ECON 570 - Big Data Econometrics

Machine Learning and Causal Inference

Ida Johnsson

University of Southern California
Spring 2021 (4 Units)

Course Description

This course focuses on predictive and causal inference methods which are suitable for “big data”, both from a theoretical and applied point of view. It introduces foundational methods in Machine Learning (ML) and explores differences between a traditional ML approach and econometric approach, as well as how ML can be applied in the context of econometric modeling and causal inference. The course also covers some of the the latest developments in the field of ML for causal inference and inference with cross-sectionally correlated dynamic panel data (“spatial”/”network” data). In addition to solid theoretical foundations the course places a strong emphasis on understanding the practical implications of the theory, such as computational complexity. Lectures will include demonstrations of how the theory works in practice - using real and simulated data.

Learning Objectives

After completing this course students should be able to:

- Have a good understanding of the theory that makes fundamental ML methods work
- Know how to implement these ML methods in Python and/or R
- Understand the commonalities, differences, and synergies between ML and econometric modeling, in particular, predictive and causal inference
- Be able to choose and apply the correct inference approaches for assessing causal effects in observational and experimental data
- Choose the appropriate modeling approach for “spatial” data
- Understand the computational complexities that working with big data entails and how to choose methods that work not only in theory but also in practice
- Be familiar with the latest academic developments in econometric methods for big data and how they are used in the industry
- Keep up to date with and implement new developments in big data econometrics

Prerequisites

The necessary background for this course is calculus (at the level of MATH 226), linear algebra (at the level of MATH 225), and graduate-level econometrics (at the level of ECON 513). Formal training in causal inference and machine learning are not assumed, though some prior exposure is helpful.

Supplementary Materials

All required material will be covered by lecture notes. Lectures will also draw on recent literature in ML and econometrics, references will be provided before each lecture. For students interested in more in-depth expositions, the following textbooks are recommended.

1. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer. [HTF]
2. Bishop, C. M. (2007). Pattern Recognition and Machine Learning. Springer. [B]
3. Imbens, G. W., and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press. [IR]
4. Angrist, J. D., and Pischke, J. (2009). Mostly Harmless Econometrics. Princeton [AP]
5. Pesaran, M.H. (2015) Time Series and Panel Data Econometrics. Oxford University Press. [P]

Course Grading

2 Problem Sets: 30%

Problem sets will be given to reinforce the concepts taught in class, as well as offer an opportunity for students to code and implement the algorithms covered. Students are encouraged to work in groups, but each person must turn in their own copy. Assessment will be based on whether the right approaches were used and whether the right solutions were obtained. Due dates for the assignments are XXX and XXX.

1 Literature Assignment: 30%

For this individual assignment, students are expected to choose an academic paper that is relevant to the topics reviewed in the course. The paper can be either theoretical or empirical. Students should demonstrate their ability to understand and implement a new method by turning in a document which discusses the following:

- What class of problems does the solution presented in the paper apply to?
- What theory is it based on (for example random forest, linear regression, etc) and how do the authors extend it (in the case of a theoretical paper) or how do the authors apply it (in the case of an empirical paper)?
- Key results/assumptions/equations in the paper and what they mean
- What other problems the student might apply this method do
- Extra points: simple code implementation that demonstrates how the method works

Note that this assignment is NOT about paraphrasing the text, but about showing an understanding by explaining relevant concepts and reasoning using original language. Guidance in terms of what kind of paper to choose and suggestions will be provided in the lectures.

1 Empirical Project: 40%

For the empirical project, students are expected to work in groups (maximum of three) and apply their learnings to a real data set to tackle an empirical problem that interests them. Each group must submit a short write-up (6 pages max) that summarizes the analysis, and computer code that reproduces the quantitative results. Assessment will be based on how appropriately the quantitative tools were applied. The due date for this project is XXX.

Software

Python and R will be the main programming languages used in this course. Lecture notes will contain code snippets, TA sessions will include Python instructions, and any code in problem set solutions will be in Python. Selected topics might additionally include demonstrations in R. Students may solve the problem sets and assignments in either Python or R. Basic best practices for building a codebase will be introduced at the beginning of the course and students will be expected to follow these best practices.

No prior exposure to Python or R is assumed. In fact, the TA sessions and problem sets should guide students through the learning process. Prior exposure to some programming language is, however, extremely helpful.

Preliminary Schedule

1. Introduction & course overview
2. Causality - introduction, experiments, unconfoundedness
3. Causal inference - estimation under unconfoundedness
4. Causal inference continued - IVs, Diff-in-Diff, Regression Discontinuity, Synthetic Control
5. High-dimensional data & dimensionality reduction, clustering
6. High-dimensional linear regression, applications to causal inference
7. Q&A session - learnings to date, how these methods are relevant in the job market, job search recommendations, and other questions
8. Tree-based methods, bagging, boosting, applications to causal inference
9. Random forests and ensemble methods, applications to causal inference
10. More "big data" methods for causal inference
11. Spatial econometrics
12. Neural networks
13. Text analysis
14. Examples, Q&A