Project assignment

High Dimensional Data Analysis 2021

Adapted by Milan Malfait

18 Nov 2021 (Last updated: 2021-11-12)

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 from the study by Einecke et al. (2010). The original data consists of microarray measurements from 54675 genes across 282 patients. For the purpose of this project, a random subset was made of 250 patients and 10.000 from the 25% most variable genes.

You can access the data through the ***HDDAData*** package, which can be installed from GitHub by running the following code in an R session:

```
if (!requireNamespace("remotes", quietly = TRUE)) {
    install.packages("remotes")
}
remotes::install_github("statOmics/HDDAData")
```

Note that you need to run this piece of code only once. *Don't* include it in your R script, just run it once in an R console and you're good to go.

You can then access the data in your script as follows:

```
library(HDDAData)
data("Einecke2010Kidney")

## Data dimensions
dim(Einecke2010Kidney)
#> [1]   250 10001

## Showing first 6 rows and 10 columns
head(Einecke2010Kidney[, 1:10])
#>           Reject_Status X211352_s_at X239275_at X1554536_at X1558452_at
#> GSM534173             0    2.3692560  1.0549150  1.24906517   0.8421884
#> GSM534125             0   -1.4882009 -0.6057843 -2.04274530   0.9472119
#> GSM534044             0    0.5561377  0.8382071 -0.72872103   0.3544353
#> GSM533968             0    0.1271639  1.4809816  0.19142188  -1.1555505
```

```
#> GSM534021             0   -0.3795643  0.3971213 -0.05223003   1.2730340
#> GSM533930             0   -0.3245208  2.1372173 -0.65713129  -0.3552231
#>           X1554249_a_at X234382_x_at X1557126_a_at X1554865_at  X224489_at
#> GSM534173     0.9914628    0.4619668    -0.6374899   1.4502116 -0.08625628
#> GSM534125    -0.9669676    1.3172260     0.6513907  -3.0633906 -0.89828665
#> GSM534044     1.5439325   -0.4722676    -0.3430743  -0.8369873 -1.30720408
#> GSM533968     0.6013397   -0.2955006     1.9974828   1.5790603  1.35680474
#> GSM534021     1.4170661    1.2925662     0.6802998  -0.3611688  2.38291415
#> GSM533930    -1.2812026   -0.4121304     0.6763814  -0.5040343 -0.59533172
```

More information about the data and its format can be found in the documentation, see `?Einecke2010Kidney`.

The first column, `Reject_Status`, consists of 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 10.000 genes. The `rownames` of the data frame contains the patient identifiers.

```
## Extract gene expression data as matrix X
X <- as.matrix(Einecke2010Kidney[, -1])
dim(X)
#> [1]   250 10000
str(X)
#>  num [1:250, 1:10000] 2.369 -1.488 0.556 0.127 -0.38 ...
#>  - attr(*, "dimnames")=List of 2
#>   ..$ : chr [1:250] "GSM534173" "GSM534125" "GSM534044" "GSM533968" ...
#>   ..$ : chr [1:10000] "X211352_s_at" "X239275_at" "X1554536_at" "X1558452_at" ...

## Extract Reject_Status column as vector
reject_status <- Einecke2010Kidney$Reject_Status
names(reject_status) <- rownames(Einecke2010Kidney)
table(reject_status) # number of 0's (accepts) and 1's (rejects)
#> reject_status
#>   0   1
#> 183  67
```

Assignment

You must work in groups of four students.

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 expression levels? 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`.

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. The reader should understand what you did without looking at the code!

It's recommended to prepare the report as an **RMarkdown** file. If you choose to use another format, then the R code should be submitted as a separate file (please properly 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).

***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 should 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

Note that the response variable (the rejection status) is *binary*, so you will have to use **logistic 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. An example of how to do this for PC Regression can be found on UFora.

Once you have selected the optimal PCR, Ridge and Lasso models, you have to decide with what model you want to continue. For this final model you have to determine a good threshold c for the prediction cut-off, where c threshold where when a predicted probability from your final model $p > c$, we predict kidney rejection ($y = 1$) and non-rejection ($y = 0$) otherwise. c should be chosen such that the *misclassification error* is minimal. This should be evaluated on the *test data*, i.e. the data not used for training the model. Also report the *specificity* and *sensitivity* of your final model and cut-off c.

Hint: You can find an example function to calculate the misclassification errors for a range of cut-off values below. This should help you find an optimal value for c.

```r
## Inputs:
##   * obs: vector of test observations
##   * pred: vector of model predictions
##   * cutoff_values: vector of prediction thresholds c
calculate_misclass_error <- function(obs, pred, cutoff_values) {
  stopifnot(lenth(obs) == length(pred))

  misclass_errors <- rep(NA, length(cutoff_values))

  for (i in seq_along(cutoff_values)) {
    cutoff <- cutoff_values[i]
    ypred <- as.numeric(pred > cutoff) # translates TRUE/FALSE to 1/0

    misclass_errors[i] <- mean(ypred != obs) # proportion of misclassifications
  }

  data.frame(
    "cutoff" = cutoff_values,
    "misclass" = misclass_errors
  )
}
```

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 16/12/2021 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.