Machine Learning for econometrics

Causal perspective

Matthieu Doutreligne

January 10, 2025

Table of contents

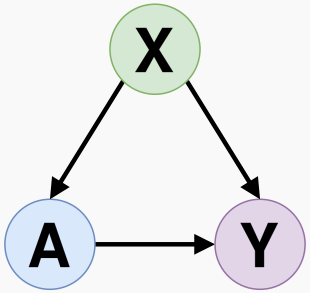
1. Introduction
2. How to ask a sound causal question: The PICO framework
3. Causal graphs
4. Four steps of causal inference : identification, estimand, causal and statistical inference, vibration analysis
5. Potential outcomes
6. Causal inference
7. Statistical estimand

Table of contents

8. Statistical inference ie. estimation
9. Related concepts

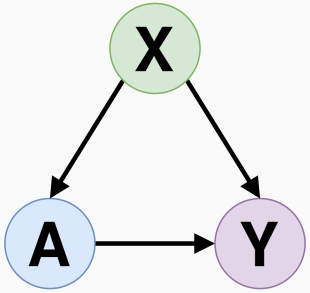
Introduction

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



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

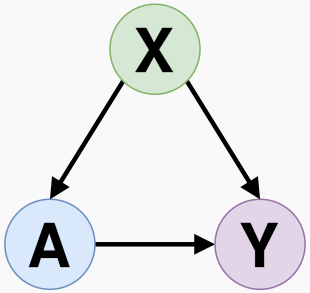
Causal inference: subfield of statistics dealing with "why questions"



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

Now, bridging with machine learning (Kaddour et al., 2022) : Fairness, reinforcement learning, causal discovery, causal inference for LLM, causal representations...

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



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

This course: Basis of causal inference using ML appraoches (semi-parametric), inspiration from epidemiology and application for applied econometrics.

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? (1/2)

- Prediction (most part of ML): What usually happens in a given situation?

Why is prediction different from causation? (1/2)

- Prediction (most part of ML): What usually happens in a given situation?

Assumption Train and test data are drawn from the same distribution.

Prediction models (X, Y)

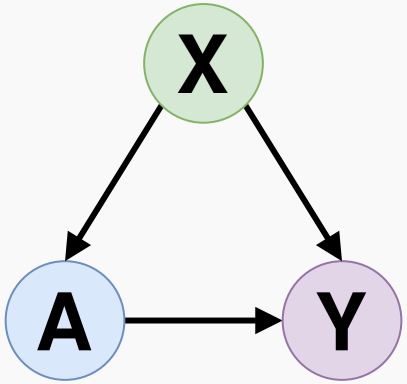
Why is prediction different from causation? (2/2)

- Causal inference (most part of economists) : What would happen if we changed the system ie. under intervention?

Why is prediction different from causation? (2/2)

- Causal inference (most part of economists) : What would happen if we changed the system ie. under intervention?

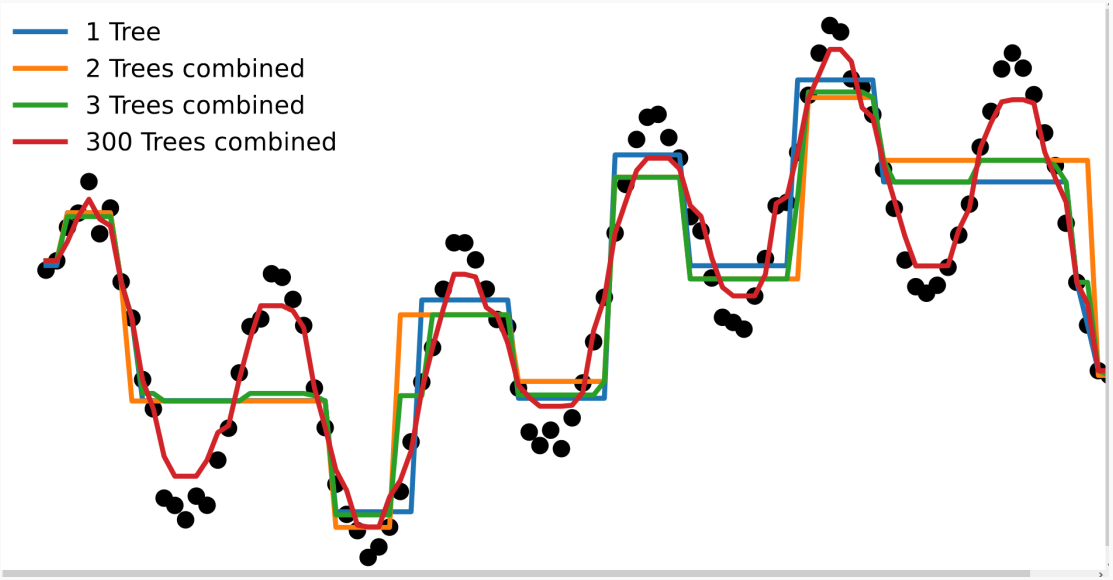
Assumption: No unmeasured variables influencing both treatment and outcome $\longrightarrow$ confounders.



