Miniproject BIO-322

November 2022

1 Introduction

In this miniproject you can demonstrate your conceptual knowledge and your coding skills on an real data set. To make it fun for you – and following a tradition in machine learning – we organise this miniproject around a competition (see `https://www.kaggle.com/competitions/epfl-bio322-2022`).

In an experiment on epigenetics and memory, Giulia Santoni (from the lab of Johannes Gräff at EPFL) measured the gene expression levels in multiple cells of a mouse brain under three different conditions that we call KAT5, CBP and eGFP. In this challenge, the goal is to predict – as accurately as possible – for each cell the experimental condition (KAT5, CBP or eGFP) under which it was measured, given only the gene expression levels.

The training data contains the normalized counts for 32285 genes in 5000 different cells together with the experimental condition under which each cell was measured. The test data contains only the normalized counts and your task is to predict the experimental condition. You can download the data here:

`https://lcnwww.epfl.ch/bio322/project2022/train.csv.gz`

`https://lcnwww.epfl.ch/bio322/project2022/test.csv.gz`

You can upload predictions on the test set to `https://www.kaggle.com/competitions/epfl-bio322-2022` in the format described here:

`https://www.kaggle.com/competitions/epfl-bio322-2022/overview/evaluation`.

Your submissions are measured with the classification accuracy metric, i.e. the number of correct predictions divided by the total number of predictions.

Although you can get bonus points based on your rank in this competition, the main evaluation criteria of your miniproject are based on reproducibility of your results, readability of your code and a report that summarises your approach and findings.

2 Rules

- You are allowed to compete alone but we really encourage you to collaborate in teams of 2 students. Collaboration across teams is not allowed. We may run plagiarism detection software on your submissions.
- In teams of 2 students, each team member has to contribute significantly to the project. We may interview team members, if we suspect one of them got a free ride.
- Your rank on the competition leaderboard counts only if your solution is fully reproducible with the code you submit.
- You host your code and your report on a private git repository on `https://github.com/` and give read access to the github user epfl-bio322.
- You write a report of your findings on at most 2 pages A4 with font size at least 11 points. The report has to be included as a pdf file in your private git repository.
- Your git repository contains a README, where you explain briefly, how your repository is organised and how your results can be reproduced, your code – in the form of one or multiple notebooks or scripts – and your report as a pdf file.

3 Deadlines

- 2 December, 18h00: **Communication of team members and git repository**. As soon as you have formed your team and set up a private git repository – it should not be visible to the public – give read access to your repository for user epfl-bio322 and communicate us the team members and the address of your git repository through the questionnaire `https://moodle.epfl.ch/mod/questionnaire/view.php?id=1182315` (per team one entry in the questionnaire).
- 23 December, 18h00: **Final submission**. At this moment the competition on kaggle closes and we will pull the content of your main branch from your private repository. The evaluation of your miniproject will be based on the content of your main branch. Do not forget to include your report as a pdf in your repository.

4 Evaluation Criteria

Your code and report must contain

1. the results of some exploration and visualisation of the data,
2. the best results you obtained with a linear method,
3. the best results you obtained with at least one non-linear method.

For points 2 and 3 you may want to experiment with regularisation or feature engineering.

Report (8 out of 20 points)

The goal of the written report is to show what you did and that you understand what you are doing. This does not mean that you need to explain everything that was shown in class. You can assume that the reader knows what cross-validation or a neural network are. However, the reader should be able to reproduce your work, your results and understand why you made particular choices. Your choice of method need to be well motivated and you need to show evidence that your work has an effect. The simplest way to do so is to start with a simple model as a baseline, evaluate it, find a way to improve it, evaluate again and repeat. Explain the process that leads to your various improvements, evaluate the results carefully and present evidence.

Your report could contain answers to the following questions:

- Is a linear methods sufficient to classify the data or are non-linear methods needed for high classification accuracy?
- Are some features (gene expression levels) more important than others to make accurate predictions? Which (combination of) gene expression levels is highly predictive and sufficient to correctly classify cells?
- How many clusters can be identified in the input data? What could the different clusters correspond to?
- Is classification easier for some clusters and more difficult for other clusters?

We do not want to limit your creativity. Many things can be done with the data and we appreciate additional creative questions and answers!

- **Content**: accurate and succinct description of your project; the report is in agreement with the submitted code. (6 points)
- **Structure**: Your report is well structured, e.g. introduction, results, conclusions (1 point)
- **Language**: Your report is written in good English. Hint: use a spell-checker and re-read your text carefully. (1 point)

Code (12 out of 20 points)

- **Content**: runnable code and reasonable hyper-parameter tuning. (9 points).
- **Readability**: The code is well structured and readable (1 point).
- **Reproducibility**: By running your code on our machines we can reproduce exactly the files you submitted to the kaggle competition. (2 points)

Leader Board (up to 4 bonus points)

As soon as you have some results you can upload them to kaggle. You can **make at most 5 submissions per day** (so don't wait until the last day with your submissions). On kaggle there is a public and a private leaderboard. The public leaderboard shows evaluations of all your submissions based on 20% of the test set. The private leaderboard shows evaluations of two submissions on the remaining 80% of the test set (this is the real test set) after the end of the competition. **Don't overfit the public leaderboard; the final ranks based on the private leaderboard may differ.** On kaggle you can select the submissions you believe are the best ones in the section "My Submissions" `https://www.kaggle.com/competitions/epfl-bio322-2022submissions`; otherwise your best entries on the public leaderboard are taken. If you get full reproducibility, you can collect up to 4 bonus points on the private leader board on kaggle. The bonus points are given by

$$\max(0, (17 - \text{your.rank.on.the.private.leader.board})/4)\,.$$

We hope you enjoy the project! Good luck!