Foundations of Machine Learning

M.Sc. in Data Sciences and Business Analytics
CentraleSupélec, Fall 2024

Instructor: Fragkiskos Malliaros
Email: fragkiskos.malliaros@centralesupelec.fr **Date:** October 23, 2024

Project Description

One of the main goals of the course is to provide the fundamental tools and algorithms that one can use in order to analyze data from various domains and applications. To this direction, in the project of the course, you will have the opportunity to develop quantitative and qualitative skills on machine learning (ML) methods and algorithms, and to obtain practical experience working with software and tools for data science. Furthermore, those of you that are interested in research, the project will give a taste of machine learning and data science research, in general.

In particular, there are several types of projects that one can choose to work on (some of them are related to each other; also think of interesting combinations among them):

- Experimental evaluation of ML algorithms and models on an interesting dataset. For example, select a ML task that you are interested in, pick 2-3 different algorithms for this problem and do an empirical experimental evaluation (e.g., take three different classification algorithms, find a good dataset or construct your own dataset and compare the performance of the algorithms).
- A new algorithm or extension of an already existing one for a ML task, and evaluation on real/artificial data (e.g., a new clustering algorithm for time series data).
- Pick a dataset that you are interested in. Formulate a predictive or analytics task over that dataset using ML methods, propose an algorithm that solves the problem (or use/adapt an existing one) and perform an experimental evaluation. Interesting datasets from a wide range of applications can be found in the website of the course.
- A theoretical project that considers an algorithm and derives rigorous results about it.
- Efficient implementation of a ML algorithm and experimental evaluation.
- Write a literature review for a machine learning topic. Each team should read at least 5-6 related research articles and summarize them. However, the survey needs to be critical and comparative, and not just a collection of summaries of different papers. In particular, you have to present the pros and cons of each method, and also provide guidelines for when to use which method depending on the application domain, the instance of the problem, or the specific properties of the dataset. If possible (and does not require to implement all methods from scratch), the survey should include an experimental comparison of the surveyed techniques.

Note that, the above list is not exhaustive and you can come up with other interesting types of projects, which will include the following components: **dataset + ML task + model + related work + evaluation/results**. Ideally, projects will be combination of an interesting application, experimentation on real/artificial data and some theoretical analysis.

Moreover, our advice is to pick a project that you can get excited and passionate about. Do feel free to propose ambitious things that you are excited about. We will always be available to discuss with you on potential projects ideas.

In general, the projects will be evaluated based on the following:

- ***Significance:*** Is the problem "real" and "interesting", or just a "toy" problem? How original, important and well defined are the questions posed? How novel is the approach? Is the work likely to be useful and/or have impact?
- ***Technical Quality:*** Is the approach and the methods appropriate and well described? Are sufficient details provided? Is the technical material correct? Are the proposed algorithms or applications creative and interesting? Are the methods and algorithms reproducible? Is the interpretation (discussion and conclusions) well balanced and supported by the data?
- ***Organization:*** Is the final project report well organized (e.g., following the structure of a paper published in a top data mining conference - see the website of the course for examples of scientific articles)? Is the write-up clear and easy to read? Are the results presented in the most appropriate manner? Are figures and tables used appropriately?

Students should work in teams of 3-4 people (preferably 4). This number is strict and should be respected. Make sure to mention the names of all of your team members when submitting the project deliverables. There are two main deliverables for the project.

Deliverables	**Due Date**
Proposal	**November 3**
Final Report + Source code	**December 20**

All deadlines are at 23:00 (unless stated otherwise). **Please, always consult the website of the course regarding the final due dates (announcements will also be made on Edunao).**

Deliverable 1: Project Proposal

In the proposal, you have to briefly but concisely describe your project: you have to clearly define the problem your project proposes to solve and how you plan to achieve it. In general, from the proposal, one should easily identify the following:

Given <dataset X *or* crawled data Y *or* …>
Use <technique(s)>
To <achieve "XYZ">

E.g., **Given** Twitter data **Use** Clustering algorithms **To** detect groups of users

The proposal should be consisting of the following parts:

- **Motivation and Problem Definition:** What is the problem you are trying to solve? Why the problem is important? What are a few potential applications? What is the main related work for the problem?
- **Methodology:** How do you plan to address the problem? What are the steps you need to take? Which algorithms/models/tools do you plan to use, develop or extend? How your methodology is related to prior work for this problem?

