07.- Data Sampling_CU_34_05_servicios_completo_v_01

June 16, 2023

#

CUxx_Nombre del caso de uso

Citizenlab Data Science Methodology > II - Data Processing Domain *** > # 07.- Data Sampling

Data Sampling is the process to obtain a subset with the appropriate size of data (examples/rows and features/columns) for analysis, maintaining the same statistical significance as raw dataset.

0.1 Tasks

Sampling dataset - Random Sampling Without Replacement - Random Sampling With Replacement - Systematic Sampling - Cluster Sampling

Evaluate Subsets - Evaluating by Mean - By fit

0.2 Consideraciones casos CitizenLab programados en R

- Puede que algunas de las tareas de este proceso se realicen en los notebooks de los procesos de deploy al estar relacionadas con otras de esos procesos. En esos casos, en este notebook se referencia al notebook del proceso correspondiente
- Puede que el proceso no aplique a los ficheros del caso de uso, y se indique "No aplica" de forma generalizada.
- Si en el nombre de archivo del notebook no aparece ningún sufijo, el notebook se refiere al caso globalmente

0.3 File

- Input File: No aplica
- Output File: No aplica

0.4 Settings

0.4.1 Encoding

Con la siguiente expresión se evitan problemas con el encoding al ejecutar el notebook. Es posible que deba ser eliminada o adaptada a la máquina en la que se ejecute el código.

[1]:
```
Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8")
```

'LC_CTYPE=es_ES.UTF-8;LC_NUMERIC=C;LC_TIME=es_ES.UTF-8;LC_COLLATE=es_ES.UTF-8;LC_MONETARY=es_ES.UTF-8;LC_MESSAGES=en_US.UTF-

8;LC_PAPER=es_ES.UTF-8;LC_NAME=C;LC_ADDRESS=C;LC_TELEPHONE=C;LC_MEASUREMENT
8;LC_IDENTIFICATION=C'

0.4.2 Libraries to use

[2]:
```
library(readr)
library(dplyr)
library(sf)
library(tidyr)
#library(stringr)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Linking to GEOS 3.11.1, GDAL 3.6.2, PROJ 6.2.1; sf_use_s2() is TRUE

WARNING: different compile-time and runtime versions for GEOS found:

Linked against: 3.11.1-CAPI-1.17.1 compiled against: 3.8.0-CAPI-1.13.1

It is probably a good idea to reinstall sf, and maybe rgeos and rgdal too
```

0.4.3 Paths

[3]:
```
iPath <- "Data/Input/"
oPath <- "Data/Output/"
```

0.5 Data Load

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if using this option

[4]:
```
# file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[5]:
```
iFile <- "CU_34_06_06_servicios_completo.csv"
file_data <- paste0(iPath, iFile)

if(file.exists(file_data)){
    cat("Se leerán datos del archivo: ", file_data)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/CU_34_06_06_servicios_completo.csv
```

Data file to dataframe Usar la función adecuada según el formato de entrada (xlsx, csv, json, ...)

[6]:
```
data <- read_csv(file_data)
```

```
Rows: 274792 Columns: 33
  Column specification
```

```
Delimiter: ","
chr   (5): Servicio, CMUN, CDIS, CSEC, NSEC
dbl  (27): Futbol, nservicios, capacidad, tmed, prec, velmedia,
presMax, cca…
date  (1): Fecha

 Use `spec()` to retrieve the full column specification for this
data.
 Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

Visualizo los datos.

Estructura de los datos:

