About the Data

The data for this assignment represent data collected from the accelerometers from the Samsung Galaxy S smartphone. Experiments have been carried out with a group of 30 volunteers. Each person performed six activities (walking, walking upstairs, walking downstairs, sitting, standing and laying) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, the experiment captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments had been video-recorded to label the data manually. The obtained dataset was randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data.

For each record it is provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
- Triaxial Angular velocity from the gyroscope.
- A 561-feature vector with time and frequency domain variables.
- Its activity label.
- An identifier of the subject who carried out the experiment.

More data about the experiment, the data and how they were obtained is available in the readme.md file included in the repo.

Steps performed to clean data

1. Download and save the data from the UCI HAR Dataset in a separate folder named "UCI HAR Dataset" in the working directory. A browse through the README file is very helpful going forward.
2. Read data from the train folder into a "train" dataset; include both x and y variables in this dataset, since the y dataset (from original data) represents the activity variable, which is an important part of the final step. Also remember to include the data from the "subject.txt" file, since that will also be required later. I have used cbind() and created separate columns for each feature, another one for the subject ID and yet another for the activity number.
3. Repeat step 1 for data from the "test" folder, also remembering to include the y dataset for the activity variable and data from the "subject.txt" file, to enable subject tagging.
4. We then need to label each of the "test" and "train" datasets with data from the "features.txt" file. An important note while doing this – R does not recognise several characters; a few include "(", ")" and "–". A quick scan through the "features.txt" file reveals that these characters are frequently used in the document, and would therefore need to be replaced. I used a simple replacement function gsub() to search and replace these characters with "", so that they retain the purpose of labelling an activity and at the same time do not interfere with R's processing. For example, a variable originally names "tBodyAccmax()-X" would now read as "tBodyAccmaxX". The resultant string was used to assign column names to both the "test" and the "train" datasets.

5. Use a simple rbind() command to combine the two datasets. A good tip is to run quick function like str() or dim() or summary() on both the "train" and the "test" datasets to verify that they do indeed have the same number of columns. At this point, the dataset should have 561 columns (for the features), 1 column for the activity variable and 1 column for the subject ID – in total 563 columns. If there's an error here, this is a great spot to go back and clear it up.

6. After this, I read in the library for the dplyr package (library(dplyr)), so I can play around the data a lot easily. I then created the combined dataset into a tbl_df; in the current assignment, this tbl_df is named all_df.

7. In the next step, we will create a new dataset that includes ONLY the variables we need – those that pertain to the mean or standard deviation. I have assumed that a variable recorded under MeanFreq should be included in this new dataset, which is actually a mean of the frequency measures. I also realised that there is an error in the labelling of the original dataset because of which I was unable to use the select() and contains() functions properly. Instead of having to go over and resolve every error in the original dataset, I combined the select function with grep("mean|std") to create this dataset using a simple subsetting function. I've also included the subject ID and activity number columns, since these are required to finally break the data down.

8. At this point, I used the head() and tail() functions to review the data and realised that each subject performed each activity – and the subject may have done so several times. So I used dplyr's versatile group_by() function to group the dataset by both the subject ID and the activity number, which I could then use to summarise the data.

9. The next function call was summarise_each() with a mean() function, which allowed me to summarise the entire data by the grouped variables (subject ID and activity number)

10. Finally, I read in the data from "Activity_Labels.txt" to create an additional column that named the activity involved. To end the cleaning up, placed the columns in the order of subject ID, activity number, activity name and then all the relevant features.

11. The tidy data will present 1 row per subject per activity – a total of 30 x 6 rows – and one column each for subject ID, activity number and activity name and columns with mean variables for each feature pertaining to a mean or standard deviation reading from the original dataset.

12. A file named "Tidy_Data.txt" has been created using the write.table() command using row.name=FALSE.

About the Variables

Sub_ID

Description: Subject ID pertaining to the participant involved in the experiment

Values: 1 – 30

Act_Num

Description: Activity number pertaining to the activity undertaken by the test participant, subsequently mapped to column "Activity"

Values: 1 – 6

1: WALKING

2: WALKING_UPSTAIRS

3: WALKING_DOWNSTAIRS

4: SITTING

5: STANDING

6: LAYING

Activity

Description: Detailed activity name pertaining to the activity undertaken by the test participant; mapped against the value in variable Act_Num

Values: WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING

tBodyAccmeanX

tBodyAccmeanY

tBodyAccmeanZ

Description: Mean value of Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyAccstdX

tBodyAccstdY

tBodyAccstdZ

Description: Standard deviation value of Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tGravityAccmeanX

tGravityAccmeanY

tGravityAccmeanZ

Description: Mean value of Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Gravity" acceleration signal

