Code Book for run_analysis.R

Introduction

This code book is for the run_analysis.R script developed for the Data Sharing course project. The goal of the project is to write a script that takes unstructured data and create a tidy data set. According to the paper "Tidy Date" written by Hadley Wickham (J. Stat. Software, 59(10), 2014), "Tidy datasets provide a standaradized way to link the structure of a dataset (its physical layout) with its semantics (its meaning)." To accomplish this, Wickham further says that "A dataset is a collection of values.... Values are organized in two ways. Every value belongs to a variable and an observation." To keep this standard, a data table that contains tidy data has three attributes (Wickham):

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

It is following these principles that the structure of the output file that results from running the run_analysis.R script is derived.

Input files:

The data was provided free of charge and as is by:

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

To understand the operation of the script, we must first understand the structure of the input files. The run_analysis script reads in 8 separate text files from the data source. The input files and a brief description follow:

- features.txt: List of all features.

- activity_labels.txt: Establishes a numerical value to the activity name.

- X_train.txt: Training set.

- y_train.txt: Training labels.

- X_test.txt: Test set.

- y_test.txt: Test labels.

- subject_test.txt: A list of the subjects (identified by a unique number)

- subject_train.txt: A list of the subjectw (identified by a unique number)

The input data is read into data tables with the following names

features:

This is a table of 561 observation types and 2 variables. Information about the features can be found in the features_info.txt file included with the data. The data is a collection of output from the accelerometer (ACC) and gyroscope (Gyro) features of a Samsung Galaxy II smart phone. From this data, other variables are calculated from the linear acceleration and angular velocity. The leading "t" indicates the data is collected in the time domain. The leading "f" indicates that the data has been subjected to a Fast Fourier Transform, and mathematical operation that transforms data from the time domain to the frequency domain.

Based on this, only the data tBodyAcc-X, tBodyAcc-Y, tBodyAcc-Z, tBodyGyro-X, tBodyGyro-Y, and tbodyGyro-Z are the raw data collected. All other variable types are derived from these data. The README.txt and the features_info.txt included with the data provide a general description of how this data is transformed into the other variables; however, the algorithmic details are not provided. Contained in the Inertial Signals subdirectory to the test and train subdirectories, the raw body_acc and body_gyro data is provided. Again, the files included with the data do not include the algorithmic details that allow reproduction of the calculated data that is part of the X_train and X_test data. For this reason, it was decided that the Tidy data set that is produced by the Script would include all of the different variables listed in the features table (focused on those with "mean" or "std" in the descriptive name.

Variables with "Acc" in the name have standard gravity units (g). Variables with "Gyro" have units radians/second.

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

The data labeled "*BodyACC*" and "*BodyGyro*" is the raw data from the accelerometer and the gyroscope. The other data is derived from these quantities. For each of the above variables, the following quantities were determined and included in the input dataset

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 to vectors.

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean
tBodyAccMean
tBodyAccJerkMean
tBodyGyroMean
tBodyGyroJerkMean

values in this table are normalized

activity_label:

This table associates an integer with each of the types of activities that the subjects conducted in during the measurements. These are:

1 WALKING
2 WALKING_UPSTAIRS
3 WALKING_DOWNSTAIRS
4 SITTING
5 STANDING
6 LAYING

X_train:

This is a table of 7352 observations of 561 variables. The 561 variables correspond to the 561 observation types in the features table, and the 7352 observations correspond to multiple measurements of the features conducted on 21 subjects performing the 6 activities. Each row corresponds to a single activity perfomed by one subject. This table has no column or row labels. The subjects in the train group are unique from those of the train group and were adults ages 19-48 randomly selected.

