Project assignment

High Dimensional Data Analysis 2020

Adapted by Milan Malfait

19 Nov 2020

```
library(readr)
library(dplyr)
```

Introduction

Kidney transplantation or renal transplantation is the organ transplant of a kidney into a patient who has an end-stage renal disease. Scientists claim that some genes are responsible for a patient's likelihood of rejecting a kidney after transplantation.

In this project, you are to investigate this claim. You will analyze data that contains gene expression levels of 54675 genes from 282 patients, taken from the study by Einecke et al. (2010).

The data consists of a subset of the original data where only the 25% most variable genes (i.e. 13669 genes) were retained and the number of patients was subsampled to 250.

The data can be found on the GitHub page and can be read in directly using `read_csv`. However, the data is still quite large so it's recommended to download it locally to your computer, for example in a `raw-data/` folder and then loaded into R locally.

To do that, just run the code shown below in an R console. The output directory will be *relative to your current working directory* (which you can check with `getwd()`). You can also specify an absolute file path.

```
# Create raw-data/ folder if it does not exist (you can change this to whatever path you want)
out_path <- "raw-data"
if (!(dir.exists(out_path))) dir.create(out_path)

# Download data (only if doesn't exist locally)
# Delete existing file with `unlink(file.path(out_path, fname))`
# to force re-downloading
fname <- "GSE21374-kidney-data.csv.gz"
if (!file.exists(file.path(out_path, fname))) {
  data_url <-
    "https://github.com/statOmics/HDA2020/raw/data/GSE21374-kidney-data.csv.gz"
  download.file(data_url, destfile = file.path(out_path, fname))
}
```

The data can then be loaded into R as follows. The first two columns consist of the `Patient_ID` and `Reject_Status`, which is a **binary variable** encoding whether the kidney transplant was accepted (0) or rejected (1) for each patient. The other columns contain the microarray expression data for the 13669 genes.

```r
## Assumes data is stored in "raw-data/" folder under current working directory
data_dir <- "raw-data"
kidney_data <- read_csv(
  file.path(data_dir, "GSE21374-kidney-data.csv.gz"),
  col_types = cols(
    .default = col_double(),
    Patient_ID = col_character()
  )
)

## Glimpse first 10 columns
str(kidney_data[, 1:10])
#> tibble [250 x 10] (S3: tbl_df/tbl/data.frame)
#>  $ Patient_ID   : chr [1:250] "GSM533921" "GSM533922" "GSM533923" "GSM533924" ...
#>  $ Reject_Status: num [1:250] 0 0 0 0 0 1 1 0 0 0 ...
#>  $ X121_at      : num [1:250] 0.62 1.94 0.119 -0.798 0.228 ...
#>  $ X1255_g_at   : num [1:250] 0.815 1.39 -1.291 -0.368 -0.499 ...
#>  $ X1294_at     : num [1:250] -1.5589 -0.809 1.2385 -0.0565 -1.9907 ...
#>  $ X1405_i_at   : num [1:250] -0.376 -0.531 0.959 -0.638 -1.429 ...
#>  $ X1431_at     : num [1:250] -0.5659 -0.6719 -0.0688 -0.7883 -0.5423 ...
#>  $ X1552261_at  : num [1:250] 1.596 -0.281 0.735 -0.32 -0.972 ...
#>  $ X1552269_at  : num [1:250] 1.319 1.592 -0.854 -1.321 0.998 ...
#>  $ X1552274_at  : num [1:250] 1.731 0.139 -0.353 0.349 1.561 ...

## Extract gene expression data as matrix X
X <- kidney_data %>%
  dplyr::select(-Patient_ID, -Reject_Status) %>%
  as.matrix()
rownames(X) <- kidney_data$Patient_ID
dim(X)
#> [1]   250 13669
str(X)
#>  num [1:250, 1:13669] 0.62 1.94 0.119 -0.798 0.228 ...
#>  - attr(*, "dimnames")=List of 2
#>   ..$ : chr [1:250] "GSM533921" "GSM533922" "GSM533923" "GSM533924" ...
#>   ..$ : chr [1:13669] "X121_at" "X1255_g_at" "X1294_at" "X1405_i_at" ...

## Extract Reject_Status column as vector
reject_status <- kidney_data$Reject_Status
names(reject_status) <- kidney_data$Patient_ID
length(reject_status)
#> [1] 250
table(reject_status) # number of 0's (accepts) and 1's (rejects)
#> reject_status
#>   0   1
#> 174  76
```

We are interested in the following research questions:

- How do the genes vary in terms of their gene expression levels? Is the variability associated with kidney rejection? (only to be answered in a data explorative / graphical manner)
- Which genes are differentially expressed between the two kidney rejection groups? You must control the FDR at 10%.

