

# Applied Data Analysis and Machine Learning: Introduction to the course, Logistics and Practicalities

Morten Hjorth-Jensen<sup>1,2</sup>

<sup>1</sup>Department of Physics, University of Oslo

<sup>2</sup>Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

Aug 19, 2020

## Overview of first week

- Thursday August 20: First lecture: Presentation of the course, aims and content
- Thursday: Second Lecture: Start with simple linear regression and repetition of linear algebra and elements of statistics
- Friday August 21: Linear regression
- Computer lab: Wednesdays, 8am-6pm. First time: Wednesday August 26.

## Lectures and ComputerLab

- Lectures: Thursday (12.15pm-2pm and Friday (12.15pm-2pm). Due to the present COVID-19 situation all lectures will be online. They will be recorded and posted online at the official UiO [website](#).
- Weekly reading assignments and videos needed to solve projects and exercises.
- Weekly exercises when not working on projects. You can hand in exercises if you want.

- Detailed lecture notes, exercises, all programs presented, projects etc can be found at the homepage of the course.
- Weekly plans and all other information are on the official webpage.
- No final exam, three projects that are graded and have to be approved.

## Course Format

- Three compulsory projects. Electronic reports only using [Canvas](#) to hand in projects and [git](#) as version control software and [GitHub](#) for repository (or [GitLab](#)) of all your material.
- Evaluation and grading: The three projects are graded and each counts 1/3 of the final mark. No final written or oral exam.
  1. For the last project each group/participant submits a proposal or works with suggested (by us) proposals for the project.
  2. If possible, we would like to organize the last project as a workshop where each group makes a poster and presents this to all other participants of the course
  3. Poster session where all participants can study and discuss the other proposals.
  4. Based on feedback etc, each group finalizes the report and submits for grading.
- Python is the default programming language, but feel free to use C/C++ and/or Fortran or other programming languages. All source codes discussed during the lectures can be found at the webpage and [github address](#) of the course.

## Teachers

### Teachers :

- Morten Hjorth-Jensen, [morten.hjorth-jensen@fys.uio.no](mailto:morten.hjorth-jensen@fys.uio.no)
  - **Phone:** +47-48257387
  - **Office:** Department of Physics, University of Oslo, Eastern wing, room FØ470
  - **Office hours:** *Anytime!* In Fall Semester 2020 (FS20), as a rule of thumb office hours are planned via computer or telephone. Individual or group office hours will be performed via zoom. Feel free to send an email for planning. In person meetings may also be possible if allowed by the University of Oslo's COVID-19 instructions.

- Øyvind Sigmundson Schøyen, oyvinssc@student.matnat.uio.no
  - **Office:** Department of Physics, University of Oslo, Eastern wing, room FØ452
- Michael Bitney, m.s.bitney@fys.uio.no
- Kristian Wold, kriswold@student.matnat.uio.no
- Nicolai Haug, nicoha@student.matnat.uio.no
- Per-Dimitri Sønsteland, perdimitri.bs@gmail.com

## Deadlines for projects (tentative)

1. Project 1: September 28 (graded with feedback)
2. Project 2: November 2 (graded with feedback)
3. Project 3: December 7 (graded with feedback)

Projects are handed in using **Canvas**. We use Github as repository for codes, benchmark calculations etc. Comments and feedback on projects only via **Canvas**.

## Recommended textbooks

- Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, The Elements of Statistical Learning, Springer
- Aurelien Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition

## Prerequisites

Basic knowledge in programming and mathematics, with an emphasis on linear algebra. Knowledge of Python or/and C++ as programming languages is strongly recommended and experience with Jupiter notebook is recommended. Required courses are the equivalents to the University of Oslo mathematics courses MAT1100, MAT1110, MAT1120 and at least one of the corresponding computing and programming courses INF1000/INF1110 or MAT-INF1100/MAT-INF1100L/BIOS1100/KJM-INF1100. Most universities offer nowadays a basic programming course (often compulsory) where Python is the recurring programming language.

