Machine Learning for econometrics

Causal perspective

Matthieu Doutreligne

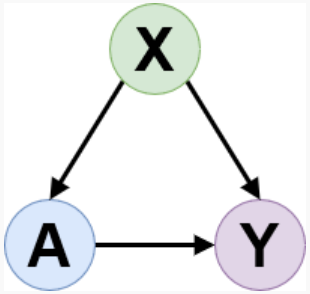
January 10, 2025

Table of contents

Table of contents

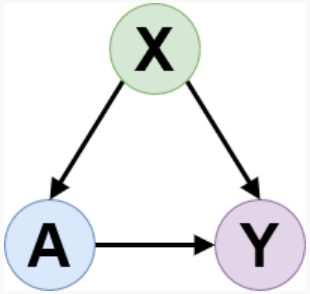
Introduction

Causal inference: subfield of statistics dealing with "why questions"



At the center of epidemiology, econometrics, social sciences...

Causal inference: subfield of statistics dealing with "why questions"



At the center of epidemiology, econometrics, social sciences...

Now, bridging with Machine Learning (Kaddour, Lynch, Liu, Kusner, & Silva, 2022)

What is a "why question"?

- Economics: How does supply and demand (causally) depend on price?
- Policy: Are job training programmes actually effective?
- Epidemiology: How does this threatment affect the patient's health?
- Public health : Is this prevention campaign effective?
- Psychology: What is the effect of family structure on children's outcome?
- Sociology: What is the effect of social media on political opinions?

This is different from a predictive question

- What will be the weather tomorrow?
- What will be the outcome of the next election?
- How many people will get infected by flue next season?
- What is the cardio-vacular risk of this patient?
- How much will the price of a stock be tomorrow?

Why is prediction different from causation?

- Prediction (most part of machine learning) focus on understanding what usually happens in a given situation.

Why is prediction different from causation?

- Prediction (most part of machine learning) focus on understanding what usually happens in a given situation.

Important assumption Train and test data are drawn from the same distribution.

Why is prediction different from causation?

- Prediction (most part of machine learning) focus on understanding what usually happens in a given situation.

- Causal inference (most part of economists) focus on what would happen if we changed the system ie. under intervention.

It models the covariate shift between treated and control units.

Why is prediction different from causation?

- Prediction (most part of machine learning) focus on understanding what usually happens in a given situation.

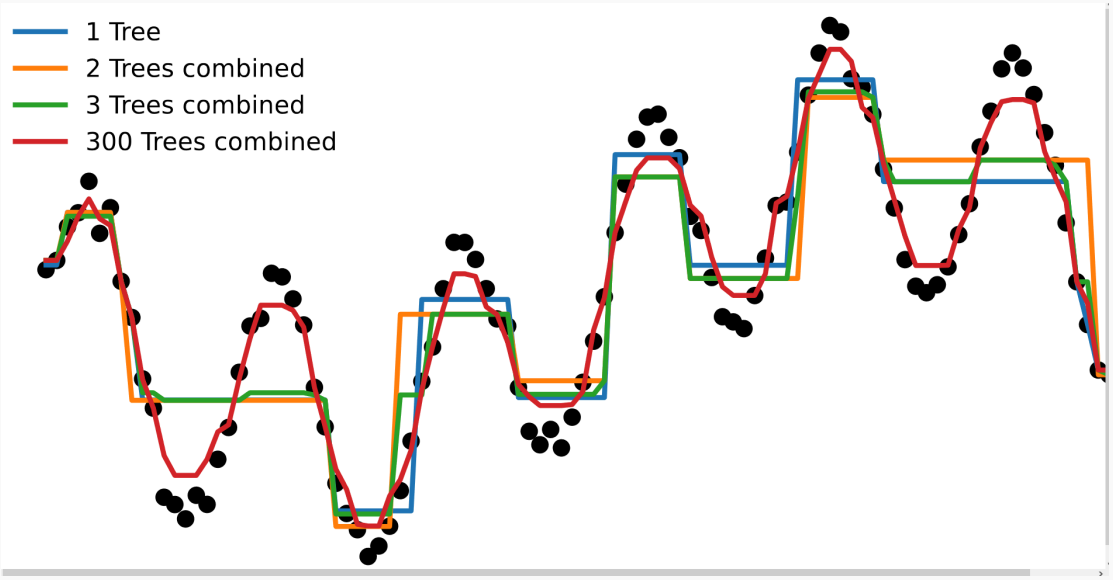
- Causal inference (most part of economists) focus on what would happen if we changed the system ie. under intervention.