- Can the kidney rejection be predicted from the gene expressions? What genes are most important in predicting the kidney transplant rejection? How well does the prediction model perform in terms of predicting rejection status?

  Note that the response variable is *binary*, so if you use regression models, you will need to use *logistic regression*, which models the response with a binomial distribution. This can generally be done in `R` by specifying `family = "binomial"` in regression functions such as `glm` and `glmnet`.

Assignment

You must work in groups of four students.

Write a scientific report that answers the research questions related to this study. The report must consist of two parts:

- An executive summary of about half a page. This summary contains the answers to the original research questions, and should be written in a non-technical manner (it is meant for researchers without a statistical background).
- A technical report that explains in detail how the results were obtained. It's recommended to prepare this technical report an **RMarkdown** file. If you choose to use another format, then the R code should be submitted as a separate file (please comment your R code).

The report is expected to be concise, but must evidently be accurate and sufficiently detailed to enable the reader to verify the correctness of the result (i.e. your results must be reproducible). The total length of the report (excluding graphs, R code and possibly appendices) should not be more than three pages. The report should not contain an explanation of the theory behind the statistical methods, and should also not contain the study description given above (you can assume that the reader already knows this).

However, *interpretation* of the results is key!

Some more specific guidelines:

- For the first research question, you may use one of the data exploration tools that we have seen in class. However, you are also free to search the literature for other techniques for data exploration and visualisation (your final mark will not depend on whether you searched the literature or not). You are only asked to *explore* whether the variability in gene expression levels is associated with rejection status; no need for hypothesis testing.
- For the second research question, you are asked to perform hypothesis testing and correct for multiple testing so as to control the FDR at 10%. The full list with differentially expressed genes may be presented in an appendix. Only list the most important results in the body of the report.
- For the third research question you are asked to predict rejection status using gene expression levels. You should randomly split the data into a test (30%) and a training (70%) dataset. Make sure you use a seed (`set.seed()` function) in R for reproducibility. The following prediction models should be evaluated:
  - Principal Component Regression (PCR)
  - Ridge Regression
  - Lasso regression

In choosing the number of PCs in PCR, and the γ in the Ridge and Lasso models, you need to use cross validation (CV) on the training dataset. You should use the **area under the receiver characteristic curve (AUC)** as a performance measure.

Once you have selected the optimal PCR, Ridge and Lasso models, you have to decide with what model you want to continue. For this model you have to determine a good threshold c for the prediction cut-off that gives a good compromise between sensitivity and specificity.

Use the test data for final performance evaluation in terms of *sensitivity* and *specificity*.

Submission

It is recommended (but not mandatory) to prepare your report in **RMarkdown**. You can render it to either HTML (`output: html_document`) or to PDF (`output: pdf_document`). In both cases the original `.Rmd` file should be included when handing in the assignment. If you don't use RMarkdown, you should include the `.R` file(s) containing your implementation and analysis scripts.

When submitting, please use the following format:

- HW-Name1-Name2-Name3-Name4.[pdf|html]
- HW-Name1-Name2-Name3-Name4.R[md]

where `Name` is your **family name**. It's also recommended to mention your full name in the report itself.

Submissions should be done **through UFora**.

The deadline for submission is 17/12/2020 at 23:59

References

Einecke, Gunilla, Jeff Reeve, Banu Sis, Michael Mengel, Luis Hidalgo, Konrad S Famulski, Arthur Matas, et al. 2010. "A Molecular Classifier for Predicting Future Graft Loss in Late Kidney Transplant Biopsies." *The Journal of Clinical Investigation* 120 (6): 1862–72.