X_test:

This is a table of 2947 observations of 561 variables. The 561 variables correspond to the 561 observation types in teh features table, and teh 2947 observations correspons to multiple measurements of the features conducted on 9 subjects performing the 6 activities. Each row corresponds to a single activity performed by one subject. This table has no column or row labels. The subjects in the test group are unique from those of the train group and were adults ages 19-48 randomly selected.

y_train:

This is a dataset of 7352 observations of 1 variable. The 7352 observations correspond to the activity type (e.g. 1 == WALKING) that the subject was participating in when the observations in the rows of the X_train table were being collected. Values in this table are limited to 1 - 6.

y_test:

This is a dataset of 2947 observations of 1 variable. The 2947 observations correspond to the activity type (e.g. 1 == WALKING) that the subject was participating in when the observations in teh rows of the X_test table were bding collected. Values in this table are limited to 1 - 6.

subject_train:

This is a dataset of 7352 observations of 1 variable. The 7352 observations correspond to the subject who was participating in the activity that lead to the observations in the rows of the X_train table. For the training data set, there were 21 different subjects chosen randomly from a group of 30 total subjects.

subject_test:

This is a dataset of 2947 observations of 1 variable. The 2947 observations correspond to the suject who was participating in the activity that lead to the observations in the rows of the X_test table. For the test data set, there were 9 different subjects choosen randomly from a group of 30 total subjects.

Input Summary:

Based on the size of the input files, the script is able to align the activity type and the unique subject with each of the 7352 and 2047 observations in the train and the test groups. This accumulation of observations with unique variable results in a table with 563 columns (e.g. variables) and 10299 rows (e.g. observations) - a total of 5,798,377 elements of the table.

Script operation summary:

The run_analysis.R script reads the data into data tables. The script first adds a header to the X_test and X_train data tables to provide a unique descriptive name to each variable. The column names are from the features data table. It then adds a column to the X_test and X_train data tables that corresponds to the activity labels (from y-test and y_train respectively) to identify the type of activity the subject was performing while a particular observation was made. Similarly, another column (from subject_test and subject_train, respectively) is added that adds the subject labels to the X_test and X_train data table to identify the subject who performed the activity while the observation was being made.

The X_test and X_train data tables are then combined using the rbind() function to create a single data table (complete_data) of all the observations of all the variables.

The data table is trimmed by selecting only the data for variables that have "mean" and "std" (mean and standard deviations). This trims the data table to 10299 observations of only 82 variables - a total of 844,518 elements in the table.

Finally, the script obtains an average value for each variable for each activity and each subject. There are 6 activities and 30 subjects. This results in a table (complete_data_average) that has 81 variables and 180 unique observations - for a total of 14,580 total elements in the table. This represents a data size that is 0.25% the size of the input file.

