Topics in Data Science

2023-2024 Academic Year
Master of Research in Economics, Finance and Management

1. Description of the subject

- Topics in Data Science — Code: 32887
- Total credits: 3 ECTS — Workload: 75 hours
  Term: 1st
- Type of subject: Elective
- Department of Economics and Business
- Teaching team: David Rossell

2. Teaching guide

- **Introduction**

This course builds on the contents of the Research Seminar in Data Science. The course follows up by overviewing flexible models (semi-parametric regression, tree-based, deep learning) and latent variable models (e.g. text data analysis models) and discussing specific applications, such as principles of causal inference with many covariates.

The motivation for the research seminar and this course is common. Statistical and Machine learning techniques are having a deep effect on many disciplines, including Economics. Due to the increasing number of applications using data science with techniques popular in Economics many World-leading institutions incorporated data science in their PhD programs. As examples see the talk by Novel prize Chris Sims, the set of lectures in the NBER Summer Institute by Guido Imbens and Susan Athey (in particular lecture 3, www.nber.org/econometrics_minicourse_2015/), the Big Data and Machine Learning syllabus at the Harvard Economics PhD program, the Chicago Booth PhD courses on Bayesian inference, Big Data and Machine Learning, or the Bocconi PhD courses in text analysis, Econometrics of networks and causal analysis.

Data analysis methods are evolving to cope with increasingly challenging problems, a case in point being high-dimensional situations where one considers a large number of parameters or models. For instance, one may have a regression or factor model where the number of covariates far exceeds the sample size, or may want to simplify the interpretation of complex data via clustering, text or latent variable analysis, or may want to predict an outcome using flexible algorithms. Engaging effectively in such research, either from a methodological or applied perspective, requires one to understand, and when needed modify or extend, such methodology. Just as importantly, they need to communicate such ideas effectively to a potentially non-expert audience.

The goal of this course is to build on the principles learned in the Data Science Research Seminar, i.e. regression with many parametes and the Bayesian paradigm, to consider increasingly complex problems. These are related to capturing non-linearities in the data to explain or forecast complex phenomena more accurately, dealing with latent variable models that are useful for clustering, modeling time series dependence or building non-parametric models, and discussing certain specific applications of interest (e.g. causal inference or text analysis basics). The learning outcomes are an improved familiarity with selected research topics in Statistics that are relevant for Data Science, at a level sufficient to critically appraise, modify and apply novel methods, and improved oral and written presentation skills. The course also intends to provide students with applied data analysis skills useful for their MRes thesis and subsequent research work.

Pre-requisites: the course is designed for MRes students who are familiar with the contents of the Research seminar in Data Science.

- **Contents**

1. Beyond linear regression. We will extend the earlier strategies to settings where one considers certain forms of causal inference, simple time series models, generalized linear models, flexible models for count data, or capturing non-linear relationships via generalized additive models (GAMs), random forests, Bayesian additive regression trees, or deep learning.
2. Mixture and flexible models. We shall discuss the use of latent variable to build more flexible models that can account for the presence of clusters/sub-populations. We will place some attention to mixture models and some non-parametric methods. We shall discuss applications to mixture-of-regressions models to account for unobserved confounders that may bias inference, to hidden Markov models for time series and to text data analysis. Time allowing, we may overview some Bayesian non-parametric methods.

- **Teaching methodology**

The course will be delivered in a combination of regular lectures, computer-based seminars where students get hands-on experience with the taught data analysis methods, and presentations by students (on published manuscripts chosen by the students and on the final project). The idea is to give students flexibility to deepen their understanding on the topics they find more interesting / relevant for their own research.

- **Assessment and Grading System**

Students will be asked to orally present 1 research paper of their choice (40% of the final mark), some selected exercises from the seminar sessions (10% of final mark) and a written report (50% of the final mark). This project will be decided by the students but must be pre-approved by the lecturer, and should involve the application, critical assessment or extension of the research methods seen in class. The content can be theoretical, empirical, a practical application or a combination of the former. Students who also take the Research seminar in Data Science can submit a single joint project for both courses, specific instructions on the format will be given in class.

References

The books below provide a good introduction to a substantial part of the topics covered in this course (and many others), however we shall complement them with a number of additional selected research manuscripts.

- Andrew Gelman, John B. Carlin, Hals S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin. Bayesian Data Analysis (3rd edition). CRC Press, 2013.

- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

- Nils Lid Hjort, Chris Holmes, Peter Müller, Stephen G. Walker. Bayesian non- parametrics. Cambridge University Press, 2010.

- Sylvia Frühwirth-Schnatter. Finite mixture and Markov switching models. Springer, 2006.