Human Activity Recognition (HAR)

Jose Ramon Hernandez Galan

18/11/2019

Introduction

This is a report showing the process and results for the creation of a model for human activity recognition (herafter HAR).

This is the capstone project of the Data Science course pursued by the author in HarvardX.

Dataset

The dataset used in this project is the **HAR Dataset for benchmarking** 1

The dataset includes measurements of **inertial sensors** attached to several person while doing normal activities during the day. It also includes data related to the person such as weight, height, etc.

A more detailed description of the dataset can be found **here**

The main goal of this project is to use machine learning techniques in order to predict the human activity. We will compare our results

We will also observe if all 4 sensors are really necessary or if we can use less sensors in order to predict the activity.

Analysis

In this section we will prepare the data to work with and explore some important characteristics of the dataset.

Data wrangling

The created har dataset has the following structure:

Name	Type	Description
user	Factor	w/ 4 levels “debora”,“jose_carlos”,..: 1 1 1 1 1 1 1 1 1 1 ...
gender	Factor	w/ 2 levels “Man”,“Woman”: 2 2 2 2 2 2 2 2 2 2 ...
age	int	46 46 46 46 46 46 46 46 46 46 ...
how_tall_in_meters	num	1.62 1.62 1.62 1.62 1.62 1.62 1.62 1.62 1.62 1.62 ...
weight	int	75 75 75 75 75 75 75 75 75 75 ...
body_mass_index	num	28.6 28.6 28.6 28.6 28.6 28.6 28.6 28.6 28.6 28.6 ...
x1	int	-3 -3 -1 -2 -1 -2 1 -1 -1 0 ...
y1	int	92 94 97 96 96 95 100 97 98 98 ...
z1	int	-63 -64 -61 -57 -61 -62 -62 -63 -63 -61 ...
x2	int	-23 -21 -12 -15 -13 -14 -10 -13 -14 -11 ...
y2	int	18 18 20 21 20 19 22 20 19 22 ...
z2	int	-19 -18 -15 -16 -15 -16 -12 -15 -17 -13 ...
x3	int	5 -14 -13 -13 -13 -13 -13 -12 -13 -13 ...
y3	int	104 104 104 104 104 104 104 104 104 104 ...

Name	Type	Description
z3	int	-92 -90 -90 -89 -89 -89 -90 -88 -90 -90 ...
x4	int	-150 -149 -151 -153 -153 -153 -151 -151 -152 -151 ...
y4	int	-103 -104 -104 -103 -104 -104 -104 -104 -103 -104 ...
z4	int	49 47 45 43 44 43 44 43 45 45 ...
class	Factor	w/ 5 levels "sitting","sittingdown",..: 1 1 1 1 1 1 1 1 1 1 ...

See the different activites listed in the dataset:

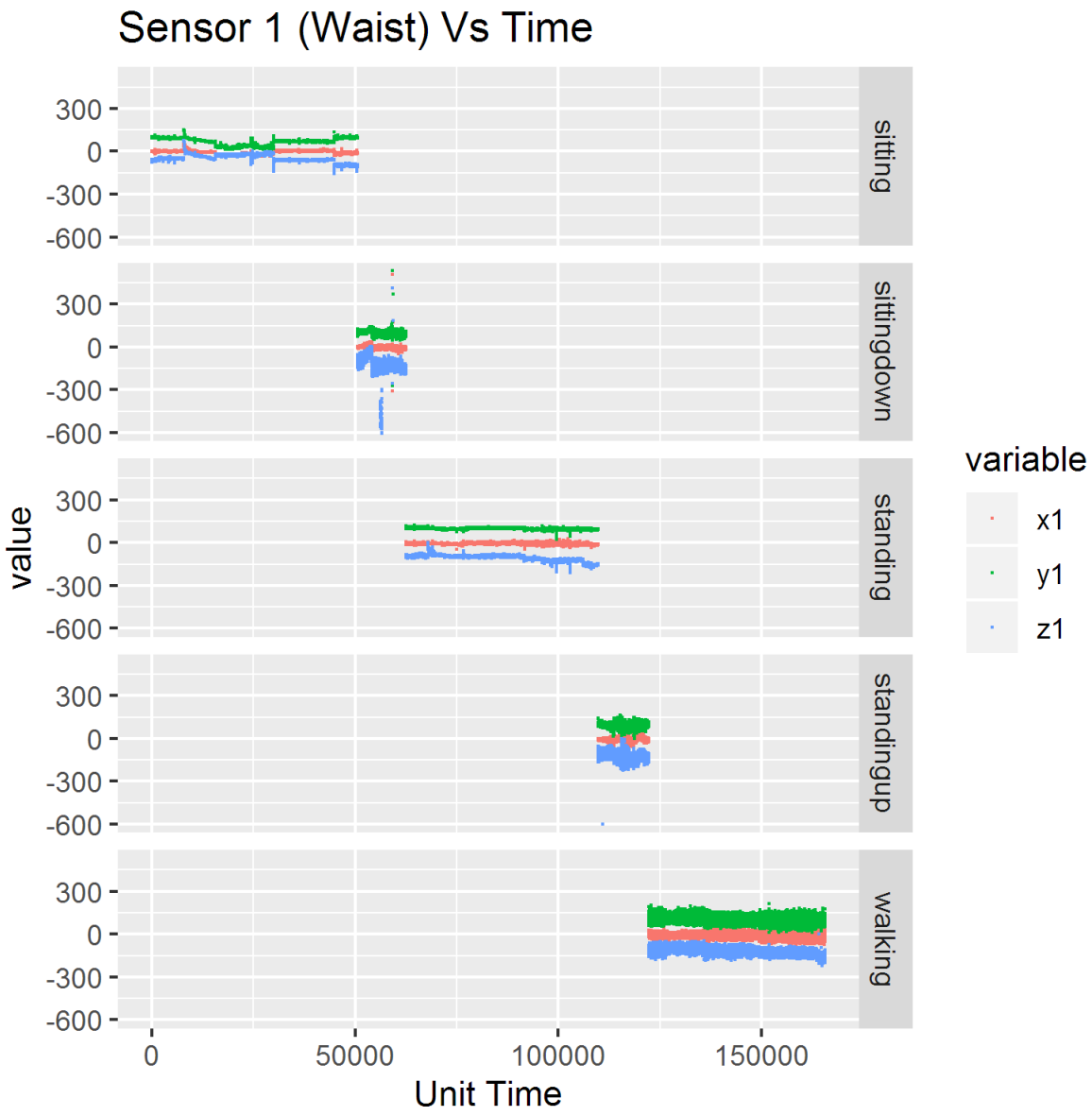
```
levels(har$class)
```

```
## [1] "sitting"     "sittingdown" "standing"    "standingup"  "walking"
```

