Bioinformatics course project: prediction of the regulatory regions active in specific cell lines through deep learning methods

Luca Cappelletti & Giorgio Valentini

April 2021

1 Project goal

The project's main objective is to develop deep neural network models for the prediction of active regulatory regions in specific cell lines [3, 2, 1]. The main focus of the project is to learn how to construct such models and how to apply correct experimental procedures to obtain sound and statistically reliable results.

2 Project content

The project (and the project report) must include the following parts:

2.1 Introduction

A brief introduction (a single page) that illustrates the problem faced, what it consists of, its importance, the current state of the art (which published works have been carried out in this area), and what the project is based on is unfolding.

2.2 Models

The section will include a description of the selected models (and/or meta-models), the comparison models if any (like Random Forest, Decision Tree etc.), the chosen parameters (or space of hyper-parameters in case of model selection).

Whenever possible, provide a reason for choosing one type of model over another.

Illustrate the models' structures with tables and images.

2.3 Experimental setup

In the experimental setup, it is necessary to clearly explain what the experiment consists of and provide enough information to reproduce it. The tasks to study are *Active enhancers vs inactive enhancers (AEvsIE)* and *Active promoters vs inactive promoters (APvsIP)* on the HG38 dataset.

The tasks **MUST** be tackled with the three neural network model architectures seen during the course: *Feed-forward neural network (FFNN)*, *Convolutional neural network (CNN)* and *multi-modal neural network (MMNN)* (either by combining the two model previously developed or just defining a new one). To each model will be assigned per task.

2.3.1 The considered tasks

Briefly explain what the tasks addressed with the models are.

2.3.2 Dataset source

Please explain what the data sources are and the procedures for downloading them from the DBs. The data to be used were presented in the lessons available on the course's GitHub repository: `https://github.com/LucaCappelletti94/bioinformatics_practice`.

The dataset for the HG38 assembly can be downloaded directly from `https://github.com/AnacletoLAB/epigenomic_dataset`.

2.3.3 Data pre-processing

How the data is preprocessed before being passed to the models, like imputation[1] and normalization[2].

2.3.4 Data correlations and distributions

Correlation analysis and study of data distributions, including any visualizations (e.g. scatter plot [3]).

[1] In the course we have used `https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html`

[2] In the course we have used `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html`

[3] The scatter plot library used during the course is `https://seaborn.pydata.org/index.html`

2.3.5 Feature selection

Where possible, perform a feature selection[4].

2.3.6 Data visualization

View the data using one or more proposed methods like t-SNE[5] or PCA[6] (or others available).

2.3.7 Holdouts

Clearly explain how the holdouts were carried out.

2.3.8 Results analysis [1 point]

Explain which metrics will be used and which statistical test will compare the models' performance.

2.4 Results

Adequately present, through bar graphs[7] and summary tables, the performance of the various models. Comment on the results obtained and validate the statements made through statistical tests[8].

3 Project organization

3.1 Students groups logistic

The groups are created based on the data provided to the form available on Google Sheet at the following URL:

`https://docs.google.com/spreadsheets/d/1Qm0KfROzXOhzC_qw0h41ut3gl5UMhS4sA3RcFqUuYLA/edit`

In the same Google Sheet, there is a page to propose groups.

3.2 Division of work within the groups

To make each group's work proportional to the number of people in a group, a cell line to be analyzed will be assigned for each person in the group.

The evaluation of each student will consider the group's work and the specific individual contribution of each member of the group, who will be responsible for the experimental part relating to a specific cell line.

3.3 Submitting the complete project

A complete report of the project must be delivered by email one week before the exam. Along with the report's PDF, a link to the code repository must be provided, which must be properly documented. Ideally, the code should be tested and test coverage added in the repository README.

3.4 Structure of the oral exam

The decisions made on the established experimental setup will be the main oral exam topics. We will then review together the statements made in the conclusions of the project and whether these are valid from a statistical point of view. If there are any errors of any kind, we will discuss how to fix those errors.

[4] During the course, we have used the `https://github.com/scikit-learn-contrib/boruta_py`

[5] A good library for computing t-SNE is `https://github.com/CannyLab/tsne-cuda`, here you can find ready a Docker image: `https://github.com/LucaCappelletti94/default_docker_images/tree/master/cuda-tsne-docker`

[6] A good library for PCA is `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`

[7] A sound library for this goal is `https://github.com/LucaCappelletti94/barplots`

[8] In the course we have used the Wilcoxon signed rank test, available from `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html`

References

[1] Luca Cappelletti, Alessandro Petrini, Jessica Gliozzo, Elena Casiraghi, Max Schubach, Martin Kircher, and Giorgio Valentini. Bayesian optimization improves tissue-specific prediction of active regulatory regions with deep neural networks. In Springer, editor, *Bioinformatics and Biomedical Engineering, IWBBIO 2020*, Lecture Notes in Computer Science, 2020.

[2] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[3] Yifeng Li, Wenqiang Shi, and Wyeth W. Wasserman. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics*, 19(1):202, 5 2018.