Values: Normalised value within [-1,1]

tGravityAccstdX

tGravityAccstdY

tGravityAccstdZ

Description: Standard deviation value of Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Gravity" acceleration signal

Values: Normalised value within [-1,1]

tBodyAccJerkmeanX

tBodyAccJerkmeanY

tBodyAccJerkmeanZ

Description: Mean value of Jerk signals derived from the Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyAccJerkstdX

tBodyAccJerkstdY

tBodyAccJerkstdZ

Description: Standard deviation value of Jerk signals derived from the Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyGyromeanX

tBodyGyromeanY

tBodyGyromeanZ

Description: Mean value of Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyGyrostdX

tBodyGyrostdY

tBodyGyrostdZ

Description: Standard deviation value of Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyGyroJerkmeanX

tBodyGyroJerkmeanY

tBodyGyroJerkmeanZ

Description: Mean value of Jerk signals derived from Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyGyroJerkstdX

tBodyGyroJerkstdY

tBodyGyroJerkstdZ

Description: Standard deviation value of Jerk signals derived from Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

tBodyAccMagmean

Description: Mean value of Magnitude of 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyAccMagstd

Description: Standard deviation value of Magnitude of 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tGravityAccMagmean

Description: Mean value of Magnitude of 3-dimensional signals obtained from the Accelerometer split into a Gravity acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tGravityAccMagstd

Description: Standard deviation value of Magnitude of 3-dimensional signals obtained from the Accelerometer split into a Gravity acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyAccJerkMagmean

Description: Mean value of Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyAccJerkMagstd

Description: Standard deviation value of Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyGyroMagmean

Description: Mean value of Magnitude of 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyGyroMagstd

Description: Standard deviation value of Magnitude of 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyGyroJerkMagmean

Description: Mean value of Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

tBodyGyroJerkMagstd

Description: Standard deviation value of Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyAccmeanX

fBodyAccmeanY

fBodyAccmeanZ

Description: Mean value of FFT applied to Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccstdX

fBodyAccstdY

fBodyAccstdZ

Description: Standard deviation value of FFT applied to Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccmeanFreqX

fBodyAccmeanFreqY

fBodyAccmeanFreqZ

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccJerkmeanX

fBodyAccJerkmeanY

fBodyAccJerkmeanZ

Description: Mean value of FFT applied to Jerk signals derived from the Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccJerkstdX

fBodyAccJerkstdY

fBodyAccJerkstdZ

Description: Standard deviation value of FFT applied to Jerk signals derived from the Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccJerkmeanFreqX

fBodyAccJerkmeanFreqY

fBodyAccJerkmeanFreqZ

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Jerk signals derived from the Accelerometer signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyGyromeanX

fBodyGyromeanY

fBodyGyromeanZ

Description: Mean value of FFT applied to Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyGyrostdX

fBodyGyrostdY

fBodyGyrostdZ

Description: Standard deviation value of FFT applied to Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyGyromeanFreqX

fBodyGyromeanFreqY

fBodyGyromeanFreqZ

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Gyroscope signals for the "X","Y" or "Z" axis, separated into a "Body" acceleration signal

Values: Normalised value within [-1,1]

fBodyAccMagmean

Description: Mean value of Magnitude of 3-dimensional signals obtained by applying an FFT to the Accelerometer values split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyAccMagstd

Description: Standard deviation value of Magnitude of 3-dimensional signals obtained by applying an FFT to the Accelerometer values split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyAccMagmeanFreq

Description: Weighted average of the frequency components to obtain a mean frequency of Magnitude of 3-dimensional signals obtained by applying an FFT to the Accelerometer values split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyAccJerkMagmean

Description: Mean value of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyAccJerkMagstd

Description: Standard deviation value of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyAccJerkMagmeanFreq

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Accelerometer split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroMagmean

Description: Mean value of FFT applied to Magnitude of 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroMagstd

Description: Standard deviation value of FFT applied to Magnitude of 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroMagmeanFreq

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Magnitude of 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroJerkMagmean

Description: Mean value of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroJerkMagstd

Description: Standard deviation value of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]

fBodyBodyGyroJerkMagmeanFreq

Description: Weighted average of the frequency components to obtain a mean frequency of FFT applied to Magnitude of Jerk signals derived from the 3-dimensional signals obtained from the Gyroscope split into a Body acceleration signal using the Euclidean norm

Values: Normalised value within [-1,1]