Projects
Machine Learning Course
Fall 2023

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion & Martin Jaggi
www.epfl.ch/labs/mlo/machine-learning-cs-433

Course Overview - CS-433

Machine learning and data analysis are becoming increasingly central in many sciences and applications. In this course, fundamental principles and methods of machine learning will be introduced, analyzed and practically implemented.

Course Goals

Define the following basic ML concepts and explain the main differences between them: Regression, classification, clustering, dimensionality reduction.

Implement and apply machine learning methods to real-world problems. Rigorously evaluate the performance of ML methods using cross-validation.

Optimize the main trade-offs such as overfitting, and computational cost vs accuracy.

Experience common pitfalls and how to overcome them, also in an interdisciplinary context (ML4science).

Explain and understand fundamental theory concepts presented for ML methods.

Syllabus

We will cover the following ML methods and concepts:

- Basic regression and classification concepts and methods:
  Linear models, overfitting, linear regression, Ridge regression, logistic regression, SVMs, and k-NN
- Fundamental concepts:
  Cost-functions and optimization, cross-validation and bias-variance trade-off, curse of dimensionality, kernel methods
- Neural networks:
  Basics, representation power, backpropagation, CNNs, transformer models, regularization, data augmentation, dropout, adversarial examples and robustness
- Unsupervised learning:
  k-Means Clustering, Gaussian mixture models, the EM algorithm, Generative models, self-supervised learning.
- Representation learning and dimensionality reduction:
  PCA, matrix factorizations, word embeddings for natural language processing, recommender systems

The syllabus provided on the website is more precise but subject to change.

Exercise sessions

Weekly every Thursday 14:15 - 16:00, in person in the following rooms. **Room assignment by lastname**:

INF1: (A-Co), INF119: (Cr-Ha), INJ218: (He-Mi), INM202: (Mo-Sh), INR219: (Si-Z).

All labs and projects will be in *Python*. See the first lab to get started. Don't worry if you have no experience in it yet, but in that case you should take enough time to thoroughly work on the first (and second) lab.

Prerequisites

We will revise some of basic ML concepts in the first and second weeks of the course. However, you are recommended to go through the list of pre-requisites here to make sure your knowledge is up to date.

Vector and Matrix Algebra. Vector and matrix multiplication, matrix inverse, rank, eigenvalue decomposition. Refer to first year courses, or the linear algebra handout on the website, or Gilbert Strang's book for example.

Vector and Matrix Calculus. Important: The definition of derivative with respect to vectors and matrices. For reference, see this blogpost explained.ai/matrix-calculus, or the Matrix Cookbook for example.

Scientific Computing Languages. Python Basics (see tutorial in the first lab).

Probability and Statistics. Conditional and joint distribution, independence, Bayes' rule, random variable and expectation, law of large numbers.

Gaussian Distribution. Univariate and multivariate, conditional, joint and marginals.

Writing Scientific Documents using Latex (not required but preferred). Many tutorials are available online, and we provide more resources when we come to Project 1.

Resources

Course Webpage

All materials will be made available on our public github repository, including annotated lecture notes, code and exercise solutions:

github.com/epfml/ML_course

Most materials will also be linked on the course website:

www.epfl.ch/labs/mlo/machine-learning-cs-433

Lecture Notes

PDF notes for each lecture will be available on the website (and github) before the day of the lecture, and will often be annotated during the lecture. For revisions in case of errors, see also on github.

Recommended Textbooks

No book is mandatory for this course. Nevertheless, the following examples contain parts relevant to the course:

- G. Strang: Linear Algebra and Learning from Data
- S. Shalev-Shwartz and S. Ben-David: Understanding Machine Learning - From Theory to Algorithms
- G. James, D. Witten, T. Hastie and R. Tibshirani: An introduction to statistical learning
- T. Hastie, R. Tibshirani and J. Friedman: Elements of statistical learning
- C. Bishop: Pattern Recognition and Machine Learning
- K. Murphy: Machine Learning: A Probabilistic Perspective
- Michael Nielsen: Neural Networks and Deep Learning

Assessment and Practical Projects

- Project 1 (10%), due Oct 30^{th}
- Project 2 (30%), due Dec 21^{st}
- Final exam (60%)

Project 1 (10%)

The goal of this project is to help you prepare for Project 2.

