Analysis of High Dimensional Data

Project assignment

Introduction

Kidney transplantation or renal transplantation is the organ transplant of a kidney into a patient who has an end-stage renal disease. Scientist 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 Gene Expression Omninibus (GEO)[1].

Read in the data. There are two datasets, `GeneExpression` and `RectionStatus`. `GeneExpression` is a 282 by 54675 dataset with gene expression data. `RectionStatus` contains a binary variable (0 (accept) and 1 (reject) kidney transplant) for each patient. The patient ID is present in each dataset.

```
load("X_GSE21374.rda")

#Gene expression data
GeneExpression<-t(X_GSE21374)
#Dimension of the data
dim(GeneExpression)
```

```
## [1]   282 54675
```

```
#View gene expression data
head(GeneExpression[,1:10])
```

```
##             1007_s_at    1053_at      117_at      121_at  1255_g_at     1294_at
## GSM533921 -0.50988780  0.9072850 -0.42639657  0.62039390  0.8151333 -1.55894428
## GSM533922  2.09927415  0.4727166 -0.97043274  1.93988527  1.3897381 -0.80895310
## GSM533923  1.51506539  0.8619308 -0.09526958  0.11933263 -1.2914004  1.23850599
## GSM533924 -1.67054330 -1.0424478 -0.19542826 -0.79811395 -0.3675015 -0.05650636
## GSM533925 -0.26615577 -1.2217764 -1.08934050  0.22782776 -0.4991056 -1.99072799
## GSM533926  0.02194967  1.4102739 -0.42394684 -0.08549078 -0.6582384  0.16832801
##              1316_at     1320_at  1405_i_at     1431_at
## GSM533921 -2.2944621 -0.02918161 -0.3755807 -0.56590799
## GSM533922 -0.1519210 -1.46531016 -0.5313299 -0.67187592
## GSM533923 -0.7926129 -1.55173291  0.9585930 -0.06883004
## GSM533924 -0.8753529 -1.69825331 -0.6383800 -0.78833139
## GSM533925 -1.7642104 -0.53798514 -1.4290300 -0.54232611
## GSM533926 -1.2425576 -1.88140132  0.2976891 -0.41866762
```

```
load("RejectionStatus.rda")
#View the response variable
head(RejectionStatus)
```

```
##   Patient_ID Reject_Status
## 1  GSM533921             0
## 2  GSM533922             0
## 3  GSM533923             0
## 4  GSM533924             0
## 5  GSM533925             0
## 6  GSM533926             1
```

The research questions can be summarised as follows.

- How do the 54675 genes vary in terms of their gene expression levels? Is the variability associated with kidney rejection? (only to be answered in a data explorative 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?

Assignment

You must work in groups of at most four students.

Write a scientific report that gives answers to the research questions related to this study. The report must contain 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. We suggest that this technical report is prepared as an R markdown 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 (exclusive 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 has knows this).

Some 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 aksed 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 in R for reproducibility. The following prediction models should be evaluated:

  1. *Principal Component Regression (PCR) prediction model*: The crucial point for building a prediction model with PCR is to determine which PCs should be included in the prediction model. Although not the best approach, you may simplify the problem by only considering models with

predictors, PC1, PC2, …, PCk (PCs ordered with decreasing variance), so that you only have to determine an optimal k (i.e. you only have to determine how many PCs have to be included).

2. *Ridge Regression prediction model*: Model selection in the Ridge model happens through searching for the optimal value of the γ tuning parameter.
3. *Lasso regression*: Model selection in the Lasso model happens through searching for the optimal value of the γ tuning parameter.

In choosing the number of PCs in PCR, and the γ in the Ridge and Lasso models, you need to use cross validation (CV) in the training dataset. You should use the area under the reciever 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 that gives a good compromise between sensitivity and specificity.

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

Submission

You must submit your files via the Minerva website (via the dropbox of the course site). Please note that only pdf, Rmd and R files will be accepted (Word doc or docx files will not be accepted) and that the name of the files should be the family names of the people in the group, ending with `Project`, e.g. `Thas-Meys-Project.pdf`. Your R code should be commented and submitted as a separate file with the same name as the pdf report file, but with the .R or .Rmd extension (e.g. `Thas-Meys-Project.Rmd`). If you prepared your report with R markdown, you should also submit the pdf version.

Evaluation

Since collaboration and communication in group work is a “learning outcome” of this course, you are asked to also submit an evaluation of your peers. We will use a tool that is integrated in Ufora (more details will follow later). Your peer assessment will be treated confidential.

References

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10