Master Degree in Physics of Complex Systems

INTRODUCTION TO DATA SCIENCE (6 ECTS)

(By S. Goldt, A. Laio, J. Barbier, R. Trotta)

COURSE DESCRIPTION: This short course will give an introduction to data science for physicists. The course will introduce the basics of data science problems, and introduce students to Bayesian techniques for inference, unsupervised learning methods for exploratory data analysis, and neural networks. On the theoretical side, students will learn fundamentals of random matrix theory and of the theory of neural networks.

EXPECTED LEARNING OUTCOMES: After this course, students will

1. Understand the fundamental ingredients of a data science problems
2. Be able to apply the basics of Bayesian inference problems to inference problems
3. Know the basics of random matrix theory, including the Wigner and Wishart ensembles as well as their spectra
4. Know basic unsupervised learning techniques
5. Know the basics of neural network architectures and neural learning dynamics in supervised learning.

PRE-REQUIREMENTS: Pre-requisites are probability, linear algebra and analysis at the level obtained after successful completion of a Bachelor degree in Physics, Mathematics, Mathematical Engineering, Computer science, or a similar course. The course does not assume any knowledge of data science.

COURSE TOPICS

I Basics of Bayesian inference [R. Trotta]

Introduction to inference: what is it? Why do we need it?
Frequentist probability vs bayesian probability
Confidence levels, posterior distributions, priors and the difference between all those
If time allows: classical hypothesis testing vs Bayesian model comparison

II An introduction to random matrix theory [J. Barbier]

The Wigner ensemble and the semi-circular law
Wishart Ensemble and Marchenko–Pastur law
Spiked matrix models and the BBP transition

III Introduction to unsupervised learning and dimensional reduction [A. Laio]

Principal Component Analysis
Multidimensional scaling and kernel methods
Intrinsic dimension estimates

IV Neural networks [S. Goldt]

Neural networks 101: types and applications + some open theoretical

problems
Learning dynamics: empirical phenomena and theoretical predictions
The impact of data structure
Unsupervised learning: from Hebbian learning to independent components

COURSE STRUCTURE: This is an entirely taught course.

READING MATERIALS: As an introduction to the course, chapters 1-3, 20-21, 38-39, 41, and 44 of the book “Information Theory, Inference, and Learning Algorithms” by David MacKay are recommended reading.

STUDY MATERIALS: A set of papers and Python notebooks accompany the course.

ASSESSMENT AND GRADING CRITERIA: The course will be assessed by a written exam of two hours, covering all four topics of the course.