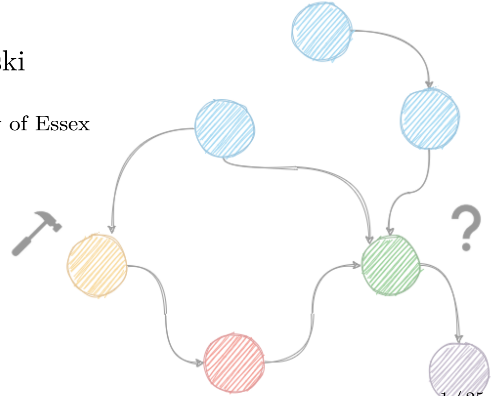
Machine Learning for Causal Inference from Observational Data

Damian Machlanski

CSEE and MiSoC, University of Essex

July 29th 2021



- Introduction
- Motivation
- Causality
- Methods
- Conclusion

Welcome!

- Agenda
  - Slides: Introduction to Causal Inference
  - Tutorial: Guided Example with Code
  - Exercise: Do It Yourself

With some breaks in the middle as necessary.

Resources

- Textbooks
  - J. Pearl, M. Glymour, and N. P. Jewell, Causal Inference in Statistics: A Primer. John Wiley & Sons, 2016.[1]
  - J. Peters, D. Janzing, and B. Scholkopf, Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press, 2017.[2]
- Online
  - Introduction to Causal Inference[3]

[1] http://bayes.cs.ucla.edu/PRIMER/

[2] https://mitpress.mit.edu/books/elements-causal-inference

[3] https://www.bradyneal.com/causal-inference-course

Tools

We are going to use the following:

- Python 3
- numpy
- pandas
- matplotlib
- scikit-learn
- EconML[4]
- Google Colab

[4] `https://github.com/microsoft/EconML`

Machine Learning

We will need the following:

- Supervised learning - predict y given (X, y) samples
  - Regression (continuous outcome)
  - Classification (binary outcome)
- Basic data exploration
- Data pre-processing
- Cross-validation
- Model selection

Problem Setting

- We want to estimate the *causal effect* of treatment T on outcome Y
  - What benefits accrue if we intervene to change T?
  - Treatment must be modifiable
  - We observe only one outcome per each individual
- Example:
  - My headache went away after I had taken the aspirin
  - Would the headache have gone away without taking the aspirin?
  - We cannot go back in time and test the alternative!
  - Treatment effect
  - Test more people and measure the average outcome?

Randomised Controlled Trials

- Data from controlled experiments
- Randomised T - people assigned $T = 0$ (control) or $T = 1$ (treated)
- This mimicks observing alternative reality
- Record background characteristics as X
- Can be expensive or even unfeasible (smoking)
- Figure

Observational Data

- Passively collected data (non-experimental)
- Abundant nowadays
- Quasi-experimental study
- Keep only X recorded before Y
- Figure

ML Perspective

- Correlation (association) vs causation
- Confounders
- Domain shift/adaptation perspective
- OOD generalisation
- Learn from given individuals, but predict unseen examples
- Cannot learn from counterfactuals
- On the surface it looks the same as supervised ML - predict Y given (X, Y) samples

Fundamentals

Table of people and factual outcomes.

$$Effect = outcome_T - outcome_C$$

But we observe only one outcome!

This is known as the fundamental problem of causal inference. Can't know the difference. Goal: approximate ITE/ATE.

Treatment Effect

Let us define the **true** outcome $\mathcal{Y}_t^{(i)}$ of individual (i) that received treatment t. The Individual Treatment Effect (ITE) is then defined as follows:

$$ITE^{(i)} = \mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}$$

The Average Treatment Effect (ATE) builds on ITE:

$$ATE = \mathbb{E}[ITE]$$

Metrics

- In practice, we want to measure how accurate our inference model is
- This is often done by measuring the amount of error (ϵ) or risk ($\mathcal{R}$) introduced by a model
- Examples:
  - ϵ_{ITE}
  - ϵ_{ATE}
  - ϵ_{PEHE}
  - ϵ_{ATT}
  - $\mathcal{R}_{pol}$

ϵ_{ATE} and ϵ_{PEHE} are the most common ones and we will focus on them.

Metrics - Predictions

Let us denote $\hat{y}_t^{(i)}$ as **predicted** outcome for individual (i) that received treatment t. Then, our predicted ITE and ATE can be written as:

$$\widehat{ITE}^{(i)} = \hat{y}_1^{(i)} - \hat{y}_0^{(i)}$$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \widehat{ITE}^{(i)}$$

Metrics - Measuring Errors

This allows us to define the following measurement errors:

$$\epsilon_{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{ITE}^{(i)} - ITE^{(i)})^2}$$

$$\epsilon_{ATE} = \left|\widehat{ATE} - ATE\right|$$

Where $PEHE$ stands for Precision in Estimation of Heterogeneous Effect, and which essentially is a Root Mean Squared Error (RMSE) between predicted and true ITEs.

Benchmark Datasets

Semi-simulated data or combinations of experimental and observaional datasets. We use metrics depending on what outcomes we have access to. Counterfactuals - ATE and PEHE. Otherwise ATT.

Well-established causal inference datasets:

- IHDP
- Jobs
- News
- Twins
- ACIC challenges

Assumptions

- No hidden confounders (we observe everything)
- All background covariates X happened *before* the outcome Y
- Modifiable treatment T
- Stable Unit Treatment Value Assumption (SUTVA):
  - No interference between units
  - Consistent treatment (different versions disallowed)
- Ignorability
  - Y conditionally independent from X and T

Modern Approaches

Mosty regression and classification (classic ML), but combined in a smart way.

- Recent surveys on modern causal inference methods [5] [6]
- Most popular:
  - Inverse Propensity Weighting (IPW)
  - Doubly-Robust
  - Double/Debiased Machine Learning
  - Causal Forests
  - Meta-Learners
  - Multiple based on neural networks (very advanced)

We will start with a simple regression, enhance it with IPW, and conclude with Meta-Learners.

[5] `https://dl.acm.org/doi/10.1145/3397269`
[6] `https://arxiv.org/abs/2002.02770`

S-Learner

Naive regression. Use all data to predict Y.

Biased Estimators

Show a biased example.

IPW Estimator

A combination of regression and classification. Sample importance. Show a biased example and how sample importance can help. PS can be misspecified, so can be the estimator. Doubly-Robust attempts to address that.

T-Learner

T and C distributions are often different. Fit two separate regressors.

X-Learner

A hybrid of the previous approaches.

Summary

Brief summary of the content.

What's next?

- Onto the practical parts
  - Tutorial
    - Predict ATE and measure ϵ_{ATE}
    - S-Learner, IPW and X-Learner
    - Random Forest as base regressors and classifiers
  - Exercise - IHDP
    - Predict ITE and ATE
    - Measure ϵ_{PEHE} and ϵ_{ATE}
  - Exercise - JOBS (optional)
    - Predict ATT and Policy
    - Measure ϵ_{ATT} and $\mathcal{R}_{pol}$
- Short break?