Causal inference models $(X, A, Y(A = 1), Y(A = 0))$, the covariate shift between treated and control units.

Machine learning is pattern matching (ie. curve fitting)

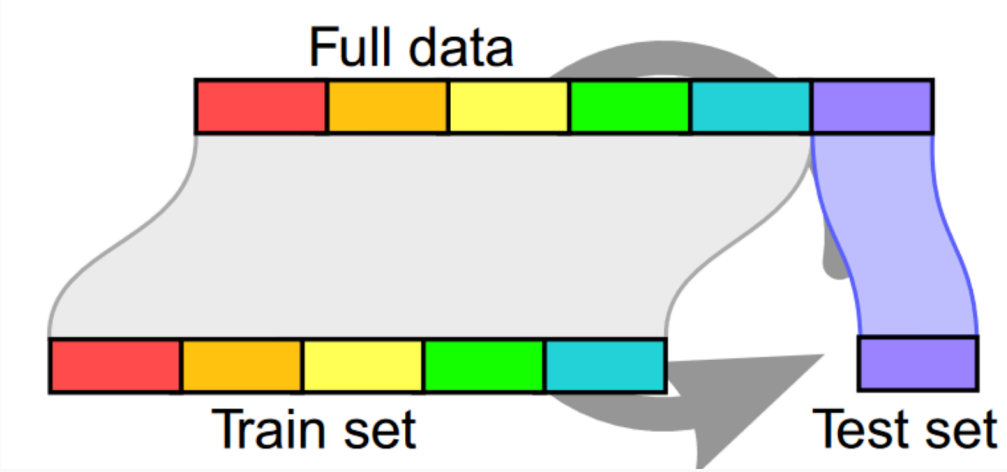
Find an estimator $f : x \longrightarrow y$ that approximates the true value of y so that $f(x) \approx y$



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

Component	Description	Notation	Example
Population	What is the target population of interest?	$X \sim P(X)$	Patients with sepsis in the ICU
Intervention	What is the treatment?	$A \sim P(A = 1) = p_A$	Crystalloids and albumin combination
Control	What is the clinically relevant comparator?	$1 - A \sim 1 - p_A$	Crystalloids only
Outcome	What are the outcomes?	$Y(1), Y(0) \sim P(Y(1), Y(0))$	28-day mortality
Time	Is the start of follow-up aligned with intervention assignment?	N/A	Intervention administered within the first 24 hours of admission

Causal graphs

Directed acyclic graphs (DAG): reason about causality

What are the important depedencies between variables?

Four steps of causal inference : identification, estimand, causal and statistical inference, vibration analysis

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

Causal inference

PICO framework and the potential outcomes

Component	Description	Notation	Example
Population	What is the target population of interest?	$X \sim P(X)$	Patients with sepsis in the ICU
Intervention	What is the treatment?	$A \sim P(A = 1) = p_A$	Crystalloids and albumin combination
Control	What is the clinically relevant comparator?	$1 - A \sim 1 - p_A$	Crystalloids only
Outcome	What are the outcomes?	$Y(1), Y(0) \sim P(Y(1), Y(0))$	28-day mortality
Time	Is the start of follow-up aligned with intervention assignment?	N/A	Intervention administered within the first 24 hours of admission

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.