Urinary Biomarkers for Pancreatic Cancer

Log Theme09 - Introduction Machine Learning

Lisa Hu
414264
Bio-Informatics
Hanzehogeschool Groningen, ILST

Dave Langers (LADR) & Bart Barnard (BABA)

September 21, 2022

Contents

1 Data description 2

1.1 Reading the data 2

1.2 Manipulate and analyse the data 3

```r
#' Setup chunk
knitr::opts_chunk$set(cache = TRUE)
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::read_chunk("EDA.R")
```

```r
#' Load all the packages
packages <- c("ggplot2", "tidyr", "dplyr", "readr", "tibble")
invisible(lapply(packages, library, character.only = T))
```

Dataset: Urinary biomarkers for pancreatic cancer

1 Data description

1.1 Reading the data

Running the following command:

```r
dataset <- read_excel("Data.xls")
```

Gave the following error:

```
filepath: <..>/Data.xls
libxls error: Unable to open file
```

Even though the filepath was correct, I tried different filenames and changing the method, but it still did not work. After scouring the internet, there was no clear solution to this. To avoid this error, open the xls-file in whichever program you have on hand for this file format. Depending on the program, `Export` (MS Excel) or `Save As` (LibreOffice Calc) the file to a CSV format (`Data.csv`).

From here, we can read the data and codebook:

```r
dataset <- read.csv("Data.csv")
codebook <- read_delim("codebook.txt", delim = "|")
knitr::kable(codebook[1:4], booktabs = T, caption = "Data values")
```

Table 1: Data values

Name	Full Name	Type	Unit
sample_id	Sample ID	chr	-
patient_cohort	Patient's Cohort	chr	-
sample_origin	Sample Origin	chr	-
age	Age of subject	dbl	-
sex	Sex of subject	chr	-
diagnosis	Diagnosis	dbl	-
stage	Stage	chr	-
benign_diagnose	Benign Sample's Diagnosis	chr	-
CA19_9	Blood plasma CA19-9	dbl	U/ml
creatinine	Creatinine	dbl	mg/ml
LYVE1	LYVE1	dbl	ng/ml
REG1B	REG1B	dbl	ng/ml
TFF1	TFF1	dbl	ng/ml
REG1A	REG1A	dbl	ng/ml

```
knitr::kable(codebook[c(1,5)], booktabs = T, caption = "Description")
```

Table 2: Description

Name	Description
sample_id	Unique string identifying each subject
patient_cohort	Cohort 1 = previously used samples; Cohort 2 = newly added samples
sample_origin	BPTB: Barts Pancreas Tissue Bank, London, UK; ESP: Spanish National Cancer Research Centre, Madrid, Spain; LIV: Liverpool University, UK; UCL: University College London, UK
age	Age in years
sex	M = male; F = female
diagnosis	1 = control (no cancer); 2 = benign hepatobiliary disease; 3 = PDA (pancreatic cancer)
stage	The stage of the disease (IA, IB, IIA, IIB, III, IV)
benign_diagnosis	The diagnosis for those with a benign diagnosis
CA19_9	Blood plasma levels of CA19-9 monoclonal antibody, usually elevated when pancreatic cancer
creatinine	Urinary biomarker of kidney function
LYVE1	Urinary levels of Lymphatic Vessel Endothelial Hyaluronan receptor 1
REG1B	Urinary levels of Regenerating Family Member 1 Beta
TFF1	Urinary levels of Trefoil Factor 1
REG1A	Urinary levels of Regenerating Family Member 1 Alpha

The information given in the codebook originates from the `Documentation.xls`. This file was given with the data file and can be found on the website.

1.2 Manipulate and analyse the data

A lot of the rows contain empty strings instead of NA, so set them to NA first. Besides that, the columns `sample_id` and `benign_sample_diagnosis` in the dataset are not valuable for the analysis and are therefor dropped.

```
# Change the empty strings to NA
dataset[dataset == ""] <- NA

# Remove unnecessary columns
drop <- c("sample_id", "benign_sample_diagnosis")
dataset <- dataset[,!(names(dataset) %in% drop)]

# Different diagnosis groups
control <- subset(dataset, diagnosis == 1)
benign <- subset(dataset, diagnosis == 2)
pdac <- subset(dataset, diagnosis == 3)

# Demographics
demograph <- data.frame(c(sum(control$sex == "F"), sum(control$sex == "M")),
                        c(sum(benign$sex == "F"), sum(benign$sex == "M")),
                        c(sum(pdac$sex == "F"), sum(pdac$sex == "M")))
colnames(demograph) <- c("Control", "Benign", "PDAC")
rownames(demograph) <- c("Female", "Male")

knitr::kable(demograph)
```

	Control	Benign	PDAC
Female	115	101	83
Male	68	107	116

```r
# Amount of blood plasma samples
nrow(dataset) - sum(length(which(is.na(dataset$plasma_CA19_9))))
```

```
## [1] 350
```

```r
blood <- subset(dataset, plasma_CA19_9 >= 0)
```