Output file, Samsung_data_tidy.txt

The output file, Samsung_data_tidy.txt is the output file that is written to disk by the script. This file should be read back into the R environment using the read.table("./Samsung_data_tidy.txt", header=TRUE). This operation will result in a data table with 180 observations of 81 variables. The variables are:

```
"1" "Subject"
"2" "Activity"
"3" "tBodyAccmeanX"
"4" "tBodyAccmeanY"
"5" "tBodyAccmeanZ"
"6" "tGravityAccmeanX"
"7" "tGravityAccmeanY"
"8" "tGravityAccmeanZ"
"9" "tBodyAccJerkmeanX"
"10" "tBodyAccJerkmeanY"
"11" "tBodyAccJerkmeanZ"
"12" "tBodyGyromeanX"
"13" "tBodyGyromeanY"
"14" "tBodyGyromeanZ"
"15" "tBodyGyroJerkmeanX"
"16" "tBodyGyroJerkmeanY"
"17" "tBodyGyroJerkmeanZ"
"18" "tBodyAccMagmean"
"19" "tGravityAccMagmean"
"20" "tBodyAccJerkMagmean"
"21" "tBodyGyroMagmean"
"22" "tBodyGyroJerkMagmean"
"23" "fBodyAccmeanX"
"24" "fBodyAccmeanY"
"25" "fBodyAccmeanZ"
"26" "fBodyAccmeanFreqX"
"27" "fBodyAccmeanFreqY"
"28" "fBodyAccmeanFreqZ"
"29" "fBodyAccJerkmeanX"
"30" "fBodyAccJerkmeanY"
"31" "fBodyAccJerkmeanZ"
"32" "fBodyAccJerkmeanFreqX"
"33" "fBodyAccJerkmeanFreqY"
"34" "fBodyAccJerkmeanFreqZ"
"35" "fBodyGyromeanX"
```

```
"36" "fBodyGyromeanY"
"37" "fBodyGyromeanZ"
"38" "fBodyGyromeanFreqX"
"39" "fBodyGyromeanFreqY"
"40" "fBodyGyromeanFreqZ"
"41" "fBodyAccMagmean"
"42" "fBodyAccMagmeanFreq"
"43" "fBodyBodyAccJerkMagmean"
"44" "fBodyBodyAccJerkMagmeanFreq"
"45" "fBodyBodyGyroMagmean"
"46" "fBodyBodyGyroMagmeanFreq"
"47" "fBodyBodyGyroJerkMagmean"
"48" "fBodyBodyGyroJerkMagmeanFreq"
"49" "tBodyAccstdX"
"50" "tBodyAccstdY"
"51" "tBodyAccstdZ"
"52" "tGravityAccstdX"
"53" "tGravityAccstdY"
"54" "tGravityAccstdZ"
"55" "tBodyAccJerkstdX"
"56" "tBodyAccJerkstdY"
"57" "tBodyAccJerkstdZ"
"58" "tBodyGyrostdX"
"59" "tBodyGyrostdY"
"60" "tBodyGyrostdZ"
"61" "tBodyGyroJerkstdX"
"62" "tBodyGyroJerkstdY"
"63" "tBodyGyroJerkstdZ"
"64" "tBodyAccMagstd"
"65" "tGravityAccMagstd"
"66" "tBodyAccJerkMagstd"
"67" "tBodyGyroMagstd"
"68" "tBodyGyroJerkMagstd"
"69" "fBodyAccstdX"
"70" "fBodyAccstdY"
"71" "fBodyAccstdZ"
"72" "fBodyAccJerkstdX"
"73" "fBodyAccJerkstdY"
"74" "fBodyAccJerkstdZ"
"75" "fBodyGyrostdX"
"76" "fBodyGyrostdY"
"77" "fBodyGyrostdZ"
```

"78" "fBodyAccMagstd"
"79" "fBodyBodyAccJerkMagstd"
"80" "fBodyBodyGyroMagstd"
"81" "fBodyBodyGyroJerkMagstd"

Those variables with Acc in the name have units of standard gravity (g), and those with Gyro in the name have units of radians/second; however, the values have been normalized so that they range from -1 to 1. Normalizing the data removes the units. The output file, Samsung_data_tidy.txt meets the criteria for a tidy data set. Each row (observation) corresponds to a unique subject/activity combination. Each column (variable) is for a unique variable of the data set.

Summary and Explanation of Choices made in the project:

One could make the argument that all of the variables except Subject, Activity, tBodyAccmeanX, tBodyAccmeanY, tBodyAccmeanZ, tBodyGyromeanX, tBodyGyromeanY, and tBodymeanZ are calculated from these variables are should not be included in the tidy data set. However, I do not agree with that interpretation. First, all of frequency based variables were calculated by a Fast Fourier Transform of time based data. To do this calculation, we would need the accelerometer and gryosope data as a function of time with a clear interval for sampling the data. The features_info.txt file included with the data says that the data were captured at a constant rate of 50Hz. Based on this, one could use the data in the Inertial Signals directory to establish the time relationship for the Accelerometer and Gyroscope data, and use that to calculate the frequency domain variables (fBodyAcc and fBodyGyro). The README txt also states that the data is normalized so that it falls between 1 and -1. Normalizing the data in teh X_text and X_train data sets removes the absolute magnitude of the data, and makes it impossible to take that data and reliably transform it into another variable. For many of the other variables in the final tidy data set, the details of the calculation that the originators of the data set used were not provided. In the absence of these details, the values in the input data set could not be reproduced. One of the concepts behind creating a script to manipulate data into a tidy data set is reproducibility.
For these reasons, I chose to keep all of the variables (with "mean" or "std" in the descriptive name) in the final tidy data set.