These classes are associated to the values obtained by the sensors (1 to 4) expressed in m/s2. X, Y, Z are the values obtained for each axis.

Exploratory Data Analysis

In order to understand how th values were taken we will view the sampled data **across the time for sensor 1 values (x,y,z)**. The data was taken during 8 hours. During this time the sensed subject was doing different activites.



This is the sensor located at the waist of the subjects.

Note the **unit time** is not specified. We have included a non-scaled time unit for creating the chart.

Some additional **checkings** have been done in order to verify the consistency of data.

```
# Is there any NAs ?
nas <- apply(har, MARGIN = 2, function(x) any(is.na(x) | is.infinite(x)))
if (any(nas) == FALSE)
{
  cat("Great, no NAs or Infinite value in the dataset")
}else{
  cat("Attention, some NA or Infinite value was found in the dataset")
}
```

```
## Great, no NAs or Infinite value in the dataset
```

```
## # A tibble: 5 x 3
##   class           n        p
##   <fct>       <int>    <dbl>
## 1 sitting     50631 0.30568
## 2 standing    47370 0.28599
## 3 walking     43390 0.26196
## 4 standingup  12415 0.074955
## 5 sittingdown 11827 0.071405
```

```
## # A tibble: 20 x 4
## # Groups:   class [5]
##    class       user            n         p
##    <fct>       <fct>       <int>     <dbl>
##  1 sitting     debora      15615 0.094275
##  2 sitting     wallace     14993 0.090519
##  3 standing    debora      14940 0.090199
##  4 standing    wallace     14467 0.087344
##  5 sitting     katia       14280 0.086215
##  6 standing    katia       14234 0.085937
##  7 walking     wallace     14037 0.084748
##  8 walking     debora      13622 0.082242
##  9 walking     katia       13556 0.081844
## 10 sitting     jose_carlos  5743 0.034673
## 11 standingup  wallace      4115 0.024844
## 12 sittingdown katia        4017 0.024252
## 13 standingup  debora       3853 0.023262
## 14 standing    jose_carlos  3729 0.022514
## 15 standingup  katia        3710 0.022399
## 16 sittingdown debora       3547 0.021415
## 17 sittingdown wallace      3486 0.021047
## 18 walking     jose_carlos  2175 0.013131
## 19 sittingdown jose_carlos   777 0.0046911
## 20 standingup  jose_carlos   737 0.0044496
```

Now, we are sure there are NA values in our dataset. Also we are sure all users (4) have been measured in all the different activities (5). Note the prevalance in some user/activity. This also can be observed in the time-based chart above.

Tranining set partitioning

We have splitted the har original dataset in several sets for training, tuning and validatiaon.

Dataset	Observations	Proportions	Description
har	165.633		Original
har_val	16.565		Set for final valiation
har_set	149.068		Set for model optimization
har_set_train	134.158		Set for model training
har_set_test	16.565		Set for model tuning

Analysis approach

Model based on Classification Trees

Decission Tress

Random Forest

Rborist

Final Results

Conclusions

References

1. Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. DOI: 10.1007/978-3-642-34459-6_6

Read more: http://groupware.les.inf.puc-rio.br/har#sbia_paper_section#ixzz65cgnrXLU