Applied Data Analysis and Machine Learning: Introduction to the course

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

Aug 23, 2018

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: Wednesday. First time: Wednesday August 29.

Reading suggestions and exercises

The recommended textbooks

- HTF: Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, The Elements of Statistical Learning, Springer
- AG: Aurelien Geron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, O'Reilly

- HTF chapters 1-3 and lecture notes
- AG chapters 1 and 2

Lectures and ComputerLab

- Lectures: Thursday (12.15pm-2pm) and Friday (12.15pm-2pm).
- Weekly reading assignments needed to solve projects.
- First hour of each lab session may be used to discuss technicalities, address questions etc linked with projects.
- Detailed lecture notes, exercises, all programs presented, projects etc can be found at the homepage of the course.
- Computerlab: Wednesday (10am-6pm), room FV329, four groups (10am-12pm, 12pm-2pm, 2pm-4pm, 4pm-6pm).
- 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.
- The last project is organized as a workshop (duration approx 5-6 hours toward the end of the semester) where each group submits a proposal for a data sets to be analyzed with the different methods discussed during the course.
    1. Each group submits a proposal or works with suggested (by us) proposals for the project.
    2. 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 (the three best posters get additional score)
    4. Based on feedback etc, each group finalizes the report and submits for grading.
- The computer lab (room FV329)consists of 16 Linux PCs, but many prefer own laptops. Python is the default programming language, but feel free to use C/C++ and/or Fortran. All source codes discussed during the lectures can be found at the webpage and github address of the course.

Teachers and ComputerLab

Teachers :

1. Kristine B. Heine
2. Morten Hjorth-Jensen
3. Bendik Samseth
4. Øyvind Sigmundson Schøyen

day	Time
Group 1: Wednesday	10am-12pm
Group 2: Wednesday	12pm-2pm
Group 3: Wednesday	2pm-4pm
Group 4: Wednesday	4pm-6pm

Deadlines for projects (end of day)

1. Project 1: October 1 (graded with feedback)
2. Project 2: November 5 (graded with feedback)
3. Project 3: November 30, tentative (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

The course introduces a variety of central algorithms and methods essential for studies of data analysis and machine learning. The course is project based and through the various 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. You will learn to develop and structure large codes for studying these systems, 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, Bayesian statistics, Monte Carlo methods, 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, from numerical integration to simulation of stock markets;
- Understand linear methods for regression and classification;
- Learn about neural network, genetic algorithms and Boltzmann machines;
- 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

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;
- Central elements from linear algebra
- Cubic splines and gradient methods for data optimization
- Monte Carlo methods, Markov chains, Metropolis-Hastings algorithm, ergodicity;
- Linear methods for regression and classification;
- Estimation of errors using blocking, bootstrapping and jackknife methods;

Topics covered in this course: Machine Learning

- Linear and non-linear regression
- Gaussian and Dirichlet processes;
- Boltzmann machines;
- Neural networks;
- Decisions trees and nearest neighbor algorithms
- Support vector machines

Extremely useful tools, strongly recommended

and discussed at the lab sessions.

- GIT for version control (see webpage)
- ipython/jupyter notebook
- Devilry for handing in projects, next week
- Anaconda and other Python environments