Miniproject BIO-322

November 2024

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 the purity level of heroin samples. 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/9d0ee709f4ef4abe83737920ceed9d72/`.

The police often need to ascertain whether confiscated substances contain illegal drugs, such as cocaine or heroin. It is also crucial to determine the purity of samples, i.e. the quantity of pure cocaine or heroin, as this information is used by law enforcement and prosecutors to determine if the case is minor (more related to consumption) or a major crime (related to drug trafficking). Since illicit drugs are rarely traded in their purest form but are typically mixed with other substances, it is important to provide techniques that enable qualification and quantification in real time. Traditional methods for accurate analysis involve expensive and time-consuming chromatography procedures in laboratories. An alternative, cost-effective method involves using portable devices that allow the police to record the infrared spectrum of the samples and employ machine learning techniques to predict the purity level based on the infrared spectrum.

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: they produce often

outputs that are subtly wrong, but look superficially correct. 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 requirements.txt or a README.md). The repository may contain other files.

3 Deadlines

- 22 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).
- 20 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.

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?
- Is it possible to identify the substances used to mix with heroin?
- ...

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/9d0ee709f4ef4abe83737920ceed9d72/submissions`; 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!