[7]:
```
data |> glimpse()
```

```
Rows: 274,792
Columns: 33
$ Fecha            <date> 2022-01-01, 2022-01-01, 2022-01-01,
2022-01-01, 2022…
$ Servicio         <chr> "Taxi", "Taxi", "Taxi", "Taxi",
"Taxi", "Taxi", "Taxi…
$ CMUN             <chr> "001", "002", "002", "003", "004",
"004", "004", "004…
$ CDIS             <chr> "01", "01", "01", "01", "01", "01",
```

```
"01", "01", "01",…
$ CSEC             <chr> "001", "001", "002", "001", "001",
"002", "003", "004…
$ Futbol           <dbl> 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1,…
$ nservicios       <dbl> 0, 0, 0, 50, 0, 50, 50, 0, 2, 0, 0,
7, 0, 1, 4, 7, 10…
$ capacidad        <dbl> 50, 50, 50, 50, 50, 50, 50, 50, 65,
65, 65, 65, 65, 6…
$ tmed             <dbl> 9.472965, 8.170448, 8.526402,
10.012282, 10.574368, 1…
$ prec             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0,…
$ velmedia         <dbl> 0.8229741, 0.9750191, 0.8874318,
2.3162929, 0.6970652…
$ presMax          <dbl> 901.6717, 970.1249, 967.2948,
849.5611, 964.2144, 965…
$ ccaa             <dbl> 13, 13, 13, 13, 13, 13, 13, 13, 13,
13, 13, 13, 13, 1…
$ CPRO             <dbl> 28, 28, 28, 28, 28, 28, 28, 28, 28,
28, 28, 28, 28, 2…
$ t1_1             <dbl> 55, 2358, 2435, 248, 2886, 3376,
2161, 1523, 1648, 10…
$ t2_1             <dbl> NA, 0.4724, 0.4830, 0.3952, 0.5135,
0.4793, 0.5016, 0…
$ t2_2             <dbl> NA, 0.5276, 0.5170, 0.6048, 0.4865,
0.5207, 0.4984, 0…
$ t3_1             <dbl> NA, 40.5161, 38.3339, 47.4597,
44.4099, 40.6810, 38.0…
$ t4_1             <dbl> NA, 0.1425, 0.1959, 0.1371, 0.1611,
0.1739, 0.2230, 0…
$ t4_2             <dbl> NA, 0.7443, 0.7002, 0.6573, 0.6067,
0.6727, 0.6395, 0…
$ t4_3             <dbl> NA, 0.1132, 0.1039, 0.2056, 0.2322,
0.1534, 0.1374, 0…
$ t5_1             <dbl> NA, 0.1569, 0.1544, 0.1129, 0.1639,
0.1440, 0.1749, 0…
$ t6_1             <dbl> NA, 0.1968, 0.1951, 0.1411, 0.2128,
0.1730, 0.2138, 0…
$ t7_1             <dbl> NA, 0.4016, 0.4244, 0.1542, 0.1470,
0.1596, 0.1668, 0…
$ t8_1             <dbl> NA, 0.3971, 0.4224, 0.1449, 0.1409,
0.1506, 0.1602, 0…
$ t9_1             <dbl> NA, 0.4377, 0.4438, 0.5280, 0.2602,
0.3234, 0.2859, 0…
$ t10_1            <dbl> NA, 0.0912, 0.1226, 0.0932, 0.1814,
0.1478, 0.1658, 0…
$ t11_1            <dbl> NA, 0.6063, 0.5955, 0.5000, 0.4213,
```

```
0.5170, 0.4943, 0…
$ t12_1            <dbl> NA, 0.6672, 0.6788, 0.5514, 0.5147,
0.6067, 0.5926, 0…
$ NSEC             <chr> "Acebeda, La - 01.001", "Ajalvir -
01.001", "Ajalvir …
$ area             <dbl> 21848790.60, 15585050.39,
4148273.41, 25674490.85, 50…
$ elevation        <dbl> 1401, 653, 715, 1150, 593, 604, 587,
570, 594, 594, 5…
$ densidad_hab_km2 <dbl> 2.517302, 151.298837, 586.991204,
9.659393, 571.86106…
```

Muestra de los primeros datos:

[8]:
```
data |> slice_head(n = 5)
```

A spec_tbl_df: 5 × 33

Fecha <date>	Servicio <chr>	CMUN <chr>	CDIS <chr>	CSEC <chr>	Futbol <dbl>	nservicios <dbl>	capacidad <dbl>	tme <db
2022-01-01	Taxi	001	01	001	0	0	50	9.47
2022-01-01	Taxi	002	01	001	0	0	50	8.17
2022-01-01	Taxi	002	01	002	0	0	50	8.52
2022-01-01	Taxi	003	01	001	1	50	50	10.0
2022-01-01	Taxi	004	01	001	0	0	50	10.5

