Advanced Topics in Data Science

2016-2017 Academic Year
Master of Research in Economics, Finance and Management

1. Description of the subject

- Advanced Topics in Data Science
- Total credits: 6 ECTS

Code: 32282
Workload: 150 hours
Term: 3rd

- Type of subject: Elective
- Department of Economics and Business
- Teaching team: David Rossell

2. Teaching guide

Introduction

Statistical and Machine learning techniques are having a deep effect on many disciplines, including economics. As a consequence of the increasing number of applications using a combination of data science with techniques popular in Economics (see references) many World-leading institutions have incorporated data science in their PhD programs. As examples see the set of lectures in the NBER Summer Institute of 2015 by Guido Imbens and Susan Athey (in particular lecture 3, www.nber.org/econometrics_minicourse_2015/), or the Big Data and machine learning syllabus at the Harvard Economics PhD program (https://locator.tlt.harvard.edu/course/colgsas-156429).

Statistical methods in data science are rapidly evolving to cope with increasingly challenging problems, a case in point being high-dimensional situations where one considers a large or even infinite number of parameters or models. Engaging effectively in research or cross-disciplinary collaborations requires data scientists to be able to absorb, 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 introduce students to some foundations behind these methods, with a certain emphasis on the Bayesian framework, expose them to and discuss research literature, and practice the skills needed for applying and presenting novel research. 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.

Pre-requisites: the course is designed for MRes students who are familiar with basic Statistical inference, specifically linear regression and maximum likelihood estimation. Basic R programming skills are beneficial, though examples and links to learning resources will be provided for students who are not familiar with R.

Contents

1. **Foundations of Bayesian inference. We review the basic Bayesian paradigm for statistical inference, its use for parameter estimation and prediction, and standard computational tools such as Gibbs or Metropolis-Hastings.**
2. **Foundations of variable selection in high-dimensional regression. We review the fundamental penalized likelihood and Bayesian frameworks for linear regression models with a large number of variables. We shall discuss the relative merits of current strategies such as SCAD or MCP penalties to help decouple variable selection from prediction or non-local priors and other Bayesian strategies to achieve good performance in high dimensions. The discussion will include theoretical and practical considerations, specifically including strategies to carry out computations in a scalable manner.**
3. **Mixture models and Bayesian non-parametrics. We shall move towards more flexible models enabled by the use of mixtures. We will consider finite and infinite mixture models, specifically for the latter we shall focus on their use as the basic building block for a broad family of Bayesian non-parametric methods, such as the Dirichlet process.**
4. **Beyond linear regression. We will discuss some strategies to extend the basic high-dimensional regression framework, such as introducing non-linearity via splines or other basis functions, the use of robust methods or mixture models to relax residual normality, or building graphical models. We shall also consider some recent results on model misspecification, that is ultimately**

any model is a mathematical abstraction and it is useful to consider how it will perform when the data are not generated by the assumed model, a currently active research topic.

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. The Elements of Statistical Learning. Data Mining, Inference and Prediction (2nd edition). Springer, 2009.**
- **Nils Lid Hjort, Chris Holmes, Peter Müller, Stephen G. Walker. Bayesian non- parametrics. Cambridge University Press, 2010.**
- **Daphne Koller, Nir Friedman. Probabilistic Graphical Models. Principles and Techniques. MIT Press, 2009.**

Some papers on data science and economics:

- **Chernozhukov, V., A. Belloni, C. Hansen, and I. Fernandez-Val (2016), "Program Evaluation with High-Dimensional Data," forthcoming in Econometrica**
- **Chernozhukov, V., A. Belloni, C. Hansen (2014), "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics", J. Economic Perspectives**
- **Chernozhukov, V., D. Chen, A. Belloni, C. Hansen (2012), "Sparse Models and Methods for Instrumental Regression, with an Application to Eminent Domain", Econometrica**
- **Einav, L. y J. Levin (2014), "Economics in the age of big data," Science, 346 (6210).**
- **Varian, H. (2014), "Big data: new tricks for econometrics," Journal of Economic Perspectives, 28 (2), 3-28.**

Teaching methodology

The course will be delivered in a combination of regular lectures, presentations of research topics given by invited experts, and presentations by students (typically critically analyzing the ideas presented in a published manuscript).

Assessment and Grading System

Students will be asked to orally present 2 research papers of their choice (40% of the final mark) and to turn in a written report for a final project (60% of the final mark). This project will be decided by the students but 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.