Projet d'evaluation : Biostatistiques approfondies (Stat3)

Aitor González & Quentin Ferré

Master Bionformatique DLAD, 2019/2020, Semestre 3

Grading of the module

This module's evaluation will consist of two parts :

- A final individual exam on paper
- A project submitted with
  - the python code
  - a report (2-4 pages) with the project introduction, results, methods, conclusion and discussion

Each of the project and the exam will account for half of your final note.

This document presents the instructions for the project.

Goal

In this project, we will provide you with a dataset and some guidelines, and we would like you to show what you learned. There is no point in trying to copy what other students have done. We will value originality.

Data source

You will work with real data on this project. The dataset is available at :

https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

The button to download the data is on the top right corner.

Instructions

The underlying question is to build a model that can predict the cancer class from the given features.

The assignment is free-form. You can select an algorithm and apply it to get the best results, or compare algorithms, or you can add descriptive statistics, or more. The choice is yours, but keep in mind that your goal is to **show us what you have learned**. Hence, we expect you to show examples of methods and approaches seen in class, such as following the correct design process of a machine learning system (learning curve, etc.)

Advice

- You have more features than examples, so be wary of the curse of dimensionality.
- The data formatting is standard : one line per example (sample, cell line, patient...) and one column per feature (gene,...). In bioinformatics, especially in genomics, this is what you will almost always work with.

Deliverables

This section lists what we expect you to send us at the end of your work. The report must be submitted to the Ametice link: Remise du rapport et projet (Informations pour les étudiants) before Jan 7, 2020 at 23:59. There will be penalties for delayed submission.

Code

The complete Python code for your analysis. Although it is not mandatory, we suggest you use the following directory structure :

- data : self-explanatory
- doc : report and documentation
- lib : functions
- output : your results
- main.py : main python code
- Readme.md : entry point for your project. *We will begin by reading this file.*

Your code should be set up so that we can run your entire project by simply typing `python3 main.py` in a console. You can assume that we have all standard Python3 modules installed, as well as *SK-learn*, *Keras*, *Numpy* and *Pandas*.

We will reward :

- **Reproductibility** including but not limited to : presenting your full code and data and a proper use of Conda or Docker,
- **Conviviality** with appropriate documentation of your code,
- **Ease of access** by using Git (GitHub or Renater SourceSup) as an alternate method of delivery for your work instead of AMETICE should you wish it.

There is a link to submit the code.

Report

We will also expect a report of roughly 2 to 4 pages, in which you will **justify your choices** and **analyze your results.** We will value the **clarity** of your report. We want the report in PDF format. There is a link to submit the report.

It can be structured as a report or a scientific article (Introduction, Materials and Methods, Results, Discussion), depending on what you think is best for your approach.

Good luck!!