Code Book

Raw data collection

Collection

Raw data are obtained from UCI Machine Learning repository. In particular we used the *Human Activity Recognition Using Smartphones Data Set*, that was used by the original collectors to conduct experiments exploiting Support Vector Machine (SVM)

Activity Recognition (AR) aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. The collectors used a sensor based approach employing smartphones as sensing tools. Smartphones are an effective solution for AR, because they come with embedded built-in sensors such as microphones, dual cameras, accelerometers, gyroscopes, etc.

The data set was built from experiments carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, 3-axial linear acceleration and 3-axial angular velocity were captured at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually

The obtained data set has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

Signals

The 3-axial time domain signals from accelerometer and gyroscope were captured at a constant rate of 50 Hz Then they were filtered to remove noise. Similarly, the acceleration signal was then separated into body and gravity acceleration signals using another filter. Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm Finally a Fast Fourier Transform (FFT) was applied to some of these time domain signals to obtain frequency domain signals.

The signals were sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window at 50 Hz). From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

The set of variables that were estimated from these signals are:

- 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(): Autoregression 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 some vectors.

No unit of measures is reported as all features were normalized and bounded within [-1,1].

List of Data Stores

Target Data Store	Sourced From
input_activity_labels	Provided txt file
input_features	Provided txt file
input_X_test	Provided txt file
input_y_test	Provided txt file
input_subject_test	Provided txt file
input_X_train	Provided txt file
input_Y_train	Provided txt file
input_subject_train	Provided txt file
tbl_subject_train	input_subject_train
tbl_y_train	input_Y_train
tbl_x_train	input_X_train
tbl_subject_test	input_subject_test

tbl_x_test	input_X_test
tbl_y_test	input_y_test
tbl_features	input_features
tbl_activity_labels	input_activity_labels
tbl_training_sensor	tbl_x_train
	bind_cols(tbl_subject_train
	bind_cols(tbl_y_train
tbl_testing_sensor	tbl_x_test
	bind_cols(tbl_subject_test%>%
	bind_cols(tbl_y_test
tbl_sensor_data	rbind(tbl_training_sensor, tbl_testing_sensor

# Provide column names to senor data table	
sensor_labels	tbl_features rbind(c(562, "Subject"%>% rbind(c(56
tbl_sensor_data	sensor_labels$V2
tbl_activity_labels	c(X1="Activity_Id", X2="Activity_Description"
tbl_sensor_data_mean_std	tbl_sensor_data[,grepl("mean|std|Subject|Activity_Id names(tbl_sensor_data]
tbl_sensor_data_mean_std	left_join(tbl_sensor_data_mean_std, tbl_activity_la "Activity_Id"
org_names	names(tbl_sensor_data_mean_std
names(tbl_sensor_data_mean_std)	Character Manipulation on names(tbl_sensor_data_mean
new_names	names(tbl_sensor_data_mean_std)
tbl_sensor_data_mean_std_2	remove from tbl_sensor_data_mean_std, colomn "Descri

tbl_sensor_data_mean_std_avg	tbl_sensor_data_mean_std_2 : group_by("subject","act calcluate the mean for 1:79

file="sensor_data_avg_grpd.txt"	Provide tidy result file thru tbl_sensor_data_mean_s

Data transformation

The raw data sets are processed with run_analisys.R script to create a tidy data set

Merge training and test sets

Test and training data (X_train.txt, X_test.txt), subject ids (subject_train.txt, subject_test.txt) and activity ids (y_train.txt, y_test.txt) are merged to obtain a single data set. Variables are labelled with the names assigned by original collectors (features.txt).

Extract mean and standard deviation variables

From the merged data set is extracted and intermediate data set with only the values of estimated mean (variables with labels that contain "mean") and standard deviation (variables with labels that contain "std").

Use descriptive activity names

A new column is added to intermediate data set with the activity description. Activity id column is used to look up descriptions in activity_labels.txt.

Label variables appropriately

Labels given from the original collectors were changed: *to obtain valid R names without parentheses, dashes and commas* to obtain more descriptive labels

Create a tidy data set

From the intermediate data set is created a final tidy data set where numeric variables are averaged for each activity and each subject.

The tidy data set contains 10299 observations with 81 variables divided in:

- an activity label (**Activity**): WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING
- an identifier of the subject who carried out the experiment (**Subject**): 1, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30
- a 79-feature vector with time and frequency domain signal variables (numeric)

The following table relates the 17 signals to the names used as prefix for the variables names present in the data set. ".XYZ" denotes three variables, one for each axis.

Name	Time domain	Frequency domain
Body Acceleration	TimeDomain.BodyAcceleration.XYZ	FrequencyDomain.BodyAcceleration.XYZ
Gravity Acceleration	TimeDomain.GravityAcceleration.XYZ	
Body Acceleration Jerk	TimeDomain.BodyAccelerationJerk.XYZ	FrequencyDomain.BodyAccelerationJerk.XYZ
Body Angular Speed	TimeDomain.BodyAngularSpeed.XYZ	FrequencyDomain.BodyAngularSpeed.XYZ
Body Angular Acceleration	TimeDomain.BodyAngularAcceleration.XYZ	
Body Acceleration Magnitude	TimeDomain.BodyAccelerationMagnitude	FrequencyDomain.BodyAccelerationMagnitude
Gravity Acceleration Magnitude	TimeDomain.GravityAccelerationMagnitude	
Body Acceleration Jerk Magnitude	TimeDomain.BodyAccelerationJerkMagnitude	FrequencyDomain.BodyAccelerationJerkMagnitude
Body Angular Speed Magnitude	TimeDomain.BodyAngularSpeedMagnitude	FrequencyDomain.BodyAngularSpeedMagnitude
Body Angular Acceleration	TimeDomain.BodyAngularAccelerationMagnitude	FrequencyDomain.BodyAngularAccelerationMagnitude

Name	Time domain	Frequency domain
Magnitude		

For variables derived from mean and standard deviation estimation, the previous labels are augmented with the terms "Mean" or "StandardDeviation".

The data set is written to the file sensor_data_avg_grpd.txt.