Information about project

Data Mining and Machine Learning I

Deadline: 16 July 2021

Discriminating between healthy individuals and patients with Pulmonary Arterial Hypertension

The importance of genetic predisposition, inflammation, and auto-immune mechanisms in the development of pulmonary arterial hypertension (PAH) is becoming increasingly clear. It is hypothesised that the development of PAH requires first a genetic susceptibility followed by one or several secondary trigger factors such as a viral infection or drug exposure. The individual's genetic susceptibility and the interaction of the genotype with the promoting factor(s) still remain areas of active research.

Project goal

For this project we will hypothesise that the analysis of gene expression profiles from peripheral blood mononuclear cells would distinguish patients with PAH from healthy individuals. The question of interest is which genes (or combinations of genes) have an impact on patient status (hypertensive or healthy).

The project aim is to fit a variety of classification algorithms that you have been taught during weeks 3 to 5 to your dataset and decide which is the best at predicting hypertensive and healthy individuals.

Data set

A data set comprised of the statistics of 7139 genes from 20 individuals (the first 14 are hypertensive while the last six are healthy). You can load the data set in R with the following command:

```
dataset <- read.csv("../Data/GDS504.clean.csv",header=TRUE)
head(dataset)
```

You can also check the dimensions of the data set using the `dim` command.

```
dim(dataset)
```

You can see that the rows of the data set refer to the genes and the columns refer to the samples. You can also notice that there isn't an class variable included for identifying which are the healthy individuals and which are the individuals with PAH. You would have to find a way to do that in R based on the previous description of the problem at the start of this section (in order for the models you choose to distinguish between these individuals).

Assignments of genes

Each student will study the performance of the classification techniques on a sub-region of the available data set. To find your assigned range, look at the row containing your student number in the following table.

ID	First gene	Last gene
2608851	1	1000
2609076	151	1150
2600639	301	1300
2603599	451	1450
2614867	601	1600
2611023	751	1750
2605175	901	1900
2600767	1051	2050
2585814	1201	2200
2601523	1351	2350
2602334	1501	2500
2602483	1651	2650
2611484	1801	2800
2560535	1951	2950
2612369	2101	3100
2592555	2251	3250
2600094	2401	3400
2505553	2551	3550
2614765	2701	3700
2611926	2851	3850
2610621	3001	4000
2603608	3151	4150
2600739	3301	4300
2601143	3451	4450
2561814	3601	4600
2611850	3751	4750
2610003	3901	4900
2466935	4051	5050
2612326	4201	5200
2611469	4351	5350
2598932	4501	5500
2611831	4651	5650
2610050	4801	5800
2599304	4951	5950
2576050	5101	6100
2610368	5251	6250
2609222	5401	6400
2610083	5551	6550
2608655	5701	6700
2613901	5851	6850
2604756	6001	7000
2599942	51	1050
2610369	201	1200
2597026	351	1350
2514088	501	1500
2424772	651	1650
2609514	801	1800
2611002	951	1950
2603873	1101	2100
2615246	1251	2250
2598820	1401	2400
2487359	1551	2550

ID	First gene	Last gene
2542827	1701	2700
2611948	1851	2850

Project assessment

The project is assessed on a report and corresponding R code submitted. This will be worth 40% of the overall grade for the course. On Moodle you should upload two files. One is your report (as a pdf document) and the other one should be an R script that allows us to reproduce all the statistical analysis you refer to within your report. The deadline for submission is **Friday July 16th 2021 at 23:00 BST** (and your assessment report and code must be uploaded on Moodle by that time).

Report and R code - tasks

After you create the training, validation and test data sets, you have to:

- Perform exploratory analysis on the training data set.
- Apply all classification techniques that you have learned during weeks 3 to 5 (e.g. k-nearest neighbours, tree-based methods and support vector machines).
- Create appropriate graphs or summaries that communicate the results from these methods.
- Comment on these and interpret the results.
- Compare the results to choose a final, "best" classification model (with justification given for the choice) and comment on this model's overall classification rate, sensitivity and specificity for future sample predictions.
- Create a 10 page report that includes all the previous information. (This refers to the maximum number of pages, and does not include a title page if you wish to have one or the R code)
- Suggested report structure: Introduction, Exploratory Analysis, Results, Discussion sections

The R code should allow us to:

- install (and load) any packages you might have used,
- clearly show which set of variables you were working on (since we will not be able to reproduce this),
- reproduce any models you have worked on and
- recreate any of the graphs and summaries you have on your report.

The R code should be well enough commented that the code is understandable to one who hasn't written it.