Projects
Machine Learning Course
Fall 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
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 problems and explain the main differences between them: Regression, classification, clustering, dimensionality reduction.

Implement and apply machine learning methods to real-world problems, and rigorously evaluate their performance using cross-validation.

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

Experience common pitfalls and how to overcome them.

Explain and understand the fundamental theory 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, and k-NN.
- Fundamental concepts:
  Cost-functions and optimization, cross-validation and bias-variance trade-off, curse of dimensionality.
- Unsupervised learning:
  k-Means Clustering, Gaussian mixture models and the EM algorithm.
- Dimensionality reduction:
  PCA and matrix factorization, word embeddings for natural language processing.
- Advanced methods:
  Generalized linear models, SVMs and Kernel methods, graphical models, neural networks and deep learning.

The syllabus provided on the website is more precise but subject to change, especially for the later parts of the lecture.

What not to expect

You will not learn ALL advanced methods.

You will not learn ALL the details.

This course is not specifically about big data or large-scale methods.

This is not a course about numerical optimization, neither is it about statistics. We will use both of these but focus on the fundamental techniques only.

Pre-requisites are stated below and in the coursebook. You have to learn that on your own.

This course does not teach you all that you need to know to be able to apply machine learning, but this course will get you started for sure.

Labs

Weekly every Thursday 14:15 - 16:00, in the following 5 rooms:
INF119 (A-Car), INJ218 (Cav-G), INF2 (H-Mo), INM202 (Mu-Sing), INR219 (Sinn-Z)

TODO

according to last-name.

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 the 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. For example, see Chapter 2 of Chris Bishop's book.

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

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

Resources

Course Webpage

We will use the course website for all material, except code

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

Code templates and solutions for the labs will be made available on our github repository

github.com/epfml/ML_course

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

There are many relevant books. None of them are strictly mandatory for this course. The following examples contain parts which are relevant to the course:

- 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.

The first three and the last books are available online for free. You do not have to buy any books, since we will only refer to few chapters in these. Some physical copies are available in the library.

Assessment and Practical Projects

- Project 1 (10%), due Oct 28^{th}
- Project 2 (30%), due Dec 19^{th}
- 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.kaggle.com/c/epfml18-higgs` ***TODO***

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 moodle 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 28th*.

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, who has agreed with us to offer their dataset and task. A list of potential projects will be made available later, but is subject to availability since the course has a lot of students. It is also possible that you reach out directly to some research groups now, and ask for a project idea. Projects ideas need to be approved by the group in question and by us, early november.

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

`https://www.crowdai.org/challenges/epfl-ml-road-segmentation`
`https://www.crowdai.org/challenges/epfl-ml-text-classification`
`https://www.crowdai.org/challenges/epfl-ml-recommender-system`

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

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

`https://reproducibility-challenge.github.io/iclr_2019/`
(link for 2020 is currently not available yet. this option is subject to the challenge being offered again in 2020)

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 19th.

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), either handwritten or 11 point minimum font size.

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

Contact Us

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

Teaching Assistants:

Okan Altingövde
Adam Gluch Grzegorz
Semih Günel
Prakhar Gupta
Anastasia Koloskova
Andreas Maggiori
Aswin Suresh
Thijs Vogels
Yubo Xie
Teresa Yeo

Sami Benhassen
Pei Wang
Wanhao Zhou

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

Credits

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