# Title: Human Activity Recognition Using Smartphones Data Set

**Introduction:**

The experiments have been performed with a group of 30 individuals within an age interval of 19-48 years. Each person wore a smartphone (Samsung Galaxy S II) on the waist, while he/she was performing 6 activities (walking, walking upstairs, waking downstairs, sitting, standing, laying).

Using embedded gyroscope and accelerometer linear accelerations, and angular velocity data have been collected in 3-axial directions at a constant rate of 50HZ. The experiments have been video recorded to label activities.

The purpose of this analysis is to identify and quantify associations between activities and recorded data by embedded accelerometer and gyroscope [1].

**Methods:**

*Data Collection*

For our analysis we used the data from machine learning repository data set [2]. The data were downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/samsungData.rda December 6 2013 using the R programming language [3].

*Exploratory Analysis*

Exploratory analysis was performed by examining tables and plots of the observed data. The name of variables (columns) have been corrected to be readable by R. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms used in the tree model for finding other important variables. The number of features in this study was 561. We applied tree classification method in R to find the most important features in data set.

We used n-fold cross validation (CV) method plus misclassification and deviance metrics to identify the best size for pruning the tree.

*Statistical Modeling*

To find the relation between the activities and important features we pruned the best tree with size 9, which was suggested by cross validation analysis from previous part [4,5].

*Reproducibility*

All analyses performed in this manuscript are reproduced in the R markdown file project_2.Rmd [6]. To reproduce the exact results presented in this manuscript the coursera version of the analysis must be used.

**Results:**

The data used in this analysis contains has 531 features that may related to the activity. We didn't identify any missing data in data set, since it was already cleared by the instructor of the course.

We used the data of subjects 27, 28, 29, and 30 as a test set and used the rest of data to the training set. The number of samples in training and test sets were 5867 and 1845 respectively.

We first applied a tree model to training data set and then we used with 10-fold CV method to measure our misclassification and deviance VS tree size. According to CV method the best size for pruning tree is 9, however this can change based on the size of training set. We pruned tree with sizes from 4 to 9. As Table 1 shows, the best result was for the size 9 with lowest "in sample" and "out sample" misclassifications.

**Table 1. Misclassification in training set and test set for different tree sizes**

| Tree Size | Training set misclassification | Test set misclassification |
|---|---|---|
| 4 | 2027 | 522 |
| 5 | 1209 | 351 |
| 6 | 636 | 205 |
| 7 | 598 | 203 |
| 8 | 596 | 200 |
| 9 | 549 | 164 |

Based on size 9 we pruned the best tree using the following 8 variables:
energy of a frequency interval measured by accelerometer in x direction in frequency domain, minimum raw signal of accelerometer for gravity in x direction, angle between y direction and average of gravity vector, total energy measured by accelerometer in frequency domain, auto regression of gravity acceleration in time domain, index of largest magnitude in z direction measured by gyroscope in frequency domain, correlation between x and y component measured in time domain, and finally gravity acceleration in time domain in y direction.

**Conclusions:**

Our analysis suggests that only 9 of those 531 features are enough for classification of subject activity.

As figure 1 suggests, only by knowing the "value of energy of a frequency interval measured by accelerometer in x direction in frequency domain" we can determine that subject is stationary (i.e., laying, sitting or standing) or moving.

On the left side of tree, by adding the "minimum raw signal of accelerometer for gravity in x direction" to the model, we can determine that subject is lying or not. Moreover having "angle between y direction and average of gravity vector" will help us to guess that subject is sitting or standing.

On the right side of tree, “total energy measured by accelerometer in frequency domain” indicate that subject is walking down or not. For distinguishing between walking up and walking we need the other 3 features. In this sense, walking up and walking are the hardest activities to be determined by the modeled tree.

**References**


1. Anguita, Davide, et al. "*Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine*." Ambient Assisted Living and Home Care. Springer Berlin Heidelberg, 2012. 216-223.

**2**. Center for Machine Learning and Intelligent Systems Page. URL: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Accessed 12/6/2013

3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org

3. Quick-R “Tree-Based Models” Page. URL: http://www.statmethods.net/advstats/cart.html. Accessed 12/6/2013.

5. Wikipedia “Decision tree” Page. URL: http://en.wikipedia.org/wiki/Decision_tree. Accessed 12/6/2013.

6. R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 1/31/2013