It models the covariate shift between treated and control units.

Important assumption Train and test data are drawn from the same distribution.

Machine learning is pattern matching (ie. curve fitting)

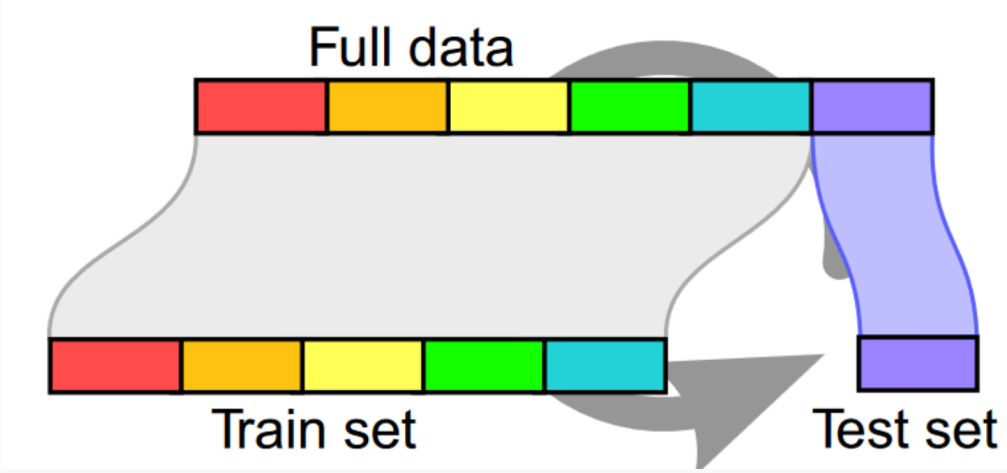
Find an estimator $f : x \longrightarrow y$ that approximates the true value of y so that $f(x) \approx y$



Boosted trees : iterative ensemble of decision trees

Machine learning is pattern matching that generalizes to new data

Select models based on their ability to generalize to new data : (train, test) splits and cross validation (Stone, 1974).



"Cross validation" (Varoquaux et al., 2017)

How to ask a sound causal question: The PICO framework

Identify the target trial

What would be the ideal randomized experiment to answer the question? (Hernán & Robins, 2016)

PICO framework

- Population : Who are we interested in?
- Intervention : What treatment/intervention do we study?
- Comparison : What are we comparing it to?
- Outcome : What are we interested in?

PICO framework, an illustration

- P
- I
- C
- O

Causal graphs

Directed acyclic graphs (DAG): reason about causality

What are the important depedencies between variables?

The four steps of causal inferenceidentification, statistical estimand, statistical inference

Causal estimand

What can we learn from the data?

Identification

What can we learn from the data?

Knowledge based

Cannot be validated with data

Potential outcomes

Statistical estimand

Statistical inference ie. estimation

Related concepts

- Structural equations:

Resources

- https://web.stanford.edu/~swager/stats361.pdf
- https://www.mixtapesessions.io/
- https://alejandroschuler.github.io/mci/

Bibliography

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. American Journal of Epidemiology, 183(8), 758–764.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2022). Causal machine learning: A survey and open problems. Arxiv Preprint Arxiv:2206.15475.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111–133.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage, 145, 166–179.