0.6 Sampling dataset

0.6.1 Random Sampling Without Repacement

No aplica

Data Rate to sampling

[9]:
```
# rate <- 0.50
```

Operation

[10]:
```
# set.seed(1)
# sdata <- slice_sample(data, prop = rate, replace = FALSE)
```

0.6.2 Random Sampling With Replacement

Data Rate to sampling

[11]:
```
# Data Rate to sampling
# rate <- 0.50
```

Operation

[12]:
```
# set.seed(1)
# sdata <- slice_sample(data, prop = rate, replace = FALSE)
```

0.6.3 Systematic Sampling

No aplica

Fixed sampling interval

[13]:
```
# interv <- 3
```

Operation

[14]:
```
# Código si es neceario
```

0.6.4 Cluster Sampling

No aplica

Number of clusters

[15]:
```
# nc <- 2
```

Operation

[16]:
```
# Código si es neceario
```

0.7 Evaluating Subsets Generated

No aplica

0.7.1 Evaluating by Mean

0.7.2 Evaluating by Fit: Classification

Random Sampling Without Repacement

Random Sampling With Repacement

Systematic Sampling

0.7.3 Evaluating by Fit: Regression

Random Sampling Without Repacement

Random Sampling With Repacement

Systematic Sampling

Cluster Sampling

0.8 Data Save

- Solo si se han hecho cambios
- No aplica

Identificamos los datos a guardar

[17]:
```
data_to_save <- data
```

Estructura de nombre de archivos:

- Código del caso de uso, por ejemplo “CU_04”
- Número del proceso que lo genera, por ejemplo ”_06”.
- Resto del nombre del archivo de entrada
- Extensión del archivo

Ejemplo: ”CU_04_06_01_01_zonasgeo.json, primer fichero que se genera en la tarea 01 del proceso 05 (Data Collection) para el caso de uso 04 (vacunas) y que se ha transformado en el proceso 06

Importante mantener los guiones bajos antes de proceso, tarea, archivo y nombre

0.8.1 Proceso 07

[18]:
```
caso <- "CU_34"
proceso <- '_07'
tarea <- "_05"
archivo <- ""
proper <- "_servicios_completo"
extension <- ".csv"
```

OPCION A: Uso del paquete “tcltk” para mayor comodidad

- Buscar carpeta, escribir nombre de archivo SIN extensión (se especifica en el código)
- Especificar sufijo2 si es necesario
- Cambiar datos por datos_xx si es necesario

[19]:
```
# file_save <- paste0(caso, proceso, tarea, tcltk::tkgetSaveFile(), proper,␣
 ↪extension)
# path_out <- paste0(oPath, file_save)
# write_csv(data_to_save_xxxx, path_out)

# cat('File saved as: ')
# path_out
```

OPCION B: Especificar el nombre de archivo

- Los ficheros de salida del proceso van siempre a Data/Output/.

[20]:
```
file_save <- paste0(caso, proceso, tarea, archivo, proper, extension)
path_out <- paste0(oPath, file_save)
write_csv(data_to_save, path_out)

cat('File saved as: ')
path_out
```

```
File saved as:
```

'Data/Output/CU_34_07_05_servicios_completo.csv'

Copia del fichero a Input Si el archivo se va a usar en otros notebooks, copiar a la carpeta Input

[21]:
```
path_in <- paste0(iPath, file_save)
file.copy(path_out, path_in, overwrite = TRUE)
```

TRUE

0.9 REPORT

A continuación se realizará un informe de las acciones realizadas

0.10 Main Actions Carried Out

- No aplica el proceso al caso

0.11 Main Conclusions

- En este caso no se realiza muestreo

0.12 CODE TO DEPLOY (PILOT)

A continuación se incluirá el código que deba ser llevado a despliegue para producción, dado que se entiende efectúa operaciones necesarias sobre los datos en la ejecución del prototipo

Description

- No hay nada que desplegar en el piloto, ya que estos datos son estáticos o en todo caso cambian con muy poca frecuencia, altamente improbable durante el proyecto.

CODE

[22]:
```
# incluir código
```