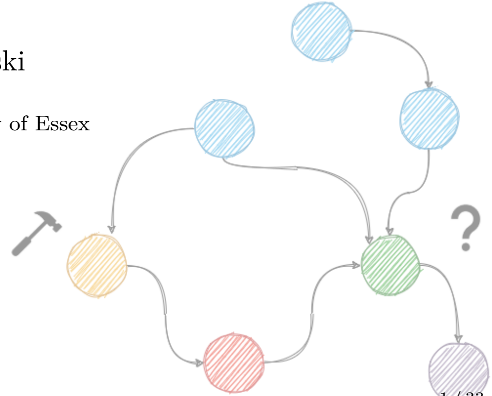
Machine Learning for Causal Inference from Observational Data

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:

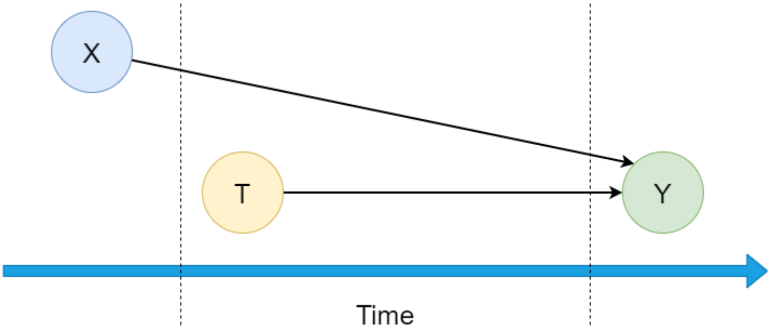
- Supervised learning - predict y given (X, y) samples
  - Regression (continuous outcome)
  - Classification (binary outcome)
- Basic data exploration
- Data pre-processing
- Cross-validation
- Model selection

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?

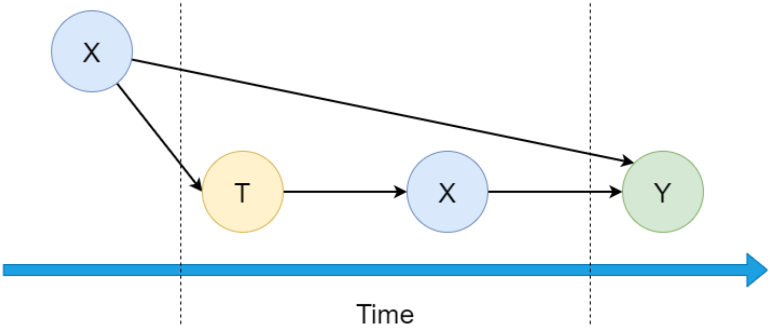
Randomised Controlled Trials

- Data from controlled experiments
- Randomised T - people assigned $T = 0$ (control) or $T = 1$ (treated)
- This mimicks observing alternative reality
- Record background characteristics as $X = [X_1, X_2, ..., X_n]$
- Can be expensive or even unfeasible (e.g. smoking)



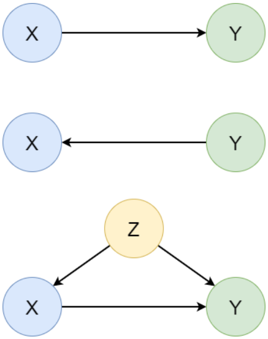
Observational Data

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



ML Perspective

- Correlation (association) vs causation
- The role of confounders
- Domain shift/adaptation perspective
- Out-of-distribution (OOD) generalisation
- Learn from given individuals, but predict unseen examples
- Cannot learn from counterfactuals
- On the surface it looks the same as supervised ML
  - ML: predict Y given (X, Y) samples
  - CI: predict **effects** given (X, Y) samples



- Learn: $[x_i, t_i, y_i]$
- Predict: $[x_i, 1 - t_i]$ -> ?

Fundamentals

$$Effect = Y_1 - Y_0$$

#	X_1	X_2	X_3	T	Y_0	Y_1
1	1.397	0.996	0	1	**?**	4.771
2	0.269	0.196	1	0	2.956	**?**
3	1.051	1.795	1	1	**?**	4.164
4	0.662	0.196	0	1	**?**	6.172
5	0.856	1.795	1	0	7.834	**?**

But we observe only one outcome!

This is known as the fundamental problem of causal inference. We cannot *know* the difference. But we can **approximate** it.