- **Evaluation:** How will you evaluate your work? What experiments do you plan to do? What dataset will be used (existing datasets, or are you planning to create new ones)?
- **References:** The proposal should include a preliminary list of at least 4-5 related papers that have been published in conferences/journals.

Formatting and Page Limits: The suggested length of the proposal is **3 pages** and should be in PDF format. Please include in the header of the report the title of your project and the names of the members of the team. Also, indicate that this is the project proposal (adding "Project Proposal" as subtitle). All reports should be formatted according to the ICLR template (available in Overleaf):
https://www.overleaf.com/latex/templates/template-for-iclr-2021-conference-submission/mmpfhsxmqdkp

How to submit? Please, submit the PDF file in **gradescope** (Project Proposal) as group submission (Entry Code: **R73RYE**).

Deliverable 2: Project Final Report

The final report should represent all the completed work, having the following structure:

- **Abstract:** Short (200-250 words) abstract of your project.
- **Introduction/Motivation:** What is the project about? What is the problem you are trying to solve? What are the questions you want to answer? Why the problem is important? What are a few potential applications?
- **Problem Definition:** Introduce notation, provide formal definitions as needed, define any constraints or restrictions, define what you try to optimize (e.g., maximize or minimize an optimization function, or an accuracy/error function). Describe the problem in a formal way.
- **Related Work:** Position the problem among the body of existing research. How does your project relate to previous research? How is your project replicating/different/complementary to previous research? For example, what are the state-of-the-art methods currently employed to deal with this problem? References to papers you cite should be explicit followed by a comment that describes how it is relevant.
- **Methodology:** How did you address the problem? What are the steps you had to take? Describe the data collection process. How did you process the data? Provide any mathematical background necessary for the methods. Describe any algorithms or variations of the methods. Did you have any issues due to scalability, overfitting, etc.? Describe limitations or difficulties with your approach. Formally describe any important algorithms used from the literature. Try to be as specific as possible.
- **Evaluation:** How did you evaluate your work? What experiments did you run? Describe clearly your findings.
- **Conclusions:** What are the conclusions of your work? Are there any highlights? What are some ideas for future work?
- **References:** In the final report, you should provide full list of references.

Note: Keep in mind however, that if there is a good reason why your project doesn't match the above description, we will take that into consideration when grading your report. For example, we recognize that purely theoretical or pure data analysis projects may not fit the rubric above perfectly.

The evaluation of the final report will be based on the following guidelines:

Outline	Weight
Introduction/Motivation/Problem Definition	15%
Related Work	10%
Model/Methodology/Algorithm	20%
Evaluation/Results	30%
Style and writing	15%
Code of the project	10%

Formatting and Page Limits: The suggested length of the final report is **8-9 pages** and should be in PDF format (same template as above). Please **include all team member names** (as authors) and indicate that this is the final project report (adding "Final Project Report" as subtitle).

How to submit? Please, submit the PDF file of your report in **gradescope** (Project Final Report). It is mandatory to include all team members in both the pdf file and in gradescope. **Also**, please send the code and data (in case you have not used publicly available datasets) as well as the recording of your presentation by email to **centralesupelec.fmlclass@gmail.com** (or preferably, the link to a repository where you have uploaded the code/data).

Resources

Please, check the **Resources** section of the website for datasets, software and other material that might help you in the project. We are also very happy to discuss with you any aspect of the project :-)

Acknowledgments: Ideas for the project reports were borrowed from J. Leskovec (Stanford University), M. Papagelis (York University), and Julian McAuley (UC San Diego).