Data in Public and Social Services - 2nd practical exercise class

8th April 2024
by Davide Chicco
davide.chicco@unimib.it

Practical exercise class lecturer: Vasco Coelho v.coelho@campus.unimib.it

The goals of this practical exercise class are the following:

A. Apply the concepts on exploratory data analysis (EDA) seen during the last class lecture on an EHRs dataset:
   a. Quantitative description
   b. Statistical correlations
   c. Dimensionality reduction

R setup main commands

0) Put the following commands as header of your script:

```
setwd(".") #we use the current folder of the script as working directory
options(stringsAsFactors = FALSE) # we set the input strings not to be considered
set.seed(10) # we set a seed to be able to replicate our tests
options(repos = list(CRAN="http://cran.rstudio.com/")) # we set the URL where to download the packages
```

1) load the pacman library for an easier installation and loading of the libraries:

```
# install.packages("pacman", dependencies = TRUE)
library("pacman")
```

# Load and/or install other packages we need

```
p_load("dlookr", "dplyr", "ggplot2",  "pastecs", "tableone", "umap",
"textshape")
```

Application of the main exploratory data analysis (EDA) steps to a dataset of electronic health records of patients with diabetes type 1 from Japan

2) Apply the concepts on data cleaning and data preparation seen during the first three class lectures on a real EHRs dataset:.

2.1) Takashi 2019 diabetes type 1 dataset: we download the preprocessed version, that is the output of the first practical exercise class

Takashi2019_diabetes_type1_dataset_preprocessed.csv file to download

> Takashi Y, Ishizu M, Mori H, Miyashita K, Sakamoto F, Katakami N, et al. (2019) "Circulating osteocalcin as a bone-derived hormone is inversely correlated with body fat in patients with type 1 diabetes". PLOS ONE 14(5): e0216416. https://doi.org/10.1371/journal.pone.0216416
>
> Cerono G, Chicco D. 2024, "Ensemble machine learning reveals key features for diabetes duration from electronic health records". PeerJ Computer Science 10:e1896 https://doi.org/10.7717/peerj-cs.1896

A. Load the dataset

```
fileName <- "Takashi2019_diabetes_type1_dataset_preprocessed.csv"
patients_data <- read.csv(fileName, header = TRUE, sep =",")
```

B. Quantitative description

```
# We want to generate the descriptive statistics of all the features involved

patients_data %>% dim()
patients_data %>% summary()
patients_data %>% str()
patients_data %>% colnames() %>% sort()

patients_data %>% pastecs::stat.desc()

tableone::CreateTableOne(data=patients_data)

summary(tableone::CreateTableOne(data = patients_data))

# What are the outputs of stat.desc() and summary() telling us?
```

C. Histograms

```
# Let's see the histogram of age, by changing some parameters
```

```
ggplot(patients_data, aes(x=age)) + geom_histogram()

ggplot(patients_data, aes(x=age)) + geom_histogram(binwidth=10)

age_histogram <- ggplot(patients_data, aes(x=age)) +
geom_histogram(binwidth=1, fill="blue", color="black")

histogram_file_width <- 10
histogram_file_height <- 5
output_age_histogram_file_name <- "age_histogram.pdf"
ggsave(age_histogram, width=histogram_file_width,
height=histogram_file_height, file=output_age_histogram_file_name)
```

# Generate the histograms for body mass index, eGFR, ucOC and save them into files

# Interpret these plots: what are they telling you?

D. Correlation matrix (or correlation heatmap)

# We can use the dlookr package to generate a correlation matrix (or correlation heatmap) by using the Pearson correlation coefficient, the Kendall distance, or the Spearman coefficent

```
pearson_correlation_matrix <-  patients_data %>% correlate(.,
method="pearson") %>% plot()

corr_matrix_file_width <- 15
corr_matrix_file_height <- 15
output_corr_matrix_file_name <- "Pearson_corr_matrix.pdf"

ggsave(pearson_correlation_matrix, width=corr_matrix_file_width,
height=corr_matrix_file_height, file=output_corr_matrix_file_name)
```

# Generate the correlation heatmaps for the Kendall distance and the Spearman coefficient, and save them into PDF files

# Interpret these images: what are they telling you?

E. Dimensionality reduction

# Let’s use UMAP to perform dimensionality reduction

```
thisNeighborsNumber <- 20
min_distance <- 0.01
```

```
# we need the ID column for the visualization
patients_data$"ID" <- row.names(patients_data) %>% as.numeric()


umap_fit_patients <- patients_data  %>%  column_to_rownames("ID") %>%
scale() %>%  umap(., n_neighbors = thisNeighborsNumber, min_dist =
min_distance)

umap_fit_patients %>% str()

# we merge the first two principal components of UMAP with the rest of the
patients_data variable

umap_df_patients <- umap_fit_patients$"layout" %>%
    as.data.frame() %>%
    rename(UMAP1="V1",UMAP2="V2") %>%
    mutate(ID=row_number()) %>%
    inner_join(patients_data, by="ID")

umap_df_patients %>% dim()
umap_df_patients %>% head()
umap_df_patients %>% colnames() %>% sort()


# we plot two first two principal components of UMAP, by highlighting age and sex as
color and shape of the points

pointSize <- 10
setFontSize <- 20

plot_title <- "UMAP plot on diabetes 1 dataset"

umap_plot_single <- umap_df_patients %>%
    ggplot(aes(x = UMAP1,
    y = UMAP2,
    color = age,
    shape = as.factor(sex_0man_1woman))) +
    geom_point(size=pointSize, alpha=0.5)+
    labs(x = "UMAP1",
    y = "UMAP2",
    subtitle = plot_title) +
    theme(text=element_text(size=setFontSize))


umap_plot_single

umap_plot_file_name <- paste0("umap_plot_neighbors",
thisNeighborsNumber, "_minDistance", min_distance, ".pdf")
```

```
plot_width <- 25
plot_height <- 15
ggsave(umap_plot_single, file=umap_plot_file_name, width=plot_width, height=plot_height)

# Generate new UMAP plots by changing the number of neighbors and the minimal distance
# Plot the UMAP results using another real variable different from age and another ordinal variable different from sex
```

F. Insert comments for all the previous R commands you used

Application of the main data cleaning and data preparation steps to a dataset of electronic health records of patients with diabetes type 2 from Saudi Arabia

Repeat all the steps of the previous analysis [A, B, C, ..., I]
Convert the file from XLSX to CSV, first.
Use Diabetic retinopathy (DR) as target for the data unbalance phase

For one-hot encoding, use the `one_hot()` function of the `nestedcv` package:
https://search.r-project.org/CRAN/refmans/nestedcv/html/one_hot.html

AlOlaiwi LA, AlHarbi TJ, Tourkmani AM (2018) Prevalence of cardiovascular autonomic neuropathy and gastroparesis symptoms among patients with type 2 diabetes who attend a primary health care center. PLOS ONE 13(12): e0209500.
https://doi.org/10.1371/journal.pone.0209500

Cerono G, Chicco D. 2024, "Ensemble machine learning reveals key features for diabetes duration from electronic health records". PeerJ Computer Science 10:e1896
https://doi.org/10.7717/peerj-cs.1896

2.2) pone.0209500.s001.xlsx file to download