Treatment Effect

Let us define the **true** outcome $\mathcal{Y}_t^{(i)}$ of individual (i) that received treatment $t \in \{0, 1\}$. 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:

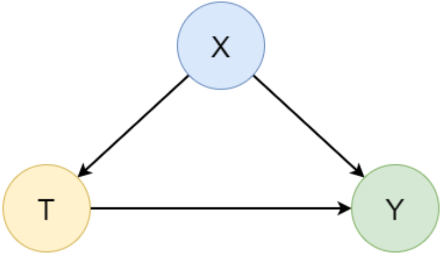
- IHDP
- Jobs
- News
- Twins
- ACIC challenges

Assumptions

- Ignorability:
  - 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)

Assumptions (2)

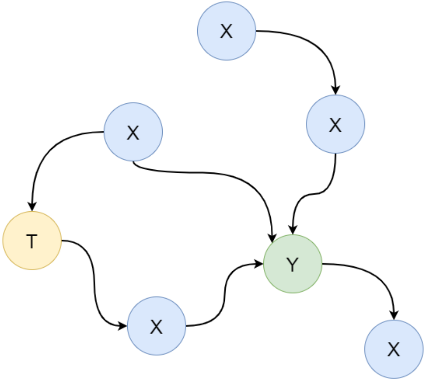
- Most CI estimators assume the *triangle* graph



- This is a very simplistic view of the world
- Actual reality can be much more complex

Assumptions (3)

- Can we infer graphs from data?
- Causal discovery



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

We want to estimate

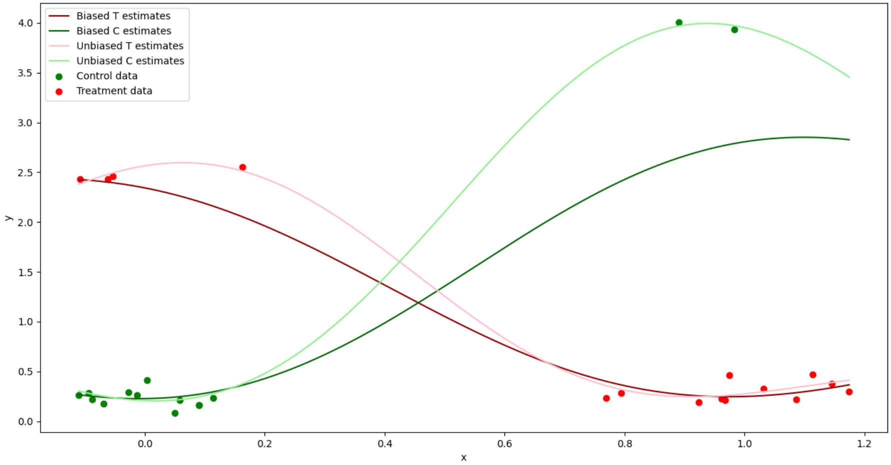
$$\mu(t, x) = \mathbb{E}[\mathcal{Y} | X = x, T = t]$$

1. Obtain $\hat{\mu}(t, x)$ estimator.
2. Predict ITE as

$$\widehat{ITE}(x) = \hat{\mu}(1, x) - \hat{\mu}(0, x)$$

- *Single* model approach
- Allows heterogenous treatment effects
- Can be biased (next slide)

Biased Estimators



Propensity Score

$$e(x) = P(t_i = 1 | x_i = x)$$

- Probability of a unit i receiving the treatment $(T = 1)$
- For discrete treatments, this is a classification problem
- Binary classification in most cases as $t \in \{0, 1\}$
- We denote $\hat{e}(x)$ as our estimation

IPW Estimator

Using the propensity score $\hat{e}(x)$, we can obtain the following weights

$$w_i = \frac{t_i}{\hat{e}(x_i)} + \frac{1 - t_i}{1 - \hat{e}(x_i)}$$

- These are called Inverse Propensity Weights (IPW)
- Use the weights to perform **weighted** regression
- Similar to S-Learner, but combines regression and classification
- Sample importance (pay attention to scarce data points)
- Either $\hat{e}(x)$ or $\hat{\mu}(x)$ can still have bias (misspecification)
- Doubly-Robust method attempts to address that

T-Learner

- Treated and control distributions are often different
- Solution: fit *two* separate regressors

$$\mu_1(x) = \mathbb{E}[\mathcal{Y}|X = x, T = 1]$$

$$\mu_0(x) = \mathbb{E}[\mathcal{Y}|X = x, T = 0]$$

1. Learn $\mu_1(x)$ from treated units, obtain $\hat{\mu}_1(x)$.
2. Learn $\mu_0(x)$ from control units, obtain $\hat{\mu}_0(x)$.
3. Predict ITE as

$$\widehat{ITE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

X-Learner

A hybrid of the previous approaches. There are three main stages.

