Economics 570: Big Data Econometrics

Fall 2023

Instructor:	Yuehao Bai
Office:	Kaprielian (KAP) Hall [TBA]
Email:	yuehao.bai@usc.edu
Office Hours:	[TBA]
TA:	[TBA]
Office Hours:	[TBA]

Course Time and Location:	Tuesday, 4pm-6:50pm, Leavey Library 17
Course Webpage:	Blackboard

Course Description

This course provides an introduction to machine learning and related methods for big data from the perspective of economics. Students will be introduced to modern estimation methods for high-dimensional data, which will be illustrated through applications to causal inference and prediction problems in economics, business, and related fields. Students will gain experience working with these methods through programming assignments. The course will be focused on methodology and its practical application and will culminate in an empirical project in which students apply course concepts to real-world data. By the end of the course, students should be able to do the following.

- Apply machine learning methods to estimate causal effects in experimental and observational settings and solve prediction problems.
- Understand the uses and limitations of machine learning for answering economic questions.
- Implement machine learning algorithms using R.

Optional Textbooks

Much of the lecture material draws from the first textbook below.

BDS Taddy, M. (2019) *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*. McGraw Hill.

ISL Gareth, J.; Witten, D.; Hastie, T.; and Tibshirani, R. (2017) *An Introduction to Statistical Learning*. Springer.

MM Angrist, J. and Pischke, J. (2014) *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Prerequisites

Econometrics at the level of ECON 513 is required, along with calculus at the level of MATH 226 and linear algebra at the level of MATH 225.

Software

This course will use R, a free statistical programming language. While you do not need to have prior knowledge of R to take this course, you will be expected to learn R on your own time not only through lectures and homework, but also by searching independently for relevant commands, tutorials and documentation using Google, StackOverflow, the R help() command, etc. Programming experience in another language should suffice, so long as you are willing to learn R on your own time.

Evaluation

Grades will be based on problem assignments (60%) and a final paper (40%).

Homework

Regular programming exercises will be assigned to reinforce the concepts taught in class, as well as offer an opportunity for students to code and implement the algorithms covered. Students will work in groups of 4-5, which will be assigned at the beginning of the course, and then reassigned about halfway through the course. While each student is expected to contribute, only one problem set will be turned in per group. R must be used as the language for programming exercises. Problem sets will be turned in through Blackboard, along with code and any graphs generated. Assessment will be based on whether the right approaches were used and whether the right solutions were obtained.

Final Project

Students are expected to work in groups of 4-5 students each and apply course concepts to a real-world data set to tackle an empirical problem that interests them. Students are expected to collect their own data as part of the project. Suggestions for starting points for publicly available data will be posted on Blackboard. Each group will submit a writeup, as well as code to reproduce the analysis. Groups will give short presentations of their work in class during the last few lectures. Assessment will be based on how appropriately the quantitative tools were applied. Due date for this project is TBA.

Homework Policies

Problem sets will be discussed in class a week or more after the original due date. The absolute deadline for turning in a problem set is when solutions are made available: homework turned in after this second deadline will receive zero points. 15% will be taken off for problem sets that are turned in late, but before the problem set is discussed in class.

Course Outline

1. Introduction (0.5 weeks; BDS Intro; ISL Ch. 2).
    (a) Prediction vs. causal inference.
    (b) Intro to R.
2. Sampling (1 week; BDS Ch. 1; ISL Ch. 5.2).
    (a) CLT and standard errors.
    (b) Bootstrap.
3. Regression (1.5 weeks; BDS Ch. 2; ISL Ch. 3).
    (a) Linear conditional mean model.
    (b) Logistic regression.
4. Fundamental concepts in ML (2.5 weeks; BDS Ch. 3; ISL Ch. 2, 5.1, 6).
    (a) Cross-validation.
    (b) Regularization.
    (c) Bias-variance trade-off.
5. Classification (1.5 weeks; BDS Ch. 4; ISL Ch. 4.3).
    (a) Logistic lasso.
    (b) Multinomial logit.

6. Causal inference (2.5 weeks; BDS Ch. 5, 6; MM Ch. 1, 2).
    (a) Randomized control trials.
    (b) High-dimensional controls.
7. Woodlands (1 week; BDS Ch. 9; ISL Ch. 8).
    (a) Classification and regression trees.
    (b) Random forests.
8. Unsupervised learning (2 weeks; BDS Ch. 7; ISL Ch. 10).
    (a) Principal components analysis.
    (b) Partial least squares.
9. Natural language processing (1.5 weeks; BDS Ch. 8).
    (a) Tokenization and bag-of-words.
    (b) Topic models.
10. Deep learning (1 week; BDS Ch. 10).
    (a) Architecture of neural networks.
    (b) Stochastic gradient descent.