In this first project, you will work in a small group of 3 students (2 only in exceptional cases, pending approval).

You will implement the most important methods covered in the lectures and labs so far.

Additionally, we will provide you with an interesting real-world dataset, and organize our own competition here:

`https://www.aicrowd.com/challenges/epfl-machine-learning-higgs`

A detailed project description will be posted on the website very soon.

Team assignment: Your choice. We recommend working in interdisciplinary teams, since the projects require many aspects. Use the discussion forum to find team-mates and form a group of 3.

You will also submit your Python code, and a 2 page PDF report. *Deadline: Oct 30th.*

Project 2 (30%)

Project 2 is the final project and gives you more freedom and responsibilities.

Again, you will work in a group of 3 people.

You can freely choose between three options:

A) **Machine Learning for Science**: Pick a real-world challenge offered by any research group of the EPFL campus, or any academic institution in Switzerland.

`https://www.epfl.ch/labs/mlo/ml4science/`

A list of potential project ideas will be made available to students later, and is subject to availability. It is also possible that you reach out directly to some research groups early in the semester, and ask for a potential project idea. Labs who are interest to host a student group can contact us here to offer their project idea, dataset or task. Projects ideas need to be approved by the group in question and by us, early November.

B) Pick one of two **pre-defined challenges** with real-word data problems:

`https://www.aicrowd.com/challenges/epfl-ml-text-classification`
`https://www.aicrowd.com/challenges/epfl-ml-road-segmentation`

Submitting your predictions to the CrowdAI platform allows you to get immediate feedback on your performance.

C) Your team participates in the **Reproducibility Challenge**. Here the goal is to select a recently submitted paper, and try to reproduce (parts of) its experiments:

`https://paperswithcode.com/rc2022`
(the papers can be from any conference including NeurIPS, ICML, ICLR, ACL-IJCNLP, EMNLP, CVPR, ICCV, AAAI and IJCAI)

In cases A) and B), you will submit your project as Python code, and a 4 page PDF report. For C), see separate instructions on the webpage.

Deadline for all cases: Dec 21nd.

Final exam (60%)

A very standard final exam.

It will contain questions on what you have learned during the lectures and exercise sessions.

We will give you a sample exam before for you to practice.

You are allowed to bring one cheat sheet (A4 size paper, both sides can be used).

No calculator, No collaborations. No cell phones. No laptops etc.

Contact Us

Please use the discussion forum for any questions and feedback on the course material or exercises, or email the respective assistants and teachers.

Teaching Assistants:

Lara Orlandic (Organizing TA)

Maksym Andriushchenko
Francesco D'Angelo
Corentin Dumery
Dongyang Fan
Simin Fan
Hojjat Karami
Atli Kosson
Skander Moalla
Hristo Papazov
Aditya Varre
Oguz Yüksel

Erwan Emlil
Vinko Sabolcec
Mikhail Seliugin
Guanyu Zhang
Yauheniya Karelskaya
Alejandro Hernandez Cano
Naisong Zhou
Aybars Yazici
Philippe Servant
Yiyang Feng
Leonardo Trentini
Simon Halstensen
Mathis Randl
Mohamed Hadhri
Dong Chu

The assistants will be helping you during the exercise sessions and projects.

Credits

Teaching material by Emtiyaz Khan, Rüdiger Urbanke, Martin Jaggi, Nicolas Flammarion, Tatjana Chavdarova. Additional material and code by Tao Lin, Frederik Kunstner, Yannic Kilcher, Aurelien Lucchi, Thijs Vogels, the assistants from the previous years, and others...