CodeBook

========================================================

Data description

The data for this assignment comes from the Human Activity Recognition Using Smartphones Dataset, Version 1.0, collected by Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, available at

<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>.

The data was collected on 30 volunteers whose movements were measured to train and test models to predict in which of six "human activities" a subject was engaged:

Walking

Walking upstairs

Walking downstairs

Sitting

Standing

Laying

The subjects were fitted with a Body Accelerometer and a Body Gyroscope. Each subject engaged in all six activities while being monitored among 17 different measured variables. The available data sets contain the following variable prefixes:

tBodyAcc-XYZ

tGravityAcc-XYZ

tBodyAccJerk-XYZ

tBodyGyro-XYZ

tBodyGyroJerk-XYZ

tBodyAccMag

tGravityAccMag

tBodyAccJerkMag

tBodyGyroMag

tBodyGyroJerkMag

fBodyAcc-XYZ

fBodyAccJerk-XYZ

fBodyGyro-XYZ

fBodyAccMag

fBodyAccJerkMag

fBodyGyroMag

fBodyGyroJerkMag

where "X", "Y", and "Z" refer to spatial planes in which the variable was measured.

The measures for each variable captured in the data were:

mean(): Mean value

std(): Standard deviation

mad(): Median absolute deviation

max(): Largest value in array

min(): Smallest value in array

sma(): Signal magnitude area

energy(): Energy measure. Sum of the squares divided by the number of values.

iqr(): Interquartile range

entropy(): Signal entropy

arCoeff(): Autorregresion coefficients with Burg order equal to 4

correlation(): correlation coefficient between two signals

maxInds(): index of the frequency component with largest magnitude

meanFreq(): Weighted average of the frequency components to obtain a mean frequency

skewness(): skewness of the frequency domain signal

kurtosis(): kurtosis of the frequency domain signal

bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.

angle(): Angle between two vectors.

In total, each observation reflects one subject engaged in one activity, and consists of 561 measures labeled as "[variable prefix]-[measure]", eg, the first variable is labeled "tBodyAcc-X-mean()", the second "tBodyAcc-X-std()", etc.

The measured data sets are split into training and testing subsets in roughly a 70/30 ratio. While more data sets were available, for this project I used the following eight:

X\_train X\_test

y\_train y\_test

subject\_train subject\_test

activity\_labels

features

Study design

========================================================

run\_analysis.R

The script run\_analysis.R first reads in all eight data sets, and names the 561 columns in the train and test data sets with the 561 measures listed in the features data set.

Next, it creates subsets of the train and test data sets, selecting only the columns with either "-mean()" or "-sd()" in their name.

Next, it binds the appropriate subject ID vector and verbose activity label vector to the beginning of both the train and test data sets.

Next, it merges the train and test data sets to create a single data set. The two data sets' column variables are identical so there is no massaging to be done.

Next, it reduces the merged data only to the subject IDs, activity labels, and variables containing measure "-mean()". The resulting data set is called meanData.

Since the suffix "-mean()" is now superfluous, the final step before writing the data is to remove "-mean()" from the column names.