CodeBook

Brett Taylor

February 22, 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. The following are the steps taken to retrieve, and clean the data. The script that performed this process is found in the file "run\_analysis.txt"

**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 label displayed in the activity\_summary.txt data set. |
| feature\_id | This is the identifier associated with the feature.txt file. It will allow you to trace back to the source column in the X\_(train |
| domain: | the measures were are either in the time or frequency domain. This was translated from -f for frequency, or a -t time. The frequency domain was converted from the signal data by fourier transform in the initial data sets. |
| signal | The measure is based on the following values {BodyAcc,BodyAccJerk, BodyGyro,BodyGyroJerk,BodyAccMag,BodyAccJerk,BodyGyroMag,BodyGyroJerkMag,GravityAcc,GravityAccMag}. Acc==Acceleration, Gyro=Gyration |
| Function | The functions applied to the signals. The list that were selected are mean(),std(),meanFreq(). |
| dimension | This indicates the direction of the vector. values are {X,Y,Z} |

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

### Primary Variables

* **Subject**
* This is an integer in the range 1:30 of the subjects that were engaged in the experiment.
* **Activity** This is the type of action the subject was performing during the measurement. Values range from {Laying,Sitting,Standing, Walking,Walking\_Downstairs,Walking\_Upstairs}

### Signal Variable Descriptions - Variable, Feature.Id and Dimension

The following is a description of the fields:

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

### Signal Variable Descriptions (cont.) - Variable, Domain and Signal

## Variable.Name Domain Signal  
## Activity <NA> <NA>  
## Subject <NA> <NA>  
## fid001\_timeBodyAcc\_mean\_X time BodyAcc  
## fid002\_timeBodyAcc\_mean\_Y time BodyAcc  
## fid003\_timeBodyAcc\_mean\_Z time BodyAcc  
## fid004\_timeBodyAcc\_std\_X time BodyAcc  
## fid005\_timeBodyAcc\_std\_Y time BodyAcc  
## fid006\_timeBodyAcc\_std\_Z time BodyAcc  
## fid041\_timeGravityAcc\_mean\_X time GravityAcc  
## fid042\_timeGravityAcc\_mean\_Y time GravityAcc  
## fid043\_timeGravityAcc\_mean\_Z time GravityAcc  
## fid044\_timeGravityAcc\_std\_X time GravityAcc  
## fid045\_timeGravityAcc\_std\_Y time GravityAcc  
## fid046\_timeGravityAcc\_std\_Z time GravityAcc  
## fid081\_timeBodyAccJerk\_mean\_X time BodyAccJerk  
## fid082\_timeBodyAccJerk\_mean\_Y time BodyAccJerk  
## fid083\_timeBodyAccJerk\_mean\_Z time BodyAccJerk  
## fid084\_timeBodyAccJerk\_std\_X time BodyAccJerk  
## fid085\_timeBodyAccJerk\_std\_Y time BodyAccJerk  
## fid086\_timeBodyAccJerk\_std\_Z time BodyAccJerk  
## fid121\_timeBodyGyro\_mean\_X time BodyGyro  
## fid122\_timeBodyGyro\_mean\_Y time BodyGyro  
## fid123\_timeBodyGyro\_mean\_Z time BodyGyro  
## fid124\_timeBodyGyro\_std\_X time BodyGyro  
## fid125\_timeBodyGyro\_std\_Y time BodyGyro  
## fid126\_timeBodyGyro\_std\_Z time BodyGyro  
## fid161\_timeBodyGyroJerk\_mean\_X time BodyGyroJerk  
## fid162\_timeBodyGyroJerk\_mean\_Y time BodyGyroJerk  
## fid163\_timeBodyGyroJerk\_mean\_Z time BodyGyroJerk  
## fid164\_timeBodyGyroJerk\_std\_X time BodyGyroJerk  
## fid165\_timeBodyGyroJerk\_std\_Y time BodyGyroJerk  
## fid166\_timeBodyGyroJerk\_std\_Z time BodyGyroJerk  
## fid201\_timeBodyAccMag\_mean time BodyAccMag  
## fid202\_timeBodyAccMag\_std time BodyAccMag  
## fid214\_timeGravityAccMag\_mean time GravityAccMag  
## fid215\_timeGravityAccMag\_std time GravityAccMag  
## fid227\_timeBodyAccJerkMag\_mean time BodyAccJerkMag  
## fid228\_timeBodyAccJerkMag\_std time BodyAccJerkMag  
## fid240\_timeBodyGyroMag\_mean time BodyGyroMag  
## fid241\_timeBodyGyroMag\_std time BodyGyroMag  
## fid253\_timeBodyGyroJerkMag\_mean time BodyGyroJerkMag  
## fid254\_timeBodyGyroJerkMag\_std time BodyGyroJerkMag  
## fid266\_frequencyBodyAcc\_mean\_X frequency BodyAcc  
## fid267\_frequencyBodyAcc\_mean\_Y frequency BodyAcc  
## fid268\_frequencyBodyAcc\_mean\_Z frequency BodyAcc  
## fid269\_frequencyBodyAcc\_std\_X frequency BodyAcc  
## fid270\_frequencyBodyAcc\_std\_Y frequency BodyAcc  
## fid271\_frequencyBodyAcc\_std\_Z frequency BodyAcc  
## fid294\_frequencyBodyAcc\_meanFreq\_X frequency BodyAcc  
## fid295\_frequencyBodyAcc\_meanFreq\_Y frequency BodyAcc  
## fid296\_frequencyBodyAcc\_meanFreq\_Z frequency BodyAcc  
## fid345\_frequencyBodyAccJerk\_mean\_X frequency BodyAccJerk  
## fid346\_frequencyBodyAccJerk\_mean\_Y frequency BodyAccJerk  
## fid347\_frequencyBodyAccJerk\_mean\_Z frequency BodyAccJerk  
## fid348\_frequencyBodyAccJerk\_std\_X frequency BodyAccJerk  
## fid349\_frequencyBodyAccJerk\_std\_Y frequency BodyAccJerk  
## fid350\_frequencyBodyAccJerk\_std\_Z frequency BodyAccJerk  
## fid373\_frequencyBodyAccJerk\_meanFreq\_X frequency BodyAccJerk  
## fid374\_frequencyBodyAccJerk\_meanFreq\_Y frequency BodyAccJerk  
## fid375\_frequencyBodyAccJerk\_meanFreq\_Z frequency BodyAccJerk  
## fid424\_frequencyBodyGyro\_mean\_X frequency BodyGyro  
## fid425\_frequencyBodyGyro\_mean\_Y frequency BodyGyro  
## fid426\_frequencyBodyGyro\_mean\_Z frequency BodyGyro  
## fid427\_frequencyBodyGyro\_std\_X frequency BodyGyro  
## fid428\_frequencyBodyGyro\_std\_Y frequency BodyGyro  
## fid429\_frequencyBodyGyro\_std\_Z frequency BodyGyro  
## fid452\_frequencyBodyGyro\_meanFreq\_X frequency BodyGyro  
## fid453\_frequencyBodyGyro\_meanFreq\_Y frequency BodyGyro  
## fid454\_frequencyBodyGyro\_meanFreq\_Z frequency BodyGyro  
## fid503\_frequencyBodyAccMag\_mean frequency BodyAccMag  
## fid504\_frequencyBodyAccMag\_std frequency BodyAccMag  
## fid513\_frequencyBodyAccMag\_meanFreq frequency BodyAccMag  
## fid516\_frequencyBodyAccJerkMag\_mean frequency BodyAccJerkMag  
## fid517\_frequencyBodyAccJerkMag\_std frequency BodyAccJerkMag  
## fid526\_frequencyBodyAccJerkMag\_meanFreq frequency BodyAccJerkMag  
## fid529\_frequencyBodyGyroMag\_mean frequency BodyGyroMag  
## fid530\_frequencyBodyGyroMag\_std frequency BodyGyroMag  
## fid539\_frequencyBodyGyroMag\_meanFreq frequency BodyGyroMag  
## fid542\_frequencyBodyGyroJerkMag\_mean frequency BodyGyroJerkMag  
## fid543\_frequencyBodyGyroJerkMag\_std frequency BodyGyroJerkMag  
## fid552\_frequencyBodyGyroJerkMag\_meanFreq frequency BodyGyroJerkMag

## Note

It was attended that the variable descriptions would be in a single table. This had to be modified due to issues with the ability to fit the six columns in a single table. This will be researched in the future.

# 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.