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

Morten Hjorth-Jensen[1,2]

[1]Department of Physics, University of Oslo

[2]Department of Physics and Astronomy and National Superconducting Cyclotron Laboratory, Michigan State University

Jul 30, 2019

Overview of first week

- Thursday: First lecture: Presentation of the course, aims and content
- Thursday: Second Lecture: Start with simple linear regression and repetition of linear algebra
- Friday: Linear regression
- Computer lab: Tuesday. First time: Tuesday August 27.

Lectures and ComputerLab

- Lectures: Thursday (12.15pm-2pm) and Friday (12.15pm-2pm).
- Weekly reading assignments needed to solve projects and exercises.
- Weekly exercises when not working on projects. You can hand in exercises if you want.
- First hour of each lab session may be used to discuss technicalities, address questions etc linked with projects and exercises.
- Detailed lecture notes, exercises, all programs presented, projects etc can be found at the homepage of the course.

- Computerlab: Tuesday (8am-4pm), VB IT-auditorium 3. Depending on how many enlist we may extend the lab sessions
- 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 devilry to hand in projects and Git for repository and 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 programmin languages. All source codes discussed during the lectures can be found at the webpage and github address of the course.

Teachers and ComputerLab

Teachers :

1. Hanna Svennevik
2. Morten Hjorth-Jensen
3. Lucas Charpentier
4. Stian Bilek

day	Time
Group 1: Tuesday	8am-10am
Group 2: Tuesday	10am-12pm
Group 3: Tuesday	12pm-2pm
Group 4: Tuesday	2pm-4pm

Deadlines for projects (tentative)

1. Project 1: September 30 (graded with feedback)
2. Project 2: November 4 (graded with feedback)
3. Project 3: December 2 (graded with feedback)

Projects are handed in using devilry.ifi.uio.no. We use Github as repository for codes, benchmark calculations etc. Comments and feedback on projects only via devilry.

Learning outcomes

- 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
- Gain knowledge of central aspects of Monte Carlo methods, Markov chains, Gibbs samplers and their possible applications
- Understand linear methods for regression and classification, from ordinary least squares, via Lasso and Ridge to Logistic regression
- Learn about various neural networks and deep learning methods for supervised and unsupervised learning
- Learn about about decision trees and random forests
- Learn about support vector machines and kernel transformations
- Reduction of data sets, from PCA to clustering, supervised and unsupervided methods
- 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++

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

- 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, Metropolis-Hastings algorithm
- Linear methods for regression and classification
- Estimation of errors using cross-validation, blocking, bootstrapping and jackknife methods
- Practical optimization using Singular-value decomposition and least squares for parameterizing data

Topics covered in this course: Machine Learning

The following topics will be covered

- Linear Regression and Logistic Regression
- Neural networks and deep learning
- Decisions trees and nearest neighbor algorithms
- Support vector machines
- Bayesian Neural Networks
- Boltzmann Machines
- Dimensionality reduction, from PCA to cluster models

Extremely useful tools, strongly recommended

and discussed at the lab sessions.

- GIT for version control, highly recommended
- Devilry for handing in projects, next week
- Anaconda and other Python environments, see intro slides

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