Machine Learning and Causal Inference

Course Overview

This course will cover statistical methods based on the machine learning literature that can be used for causal inference. In economics and the social sciences more broadly, empirical analyses typically estimate the effects of counterfactual policies, such as the effect of implementing a government policy, changing a price, showing advertisements, or introducing new products. Recent advances in supervised and unsupervised machine learning provide systematic approaches to model selection and prediction, methods that are particularly well suited to datasets with many observations and/or many covariates. This course will review when and how machine learning methods can be used for causal inference, and it will also review recent modifications and extensions to standard methods to adapt them to causal inference and provide statistical theory for hypothesis testing. Applications to the evaluation of large-scale experiments, including online A/B tests and experiments on networks, will receive special attention. We will also consider topic modeling and a brief overview of textual analysis.

What this course is NOT

This course is not intended to be a *substitute* for an econometrics course or for a machine learning course. Instead, this course is designed as a complement to these courses. The course will also spend very limited time on the discovery of causal relationships (the question of how to test whether A causes B), as this type of analysis is not commonly studied in social science. Instead, we will focus on the measurement of causal effects and the ability to draw inference about estimated effects.

Target audience of the course

This course is intended primarily to help empirical researchers learn about how machine learning methods can be used to answer causal questions; the main applications will be in economics and social science, although we will also spend some time on applications from medicine and other fields. Econometricians or statisticians might also find the course useful for identifying open questions and learning about what tools are likely to be most useful to applied researchers interested in causal questions.

Background Required

This course will present machine learning concepts without assuming any background knowledge, other than required readings. The course assumes that students have prior exposure to statistics and data analysis; economics students should have completed the first-year econometrics sequence, while statistics students should be familiar with some applied empirical work as well as language and concepts commonly used in empirical work in social sciences. Students who are entirely unfamiliar with machine learning prior to the class should spend a little bit of time reviewing one of the recommended machine learning textbooks over the first two weeks of the class. Students who are entirely unfamiliar with the potential outcomes model for causal inference should review the early chapters of the Imbens and Rubin book on causal inference. For a more accessible introduction, *Mostly Harmless Econometrics* by

Angrist and Pischke provides a lot of intuition about causal inference in economics. In principle, the course should be accessible to master's students in statistics or computer science (given lots of exposure to machine learning), but some may find the pace challenging.

The homeworks will consist of applications of methods discussed in class to data sets provided by the instructor. It is possible to do well in the class without a strong background in proving theorems, but some class time will be devoted to theoretical concepts.

Contact Info	Susan Athey Office: E311 athey@stanford.edu
	Stefan Wager Office: E328 swager@stanford.edu
Office Hours	By appointment
Faculty Support	Danielle Tamagni dtamagni@stanford.edu

COURSE REQUIREMENTS

I. Regular Homework Assignments (30%)

II. Final Project Assignment (70%)

I. REGULAR HOMEWORK ASSIGNMENTS (30%)

The class will require several shorter homework assignments during the term of the class. These exercises give students practice with writing and testing code, and discussing results. Students may collaborate on code, but should submit individual write-ups of what they learned.

II. FINAL PROJECT ASSIGNMENT (70%)

Your final project assignment consists of **two parts.** Both must be submitted on the Canvas website by June 6 at 11:59pm.

The **first part** is an empirical analysis of a question involving causal inference (including prediction policy questions), using Machine Learning methods or other methods introduced in the course. Ideally you would compare at least 2-3 alternative methods, although a multi-step project that uses different methods at each step is also possible. You can bring your own data, or use one of the data sets provided for the class. You may work on this in a group of up to 4 students. You should submit your code and output, preferably produced as a knit file that shows the code and the results together.

The **second part** is a 3-5 page writeup of your findings (12 pt., 1.5 spacing, 1 inch margins, page count not including any figures or references). The write-up should explain your question, your choice of methods, and what conclusions you may draw.