Data Analysis and Machine Learning with Numerical Projects

Department of Geoscience, Department of Mathematics and Department of Physics, University of Oslo

Planned start: Fall semester 2018

Data analysis and machine learning: a new 10-ECTS course for both CS students and others

Course content. Probability theory and statistical methods play a central role in science. Nowadays we are surrounded by huge amounts of data. For example, there are about one trillion web pages; more than one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day; the genomes of 1000s of people, each of which has a length of 3.8×10^9 base pairs, have been sequenced by various labs and so on. This deluge of data calls for automated methods of data analysis, which is exactly what machine learning provides. In this course the approach is to define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. Since many of these problems can be studied using tools of probability theory, the aim of this course is to expose you to central methods in probability theory linked with machine learning.

This course covers thus topics like Monte Carlo methods and Markov chains, Bayesian statistics, error estimates, various linear methods, optimization of data and error analysis and central algorithms in machine learning. The course has several numerical projects and numerical exercises that are meant to illustrate the theory.

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, the students will be exposed to fundamental research problems in these fields, with the aim to reproduce state of the art scientific results. The students 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 basis 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;
- Have a basic knowledge of Bayesian statistics and learning and common distributions;
- 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 students are expected to know modern programming languages like Python or C++.

Prerequisites

Basic knowledge in programming and numerics. 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-INF1xxx.

Overlapping courses (?). To be determined.

The course has two central parts

1. Statistical analysis and optimization of data
2. Machine learning

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;
- Monte Carlo methods, Markov chains, Metropolis-Hastings algorithm, ergodicity;
- Linear methods for regression and classification;
- Estimation of errors using blocking, bootstrapping and jackknife methods;
- Practical optimization using Singular-value decomposition and least squares for parameterizing data.

Machine learning. The following topics will be covered

- Gaussian and Dirichlet processes;
- Boltzmann machines;
- Neural networks;
- Genetic algorithms.

All the above topics will be supported by examples, hands-on exercises and project work.

Computational aspects play a central role and the students are expected to work on numerical examples and projects which illustrate the theory and methods. Some of the projects can be coordinated with the high-performance programming course (course code to be added).

Practicalities

1. Four lectures per week, Fall semester, 10 ECTS;
2. Four hours of laboratory sessions for work on computational projects;
3. Three projects which are graded and count 60% of the final grade;
4. A selected number of weekly assignments which count 10% of the final grade;
5. Final written exam which counts 30% of the final grade;

6. Organized by the Departments of Geoscience, Mathematics and Physics;

7. Possible teachers first time: John Burkhart, Morten Hjorth-Jensen and XXX;

8. The course is part of the CS Master of Science program, but is open to other bachelor and Master of Science students at the University of Oslo;

9. Grading scale: Grades are awarded on a scale from A t* F, where A is the best grade and F is a fail;

10. The course will be offered as a 4XXX and 3XXX course.

Possible textbooks

General learning book on statistical analysis:

- Christian Robert and George Casella, Monte Carlo Statistical Methods, Springer
- Peter Hoff, A first course in Bayesian statistical models, Springer

General Machine Learning Books:

- Kevin Murphy, Machine Learning: A Probabilistic Perspective, MIT Press
- Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer
- David J.C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning, Springer
- David Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press