Codebook

SabaDeMa

1 Introduction

This file explains the code of the script for the assignment. It is organized by steps, labelled with "Step" and written in the script as comments. Before you start of course you have to unzip the folder with the data in the working directory in a a folder called "UCI HAR DATASET".

If not the case, and you prefer to unzip directly in the working directory, you have to delete from read.table commands the "UCI HAR DATASET/" in order to make the code work well.

I preferred to keep variables of every step, this is just a personal choice but if users might want to delete objects or overwrite them I think that my code can be modified very easily.

2 Details of the steps

The first command load dplyr package and (I think) no explanations are needed.

Step1: read all files in R by using the read.table function.

Step2: convert all objects to tbl_df objects in order to use them in an easier way and allow functions like group_by to work properly.

Step3: select only the second column (the name, which is what we need) of feature to use the result as names for the columns of dataTrain and dataTest, but before of that it must be converted into a vector and transposed with the t() function. Create names for the subjects and activities columns respectively "Sub" and "Act".

Step4: create a data frame for test and train with the data, the sub and the activities columns. Merge data frames together (data_train_fin and data_test_fin) to have a total data frame with all the sets names data_total.

Step5: select with the grepl function all columns that have the words mean and std inside their names.

Step6: in order to have a more readable and shorter columns names this step does some abbreviations and replacements. It delete tBody, fBody and parenthesis (by escaping them with a double \) from columns names and gives “GY” instead of “Gyro” and “AC” instead of “Acc”.

Step7: this dataset has some issues due to some identical columns names, this can be verified with the following commands. These commands show that the problems are not due to modifications in the script but they were there when the original dataset was created.

```
> similarcol <- anyDuplicated(data_total, MARGIN = 2)
> similarcol
[1] 0
```

Columns have same names but with the previous commands we know that they are not the same things so I think that is better to keep them but in order to avoid problems, change name to the duplicates. That is done with this command:

```
por <- make.names(nomi, unique = TRUE)
```

Step8: change levels from numbers 1 to 6 to activities names (Walking, Walking_Up, etc.)

Step9: group the dataset by using the group_by function.

Step10: create a dataset named data_4 with the mean for each variable grouped for each subject and activity.

Step11: write data_4 dataset to an external file located in the woking directory named data.txt.

3 Output file

The output file is data_4. It has 14580 observations in total, 180 rows and 81 columns. Values of first column named Act, which describes the activities done by subjects are: Walking, Walking_Up, Walking_down, Sitting, Standing, Lying.

Values of the second column named Sub, which represents the subjects are from 1 to 30, each number is a subjects.

Names of all columns are:

```
> names(data_4)
  [1] "Act"                   "Sub"
  [3] "ACmeanX"               "ACmeanY"
  [5] "ACmeanZ"               "ACstdX"
  [7] "ACstdY"                "ACstdZ"
  [9] "tGravityACmeanX"       "tGravityACmeanY"
 [11] "tGravityACmeanZ"       "tGravityACstdX"
 [13] "tGravityACstdY"        "tGravityACstdZ"
 [15] "ACJerkmeanX"           "ACJerkmeanY"
 [17] "ACJerkmeanZ"           "ACJerkstdX"
 [19] "ACJerkstdY"            "ACJerkstdZ"
```

```
[21] "GYmeanX"                "GYmeanY"
[23] "GYmeanZ"                "GYstdX"
[25] "GYstdY"                 "GYstdZ"
[27] "GYJerkmeanX"            "GYJerkmeanY"
[29] "GYJerkmeanZ"            "GYJerkstdX"
[31] "GYJerkstdY"             "GYJerkstdZ"
[33] "ACMagmean"              "ACMagstd"
[35] "tGravityACMagmean"      "tGravityACMagstd"
[37] "ACJerkMagmean"          "ACJerkMagstd"
[39] "GYMagmean"              "GYMagstd"
[41] "GYJerkMagmean"          "GYJerkMagstd"
[43] "ACmeanX.1"              "ACmeanY.1"
[45] "ACmeanZ.1"              "ACstdX.1"
[47] "ACstdY.1"               "ACstdZ.1"
[49] "ACmeanFreqX"            "ACmeanFreqY"
[51] "ACmeanFreqZ"            "ACJerkmeanX.1"
[53] "ACJerkmeanY.1"          "ACJerkmeanZ.1"
[55] "ACJerkstdX.1"           "ACJerkstdY.1"
[57] "ACJerkstdZ.1"           "ACJerkmeanFreqX"
[59] "ACJerkmeanFreqY"        "ACJerkmeanFreqZ"
[61] "GYmeanX.1"              "GYmeanY.1"
[63] "GYmeanZ.1"              "GYstdX.1"
[65] "GYstdY.1"               "GYstdZ.1"
[67] "GYmeanFreqX"            "GYmeanFreqY"
[69] "GYmeanFreqZ"            "ACMagmean.1"
[71] "ACMagstd.1"             "ACMagmeanFreq"
[73] "BodyACJerkMagmean"      "BodyACJerkMagstd"
[75] "BodyACJerkMagmeanFreq"  "BodyGYMagmean"
[77] "BodyGYMagstd"           "BodyGYMagmeanFreq"
[79] "BodyGYJerkMagmean"      "BodyGYJerkMagstd"
[81] "BodyGYJerkMagmeanFreq"
```