## Learning outcomes

This course aims at giving you insights and knowledge about many of the central algorithms used in Data Analysis and Machine Learning. The course is project based and through various numerical projects, normally three, you will be exposed to fundamental research problems in these fields, with the aim to reproduce state of the art scientific results. Both supervised and unsupervised methods will be covered. The emphasis is on a frequentist approach, although we will try to link it with a Bayesian approach as well. You will learn to develop and structure large codes for studying different cases where Machine Learning is applied to, get acquainted with computing facilities and learn to handle large scientific projects. A good scientific and ethical conduct is emphasized throughout the course. More specifically, after this course you will

- Learn about basic data analysis, statistical analysis, Bayesian statistics, Monte Carlo sampling, data optimization and machine learning;
- Be capable of extending the acquired knowledge to other systems and cases;
- Have an understanding of central algorithms used in data analysis and machine learning;
- Understand linear methods for regression and classification, from ordinary least squares, via Lasso and Ridge to Logistic regression;
- Learn about neural networks and deep learning methods for supervised and unsupervised learning. Emphasis on feed forward neural networks, convolutional and recurrent neural networks;
- Learn about decision trees, random forests, bagging and boosting methods;
- Learn about support vector machines and kernel transformations;
- Reduction of data sets, from PCA to clustering;
- Autoencoders and Reinforcement Learning;
- Work on numerical projects to illustrate the theory. The projects play a central role and you are expected to know modern programming languages like Python or C++ and/or Fortran (Fortran2003 or later).

## Topics covered in this course: Statistical analysis and optimization of data

The course has two central parts

1. Statistical analysis and optimization of data

## 2. Machine learning

These topics will be scattered throughout the course and may not necessarily be taught separately. Rather, we will often take an approach (during the lectures and project/exercise sessions) where say elements from statistical data analysis are mixed with specific Machine Learning algorithms

**Statistical analysis and optimization of data.** The following topics will be covered

- Basic concepts, expectation values, variance, covariance, correlation functions and errors;
- Simpler models, binomial distribution, the Poisson distribution, simple and multivariate normal distributions;
- Central elements of Bayesian statistics and modeling;
- Gradient methods for data optimization,
- Monte Carlo methods, Markov chains, Gibbs sampling and Metropolis-Hastings sampling;
- Estimation of errors and resampling techniques such as the cross-validation, blocking, bootstrapping and jackknife methods;
- Principal Component Analysis (PCA) and its mathematical foundation

## Topics covered in this course: Machine Learning

The following topics will be covered

- Linear Regression and Logistic Regression;
- Neural networks and deep learning, including convolutional and recurrent neural networks
- Decisions trees, Random Forests, Bagging and Boosting
- Support vector machines
- Bayesian linear and logistic regression
- Boltzmann Machines
- Unsupervised learning Dimensionality reduction, from PCA to cluster models

Hands-on demonstrations, exercises and projects aim at deepening your understanding of these topics.

## Extremely useful tools, strongly recommended

and discussed at the lab sessions.

- GIT for version control, and GitHub or GitLab as repositories, highly recommended. This will be discussed during the first exercise session
- Anaconda and other Python environments, see intro slides and first exercise session

## Other courses on Data science and Machine Learning at UiO

The link here <https://www.mn.uio.no/english/research/about/centre-focus/innovation/data-science/studies/> gives an excellent overview of courses on Machine learning at UiO.

1. [STK2100 Machine learning and statistical methods for prediction and classification.](#)
2. [IN3050 Introduction to Artificial Intelligence and Machine Learning.](#) Introductory course in machine learning and AI with an algorithmic approach.
3. [STK-INF3000/4000 Selected Topics in Data Science.](#) The course provides insight into selected contemporary relevant topics within Data Science.
4. [IN4080 Natural Language Processing.](#) Probabilistic and machine learning techniques applied to natural language processing.
5. [STK-IN4300 Statistical learning methods in Data Science.](#) An advanced introduction to statistical and machine learning. For students with a good mathematics and statistics background.
6. [INF4490 Biologically Inspired Computing.](#) An introduction to self-adapting methods also called artificial intelligence or machine learning.
7. [IN-STK5000 Adaptive Methods for Data-Based Decision Making.](#) Methods for adaptive collection and processing of data based on machine learning techniques.
8. [IN5400/INF5860 Machine Learning for Image Analysis.](#) An introduction to deep learning with particular emphasis on applications within Image analysis, but useful for other application areas too.
9. [TEK5040 Deep learning for autonomous systems.](#) The course addresses advanced algorithms and architectures for deep learning with neural networks. The course provides an introduction to how deep-learning techniques can be used in the construction of key parts of advanced autonomous systems that exist in physical environments and cyber environments.

10. STK4051 Computational Statistics
11. STK4021 Applied Bayesian Analysis and Numerical Methods