 429 frequency  
 fid452\_frequencyBodyGyro\_meanFreq\_X 452 frequency  
 fid453\_frequencyBodyGyro\_meanFreq\_Y 453 frequency  
 fid454\_frequencyBodyGyro\_meanFreq\_Z 454 frequency  
 fid503\_frequencyBodyAccMag\_mean 503 frequency  
 fid504\_frequencyBodyAccMag\_std 504 frequency  
 fid513\_frequencyBodyAccMag\_meanFreq 513 frequency  
 fid516\_frequencyBodyAccJerkMag\_mean 516 frequency  
 fid517\_frequencyBodyAccJerkMag\_std 517 frequency  
 fid526\_frequencyBodyAccJerkMag\_meanFreq 526 frequency  
 fid529\_frequencyBodyGyroMag\_mean 529 frequency  
 fid530\_frequencyBodyGyroMag\_std 530 frequency  
 fid539\_frequencyBodyGyroMag\_meanFreq 539 frequency  
 fid542\_frequencyBodyGyroJerkMag\_mean 542 frequency  
 fid543\_frequencyBodyGyroJerkMag\_std 543 frequency  
 fid552\_frequencyBodyGyroJerkMag\_meanFreq 552 frequency  
 Measure Function Dimension  
 <NA> <NA> <NA>  
 <NA> <NA> <NA>  
 BodyAcc mean X  
 BodyAcc mean Y  
 BodyAcc mean Z  
 BodyAcc std X  
 BodyAcc std Y  
 BodyAcc std Z  
 GravityAcc mean X  
 GravityAcc mean Y  
 GravityAcc mean Z  
 GravityAcc std X  
 GravityAcc std Y  
 GravityAcc std Z  
 BodyAccJerk mean X  
 BodyAccJerk mean Y  
 BodyAccJerk mean Z  
 BodyAccJerk std X  
 BodyAccJerk std Y  
 BodyAccJerk std Z  
 BodyGyro mean X  
 BodyGyro mean Y  
 BodyGyro mean Z  
 BodyGyro std X  
 BodyGyro std Y  
 BodyGyro std Z  
 BodyGyroJerk mean X  
 BodyGyroJerk mean Y  
 BodyGyroJerk mean Z  
 BodyGyroJerk std X  
 BodyGyroJerk std Y  
 BodyGyroJerk std Z  
 BodyAccMag mean <NA>  
 BodyAccMag std <NA>  
 GravityAccMag mean <NA>  
 GravityAccMag std <NA>  
 BodyAccJerkMag mean <NA>  
 BodyAccJerkMag std <NA>  
 BodyGyroMag mean <NA>  
 BodyGyroMag std <NA>  
 BodyGyroJerkMag mean <NA>  
 BodyGyroJerkMag std <NA>  
 BodyAcc mean X  
 BodyAcc mean Y  
 BodyAcc mean Z  
 BodyAcc std X  
 BodyAcc std Y  
 BodyAcc std Z  
 BodyAcc meanFreq X  
 BodyAcc meanFreq Y  
 BodyAcc meanFreq Z  
 BodyAccJerk mean X  
 BodyAccJerk mean Y  
 BodyAccJerk mean Z  
 BodyAccJerk std X  
 BodyAccJerk std Y  
 BodyAccJerk std Z  
 BodyAccJerk meanFreq X  
 BodyAccJerk meanFreq Y  
 BodyAccJerk meanFreq Z  
 BodyGyro mean X  
 BodyGyro mean Y  
 BodyGyro mean Z  
 BodyGyro std X  
 BodyGyro std Y  
 BodyGyro std Z  
 BodyGyro meanFreq X  
 BodyGyro meanFreq Y  
 BodyGyro meanFreq Z  
 BodyAccMag mean <NA>  
 BodyAccMag std <NA>  
 BodyAccMag meanFreq <NA>  
 BodyAccJerkMag mean <NA>  
 BodyAccJerkMag std <NA>  
 BodyAccJerkMag meanFreq <NA>  
 BodyGyroMag mean <NA>  
 BodyGyroMag std <NA>  
 BodyGyroMag meanFreq <NA>  
 BodyGyroJerkMag mean <NA>  
 BodyGyroJerkMag std <NA>  
 BodyGyroJerkMag meanFreq <NA>

This activity\_summary.txt file has a total of 180 observations, and 81 columns.

### License

Use of this dataset in publications must be acknowledged by referencing the following publication [1]

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

This dataset is distributed AS-IS and no responsibility implied or explicit can be addressed to the authors or their institutions for its use or misuse. Any commercial use is prohibited.

Jorge L. Reyes-Ortiz, Alessandro Ghio, Luca Oneto, Davide Anguita. November 2012.