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

Jan 29, 2019

Teachers

Teachers :

1. Kristine B. Heine
2. Morten Hjorth-Jensen
3. Bendik Samseth

Lectures and ComputerLab

- Time: January 21-31, 2019.
- Place: GANIL, Caen

Organization of the day.

1. Lectures 9am-12pm
2. Lunch 12pm-2pm
3. Computerlab 2pm-6pm

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, 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 machine learning;
- Gain knowledge of central aspects of Monte Carlo methods, Markov chains, and their possible applications;
- 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 in statistics, expectation values, variance, covariance, correlation functions and errors;
- Simpler models, binomial distribution, the Poisson distribution, simple and multivariate normal distributions;
- Central elements from linear algebra
- 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;
- Principal Component Analysis and Clustering algorithms;

Topics covered in this course: Machine Learning

- Linear and non-linear regression
- Boltzmann machines;
- Neural networks;
- Decisions trees and random forests
- Support vector machines

Extremely useful tools, strongly recommended

Discussed at the lab sessions.

- GIT for version control
- ipython/jupyter notebook
- Anaconda and other Python environments

Course Content and detailed plan

Lectures are approximately 45 min each with a small break between each lecture. There is also a coffee break of 30 min in the morning sessions. It will most likely be scheduled around 1045am and is not marked in the program below. Lunch is from 12pm to 2pm. The lab sessions start at 2pm and end at 6pm. The acronyms are

- BS: Bendik Samseth
- KBH: Kristine B. Hein
- MHJ: Morten Hjorth-Jensen

Week 1, January 21-26.

Day		Lecture Topics and lecturer
Monday 21	9am-945am	Introduction and welcome (MHJ)
	10am-1045am	Review of Python and Linear Algebra (MHJ)
	1045am-1115am	Break
	1115am-12pm	Getting started with Linear Regression (MHJ)
	12pm-2pm	Lunch +own activities
	2pm-6pm	Python installations and setups, anaconda and more
Tuesday 22	9am-945am	Linear Regression (MHJ)
	10am-1045am	Linear Regression, Lasso and Ridge (MHJ)
	1045am-1115am	Break
	1115am-12pm	Linear Regression, Lasso and Ridge (MHJ)
	12pm-2pm	Lunch +own activities
	2pm-6pm	Brief intro to **scikit-learn**
Wednesday 23	9am-945am	Summary of linear regression (MHJ)
	10am-1045am	Statistical analysis of data, bias and variance (MHJ)
	1045am-1115am	Break
	1115am-12pm	Statistical analysis of data, bias and variance (MHJ)
	12pm-2pm	Lunch +own activities
	2pm-6pm	More **scikit-learn** functionality
Thursday 24	9am-945am	Statistical analysis, cross-validation and Bootstrap (MHJ)
	10am-1045am	Statistical analysis, cross-validation and Bootstrap (MHJ)
	1045am-1115am	Break
	1115am-12pm	Optimization and gradient descent (MHJ)
	12pm-2pm	Lunch +own activities
	2pm-6pm	Statistics tools
Friday 25	9am-945am	Optimization and Gradient descent (MHJ)
	10am-1045am	Logistic Regression (MHJ)
	1045am-1115am	Break
	1115am-12pm	Logistic regression and classification and start neural networks (MH
	12pm-2pm	Lunch +own activities
	2pm-6pm	Gradient descent coding
Saturday 26	9am-945am	Neural Networks (MHJ)
	10am-1045am	Neural Networks and the back propagation algo (MHJ)
	1045am-1115am	Break
	1115am-12pm	Neural Networks and the Back propagation algo (MHJ)
	12pm-2pm	Lunch +own activities
	2pm-3pm	Only discussion of project 1

Week 2, January 28-31.

Day		Lecture Topics and lecturer
Monday 28	9am-945am	Neural networks, setting up code for back propagation (M
	10am-1045am	Neural networks,examples (MHJ)
	1045am-1115am	Break
	1115am-12pm	Neural networks, examples and convolutional NN, setting up your
	12pm-2pm	Lunch +own activities
	2pm-6pm	Introduction to Tensorflow and Keras examples
Tuesday 29	9am-945am	Solving differential equations with Neural Networks (KB
	10am-1045am	Boltzmann machines and the the many-body problem (M
	1045am-1115am	Break
	1115am-12pm	Boltzmann machines and the the many-body problem (M
	12pm-2pm	Lunch +own activities
	2pm-6pm	Tensorflow/Keras and CNNs
Wednesday 30	9am-945am	Decision Trees and Bagging (MHJ)
	10am-1045am	Bagging, Ensembles and Random Forests (MHJ)
	1045am-1115am	Break
	1115am-12pm	
	12pm-2pm	Lunch +own activities
	2pm-6pm	Continuation of SVM lectures and work on projects
Thursday 31	930am-1230am	Work on Projects 1 and 2
	1230pm-2pm	Lunch +own activities
	2pm-6pm	Work on projects

Teaching and projects

The course will be taught as an intensive course of duration of two weeks, with a total time of 30 h of lectures, 40 h of exercises and 2 projects that will be graded and form the final assessment. Each project counts 50% of the final grade. They have to be finalized two weeks after the course ended.

The final assignment will be graded with marks A, B, C, D, E and failed. The grades will be available not later than March 1.

The organization of a typical course day is as follows:

Time	Activity
9am12pm	Lectures, project relevant information and directed exercises
12pm-2pm	Lunch and own activities
2pm-6pm	Computational projects, exercises and hands-on sessions
6pm-7pm	Wrap-up of the day and eventual student presentations and/or further discussions

Suggested Literature

Possible 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