Miniproject BIO-322

November 2025

1 Introduction

In this miniproject you can demonstrate your conceptual knowledge and your coding skills on a real data set. To make it fun for you – and following a tradition in machine learning – we organize this miniproject around a competition. The task is to predict from neural activity of mice what they perceive and do. You need to create an account on kaggle with **your EPFL email address** to access the data and submit solutions:

`https://www.kaggle.com/t/c7e4ffade6c648dfbcab945ef2ee0186`.

Parviz Ghaderi from Carl Petersen's lab at EPFL recorded the behavior and neural activity of mice in response to different stimuli. The experiments were divided into different sessions (with different mice on different days), with each session consisting of multiple trials. Each trial lasted 3 seconds and consisted of different combinations of stimuli (auditory tone or whisker deflection) provided to the mouse, and a potential lick response from the mouse. The mouse received a small reward whenever it licked in response to a certain combination of stimuli, and no reward otherwise. For this project, Tâm Nguyen prepared the raw experimental data so that we can try to read the minds of the mice:

Goal: Given the average neural activity in 100ms time bins of different kinds of neurons in different brain areas, predict the trial type, i.e., the perceived stimuli and the lick response.

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 written summaries of your approach and findings.

2 Rules

The goal of this project is to prepare you for future projects where you apply machine learning methods to research or industry data. In short: anything in agreement with this goal is allowed; not allowed is everything that aims at getting a good grade without learning anything.

- You are allowed to compete alone, but we really encourage you to collaborate in teams of 2 students.
- 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.
- Code or text sharing across teams is not allowed. We may run plagiarism detection software on your submissions.

- You are allowed to use modern tools like ChatGPT or CoPilot to help with coding and writing. Make sure to check very carefully the output of these models: surprisingly often they still produce outputs that look superficially correct, but are subtly wrong. In any case, each team member must be able to explain every line of code and text. We may interview team members.
- Your rank on the competition leaderboard counts only if your solution is fully reproducible with the code you submit. Make sure to set the random seeds wherever needed.
- You submit your findings in a single jupyter notebook that contains the code, figures and text explaining your findings.
- You host your code on a private git repository on `https://github.com/` and give read access to the github user epfl-bio322. Your git repository must contain the final jupyter notebook, and, for reproducibility, details about which python version, package versions and operating system (Windows, macOS or Linux) was used, e.g. in a pyproject.toml (recommended), requirements.txt or a README.md. The repository may contain other files.

3 Deadlines

- 21 November, 18h00: **Communication of team members, kaggle team name and git repository**. As soon as you have formed your team, created a kaggle team (with your EPFL email addresses) 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 (on github go to Settings → Collaborators and search for "epfl-bio322") and communicate us the team members, the kaggle team name and the address of your git repository through the questionnaire `https://moodle.epfl.ch/mod/questionnaire/view.php?id=1182315` (one entry per team, please).
- 19 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. Make sure to push regularly to the repository, such that you have a close-to-final version well ahead of the submission deadline. Unless github is not functional at the submission deadline (you can check here: `https://www.githubstatus.com/`), we do not accept any excuses for late submissions. Late submission penalty: -0.25 grade if you submit on 19 December between 18:00 and 24:00, -1.0 grade otherwise.

4 Evaluation Criteria

Your final notebook must contain the following sections

1. Data Inspection: In this section you load, explore, visualize the data.
2. Preprocessing: In this section you have code for preprocessing, cleaning and feature engineering.
3. Linear Model: In this section you train and tune a linear method, e.g. with regularization, as a first baseline.
4. Non-Linear Models: In this section, you train and tune at least one non-linear method.
5. Summary and Conclusions: This section does not contain any code, but a summary and conclusion of your findings in the form of figures and text.

Your notebook could contain answers to the following questions:

- Is a linear method sufficient, or are non-linear methods needed for high accuracy?
- For which machine learning method are transformations of the data needed, and which kind of transformations work best?
- Which predictors are important?
- Which trials are the easiest to predict and why?
- Is it easier to predict what the mouse perceives or what it does?
- Do predictions generalize across days and mice?
- ...

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

The notebook will be evaluated on the following points:

- **Content**: runnable code, reasonable hyperparameter tuning, accurate and succinct descriptions of the overall strategy and hypotheses. (12 points).
- **Readability**: The notebook is well-structured and readable. The comments, descriptions and summaries are written in good English and consistent with the code. (6 point).
- **Reproducibility**: By running your code on our machines, we can reproduce exactly the files you submitted to the kaggle competition. (2 points)

We will not run all the time-consuming code that you submit, but we will sample some of your code and check if we can reproduce the submitted results. It is advisable to save intermediate results to disk, such that you do not always have to restart from scratch.

Leader Board (up to 2 bonus points)

As soon as you have some results, you can upload them to kaggle. You can **make at most 5 submissions per day** (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/t/c7e4ffade6c648dfbcab945ef2ee0186submissions`; otherwise, your best entries on the public leaderboard are taken. If you get full reproducibility, you can collect up to 2 bonus points on the private leader board on kaggle. The bonus points are given by

$$\max(0, (17 - \text{your.rank.on.the.private.leader.board})/8)\,.$$

We hope you enjoy the project! Good luck!