### Introduction

The purpose of this assignment is to provide a predictive model able to identify the correct action performed by a person from the data associated with his movements as provided by the sensors of a smart phone[[1]](#endnote-2).

The data have been collected and presented in a paper that has been wrote on the subject[[2]](#endnote-3). In the same paper the logic and the conventions linked to the variable names are extensively explained.

The vast majority of the method and the work flow used has heavily borrowed from the three Coursera[[3]](#endnote-4) courses the author attended this year: *Introduction to data science* (prof. Bill Howe – University of Washington)*, Computing for data analysis* (prof. Roger Peng – Johns Hopkins Bloomberg school of public health) and *Data analyisis* (prof. Jeff Leek – Johns Hopkins Bloomberg school of public health).

All mistakes are obviously mine.

### Methods

# Data collection and variables management

Data has been provided as an .rda file as part of the assignment rubric.

After loading the data from the .rda file I realized that few variable names are duplicated. The most likely cause is the missing axis for the measure. Using a function posted in the forum[[4]](#endnote-5) this issue was sorted. The other change being performed was to transforming the activity variable in a factor variable.

I didn't change any variable names; they are fairly long but the result is that it's clear what each measure is about.

The issue is fairly clear: how best to deal with a set of 561 relevant independent variables over 7352 observations?

# Training set/Test set

The prompt for the assigment also specify to large extent how to split the dataset between a training set and a test set.

Your task is to build a function that predicts what activity a subject is performing based on the quantitative measurements from the Samsung phone. For this analysis your training set must include the data from subjects 1, 3, 5, and 6. But you may use more subjects data to train if you wish. Your test set is the data from subjects 27, 28, 29, and 30, but you may use more data to test. Be careful that your training/test sets do not overlap.

I preferred to have a training set and a test set. I decided not to use a validation set.

In order to create the two sets I followed the instructions and did the following:

fixTrain <- c(1, 3, 5, 6)

fixTest <- c(27, 28, 29, 30)

otherSubj <- unique(samsungData[! samsungData$subject %in%

c(fixTrain, fixTest), "subject"])

set.seed(1234)

index <- 1:length(otherSubj)

trainindex <- sample(index, ceiling(length(index)/2))

otherTrain <- otherSubj[trainindex]

otherTest <- otherSubj[-trainindex]

subjTrain <- c(fixTrain, otherTrain)

subjTest <- c(fixTest, otherTest)

samTrain <- na.omit(samsungData[samsungData$subject %in% subjTrain, ])

samTest <- na.omit(samsungData[samsungData$subject %in% subjTest, ])

samTrain (3779 observations) and samTest (3573) frames are the two sets of data I worked with.

# Exploratory analysis

A quick run on the sample is done in order to assess if the there anything wrong with the activity variable:

laying sitting standing walk walkdown walkup <NA>

696 631 705 668 522 557 0

The sample appears fairly balanced.

The first option is to chart all the variables and have a quick look as suggested by someone[[5]](#endnote-6). In my opinion going through 561 sets of charts is hardly something exciting and, more crucially, useful.

How can you visually compare all the charts (an example is Illustration) and determine which one to pick?

What can be derived from those charts is that you can basically split the variables set between those that are good to identify *laying, sitting and standing* and those that are good for *walk, walkdown, walkup.*

In my opinion a better guidance is provided by statistics performed on each of them.

A table for the mean for each variable and for each activity can be computed and delivers interesting information: to start with you have 21 instances of statistics that appear to be duplicated. This gives you the opportunity to get of 21 variables (using them would lead to singularity errors). This is an example for the first few of them

laying sitting standing walk walkdown walkup

tBodyAcc.mean.X 0.26285168 0.27368204 0.27988742 0.27642252 0.28687697 0.258900900

tBodyAcc.mean.Y -0.01963588 -0.01140817 -0.01549443 -0.01821543 -0.01657714 -0.027137098

tBodyAcc.mean.Z -0.10765206 -0.10627132 -0.10694535 -0.11115261 -0.10722971 -0.121808766

tBodyAcc.std.X -0.95260705 -0.98233925 -0.98494183 -0.31959341 0.12359577 -0.242019425

tBodyAcc.std.Y -0.91656801 -0.92958808 -0.93913215 0.01151121 0.06322413 0.004372214

tBodyAcc.std.Z -0.92457633 -0.92886694 -0.93868548 -0.22980581 -0.15316738 -0.174602885

What you can also do is to calculate the standard deviation of the value of each row of this variable statistics. The higher that standard deviation the more likely the variable might split the activities.

fBodyAccJerk.entropy.X fBodyAccJerk.entropy.Y

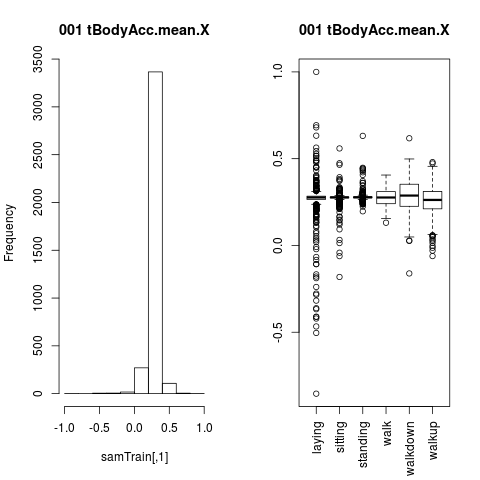
0.8121700 0.7879797

tBodyAccJerkMag.entropy fBodyAcc.entropy.X

0.7722359 0.7706269

fBodyBodyAccJerkMag.entropy fBodyAccMag.entropy

0.7208513 0.7152758

Illustration 1: Charting variable characteristics

# Statistical modelling

Unfortunately I didn't have enough time to explore the Tree method and I preferred to stick to OLS (the *lm* function) estimates which is something I am comfortable with. In order to use that I need to treat the *activity* variable as a number and then return (via rounding to the nearest integer) to the discrete numbers associated with the factors and display the activity name at the end.

What it is done as well is that when an predicted value is less than 0 it's then coerced to 1; if it is greater than 6 is coerced to 6. This was triggered in very few occasions anyway.

What appeared clear is that picking 150 variables at random and then running a stepwise regression is providing very good results. You identify correctly

### Results

### Conclusions

1. A Samsung Galaxy SII has been used: this is actually a fairly old phone (released on April 2011) by now. Most recent models (we are close to the S5 some time early next year) should provide an even better quality for those measures. [↑](#endnote-ref-2)
2. 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. Or here: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones> [↑](#endnote-ref-3)
3. http://www.coursera.com [↑](#endnote-ref-4)
4. Uwe F Mayer post on thread https://class.coursera.org/dataanalysis-002/forum/thread?thread\_id=1237 "How to deal with duplicate column names in samsungData" [↑](#endnote-ref-5)
5. Again Uwe F Mayer post on thread <https://class.coursera.org/dataanalysis-002/forum/thread?thread_id=1198> "Assignment 2: some pointers on getting started” is providing with an excellent procedure to get all the charts saved for us. [↑](#endnote-ref-6)