Untitled

Paolo Colussi

2024-11-08

This is the first attempt of make the gain algorithm work in R.

```
knitr::opts_chunk$set(echo = TRUE)

rm(list = ls())

library(reticulate)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

First we have some utilitis function we need to define.

```
# This function generate a mask for introducing the missing values
missig <- function(p, no, dim) {
  mm <- matrix(runif(no*dim),no,dim)
  mask <- ifelse((mm < p),1,NA)
  return(mask)
}


# This function has a matrix as imput and generate a df with the minimum and the maximimum of each colu
# we are going to use it to normalize the data
parameters_norm <- function(data){
  min <- apply(data, FUN = min, 2)
  max <- apply(data, FUN = max,2)

  parameters <- cbind(min,max)
  return(parameters)
}
```

```r
# this function normalize the data
normalization <- function(data, parameters){


  data_norm <- sweep(data, 2, parameters[,1], "-")
  data_norm <- sweep(data_norm, 2, parameters[,2], "/")

}

# this function calculate the RMSE
rmse_loss <- function(ori_data, imputed_data, data_m){

  parameters <- parameters_norm(ori_data)
  ori_data <- normalization(ori_data,parameters)
  imputed_data <- normalization(imputed_data,parameters)


  # Only for missing values
  nominator <- sum(((1-data_m) * ori_data - (1-data_m) * imputed_data)**2)
  denominator <- sum(1-data_m)

  rmse <- sqrt(nominator/denominator)

  return(rmse)
}
```

Now, we have the main function, that prepares the data and call the function gain.

The parameters that are suggested in the GitHub are set to default.

IT'S NOT WORKING!!!

```r
main_gain <- function(data, miss_rate = 0.2, batch_size = 128, hint_rate = 0.9, alpha = 100, iterations
  no <- length(data)
  dim <- ncol(data)
  mask <- missig(1-miss_rate, no, dim) # create the mask matrix
  data_missing <- mask*data # create the matrix with the missing values

  # put the parameters in to a list
  gain_parameters <- list(batch_size = batch_size,
    hint_rate= hint_rate,
    alpha = alpha,
    iterations = iterations)

  # save the functions from the Python code, this is obviusly the path in my computer (paolo),
  # but I think there is a way to call the document directly from the github, but i didn't try
  # to do it

  source_python("/home/paoloc/Documenti/Utrecht/uni/varie/paper/lavoro fine 2024/code/GAIN_orig/utils.py
  source_python("/home/paoloc/Documenti/Utrecht/uni/varie/paper/lavoro fine 2024/code/GAIN_orig/gain.py

  #  apply the gain algorithm
```

```
  imputed <- gain(data_missing, gain_parameters)

  # compute the RMSE
  rmse_loss(data, imputed, mask)
}
```