Laboratory exercises and project instructions

Introduction to Data Science – Academic Year 2023/2024

Laboratory exercises

Starting with the first homework assignment, given by an assistant during the auditory exercises in the 2nd week of classes, the student's task will be to try for themselves at home the Python instructions learned during the auditory exercises.

An assistant at each auditory exercises will provide a dataset and short instructions about which parts of materials covered in the auditory exercises will be expected from students to repeat (or try in case of some additional materials) on the provided dataset. Student will then provide the Python code for the given task as a homework. Student can keep the code of the homeworks on the local computer until the deadline for delivery of all the homework materials arrives.

Before the delivery deadline, the student needs to create an appropriate GitHub repository and add the course assistants as collaborators to the project (Settings → Collaborators → Add people). Detailed instruction on how to do that and the specific accounts of the assistants that need to be added will be provided during auditory exercises.

Student will then need to upload the solutions for the homeworks in the form of a single Jupyter Notebook to the opened Github repository.

The assistants (one or few of them) will then examine the homework solutions together with the student and grade the student's homeworks accordingly. Each part of the homework will be graded with points according to the table provided in the course schedule file.

Deadline for laboratory exercises homework delivery: **January 10, 2024**

Maximum number of points: **15**

Project

At some point during the semester, students need to select one of the offered scientific articles for their project work. A maximum of 5 students can apply for each article. Students who do not select an article by the article selection deadline will be randomly assigned to articles where there are vacancies left.

The articles offered (and the assistants in charge of them) are available in this table. In the same table, student makes a selection of an article by typing his/her first and last name under the desired article.

Deadline for project article selection: **January 10, 2024**

First part of the project

In the first part of the project, students work on preparing and visualizing the data associated with the article. Students should first read the selected article and then download the data that was used in the article. After that, it is necessary to familiarize oneself with the data. Guidelines on how to do this are:

- load data
- check all data types and display descriptive data statistics
- check if there are missing values and outliers
- visualize data in several different ways (e.g., a feature histogram, line charts, scatter plots, ...)
- ...

Students are not limited to the items listed above, they serve only as guidance. During the development of this part of the project, a Jupyter Notebook needs to be prepared in which the whole process of data preparation and visualization for the selected article is provided.

Students will then need to upload the data preparation and visualization project solution in the form of the single Jupyter Notebook to the previously opened GitHub repository.

The assistant in charge of the article will then examine the solution together with the student and grade the first part of the project accordingly.

Deadline for the first part of project delivery: **January 19, 2024**
Maximum number of points: **20**

Second part of the project

In the second part of the project, students need to replicate the results achieved in the paper. This part of the project is considered to be more difficult than the first part and should be pursued only by significantly involved students who wish to obtain maximal course results.

For accomplishing this task, student needs to use the approaches from the selected article to replicate the study's results. During the implementation itself, students can use already implemented functions from packages such as *numpy*, *scikit-learn*, etc. Once the methods are implemented, they need to be run on previously prepared data, correctly evaluated, compared with the results from the article and with explanation of possible differences.

For example, if a student is working on a classification problem, it is advisable to display:

- the value of metrics, such as accuracy, precision, response, etc.
- AUC/ROC curves
- confusion matrix
- . . .

Student is free to display the results in alternative ways that seem interesting, but the main goal is to replicate the results achieved by the authors of the paper. In the event that the

selected article deals with a specific issue, it is advisable to contact the assistant in charge for advice.

Once the replication is complete, student will then need to upload the solution in the form of a single Jupyter Notebook to the previously opened GitHub repository.

The assistant in charge of the article will then examine the solution together with the student and grade the second part of the project accordingly.

Deadline for the first part of project delivery: **January 24, 2024**
Maximum number of points: **5**

Additional notes

- You do all of the above using Python and Jupyter Notebooks. The final version of the notebook you will submit (at each point of the laboratory exercises / project verification) must contain comments/conclusions of all the steps taken. The notebook must be trackable without reading a lot of codes. The notebook should be submitted by placing it on the opened GitHub repository.

- If you don't have a computer with enough resources to do everything you need to do as part of the project, we suggest you use Google Colab.

- At the end of each phase of the laboratory exercises / project, the assistants will go through each student's solution separately through the submitted Jupyter notebook, and the student will need to explain the implemented code, results and conclusions. Based on this examination and the submitted solution, the assistant will award the student the appropriate number of points.

- In the event that the student is late with the submission of the solution for a particular phase of the laboratory exercises / project, that part will be scored with 0 points.

- If the student does not collect a minimum of 25% of the points from the whole laboratory exercise / project (10 points), the student is not entitled to take the final exam (as well as the regular exams).

- Both laboratory exercises and the project are considered to be individual student assignements. This means that working in a team to achieve the solutions is not allowed in this course and, if suspected to occur, can result in significant penalties for the involved students.