Stage 1 (same as T-Learner)

1. Learn $\mu_1(x)$ from treated units, obtain $\hat{\mu}_1(x)$.
2. Learn $\mu_0(x)$ from control units, obtain $\hat{\mu}_0(x)$.

X-Learner (2)

Stage 2

Define *imputed* treatment effects as:

$$\mathcal{D}_0^{(i)} = \hat{\mu}_1(X_0^{(i)}) - \mathcal{Y}_0^{(i)}$$
$$\mathcal{D}_1^{(i)} = \mathcal{Y}_1^{(i)} - \hat{\mu}_0(X_1^{(i)})$$

Use provided regressors to model $\mathcal{D}_0$ and $\mathcal{D}_1$ separately. The response functions are formally defined as:

$$\tau_0(x) = \mathbb{E}[\mathcal{D}_0 | X = x]$$
$$\tau_1(x) = \mathbb{E}[\mathcal{D}_1 | X = x]$$

We denote estimated functions as $\hat{\tau}_0$ and $\hat{\tau}_1$.

X-Learner (3)

Stage 3

The final treatment effect estimate is a weighted average of the two estimates from Stage 2:

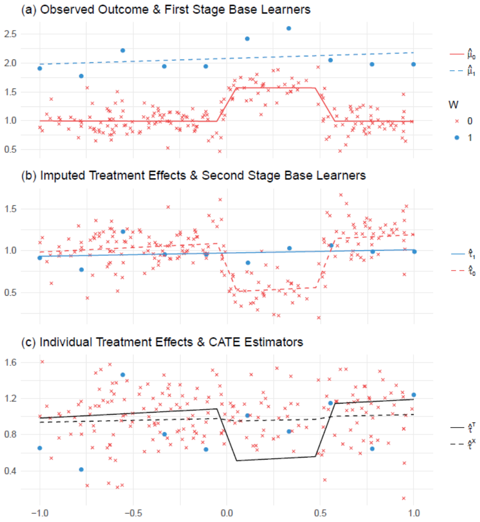
$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$$

Where $g \in [0, 1]$ is a weight function. In practice, g can be modelled as a propensity score function $e(x)$.

Using a provided classifier, we can obtain an estimate $\hat{e}$ that can be used in place of g. That is:

$$\hat{\tau}(x) = \hat{e}(x)\hat{\tau}_0(x) + (1 - \hat{e}(x))\hat{\tau}_1(x)$$

X-Learner - Intuition



Summary

- Causal inference is about measuring causal effects
  - Cannot calculate them exactly due to missing counterfactuals
  - But we can approximate them through data
- RCTs are the most reliable source of data, but can be unfeasible to get
- Non-experimental data are a great alternative, but can be *biased*
- Most methods are about finding *unbiased* estimators
- Machine Learning and Causal Inference can be both mutually beneficial
  - ML delivers better CI estimators
  - CI helps ML with OOD generalisation (domain adaptation)
- Assumptions are important and must be considered in applications

Acknowledgements

This course builds heavily on the materials from *Introduction to Machine Learning for Causal Analysis Using Observational Data* online course, delivered on June 22-23 2021 by Damian Machlanski, Dr Spyros Samothrakis and Professor Paul Clarke.

References

- J. M. Robins, A. Rotnitzky, and L. P. Zhao, 'Estimation of Regression Coefficients When Some Regressors are not Always Observed', Journal of the American Statistical Association, vol. 89, no. 427, pp. 846–866, Sep. 1994.
- U. Shalit, F. D. Johansson, and D. Sontag, 'Estimating individual treatment effect: generalization bounds and algorithms', in International Conference on Machine Learning, Jul. 2017, pp. 3076–3085.
- V. Chernozhukov et al., 'Double/debiased machine learning for treatment and structural parameters', The Econometrics Journal, vol. 21, no. 1, pp. C1–C68, Feb. 2018.
- S. Wager and S. Athey, 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests', Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228–1242, Jul. 2018.
- S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, 'Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning', Proc Natl Acad Sci USA, vol. 116, no. 10, pp. 4156–4165, Mar. 2019.
- R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, 'A Survey of Learning Causality with Data: Problems and Methods', ACM Comput. Surv., vol. 53, no. 4, p. 75:1-75:37, Jul. 2020.
- L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, 'A Survey on Causal Inference', arXiv:2002.02770 [cs, stat], Feb. 2020.

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?