Causal Inference

Machine Learning for Causal Inference from Observational Data

February 2, 2022

1 Project description

This *Causal Inference* (CI) project is about causal (treatment) effect estimation. That is, the problem is to infer effects of interventions on a system given only observed (factual) outcomes, without access to 'alternative reality' results (counterfactuals) — the issue known as the fundamental problem of causal inference.

To give you a better intuition behind the problem, consider the following example. Imagine I have a headache. I decide to take aspirin to get rid of it. After some time, the headache goes away. Now I ask myself: Would the headache go away without me taking the aspirin? Would it go away as fast? The act of taking the aspirin is our intervention (also called *treatment*). We have two actions possible: take it ($t = 1$) or not take it ($t = 0$). I decided to take it, so after a while we are able to observe the result (*outcome*) of that action (y_1). Now, the core problem here is that in order to answer the above questions about the effectiveness of the aspirin (the causal effect), we have to compare the outcomes of both scenarios (when I take the aspirin and when I do not), that is, we need access to both y_0 and y_1. Unfortunately, the way our world works is that once we apply an action, it is impossible to go back in time and change our decision. Thus, the only outcome we will ever observe is the one that actually happened (called *factual*). We cannot observe the other one that did not happen (*counterfactual*). This known problem is partially addressed through Randomised Controlled Trials (RCTs), but those experiments are usually very expensive, at times even impossible due to ethical reasons (we cannot make participants smoke/drink alcohol for a number of years). An alternative to RCTs are passively collected observational data, for instance, historical information about past patients in a hospital. However, we have to be careful about how we use such datasets as they are not randomised as RCTs are, causing some issues as a consequence. Some of the most common ones are *selection bias* (e.g. underrepresented groups of people), *hidden confounders* (omitted important features), or *covariate shifts* (discrepancies between the distributions of treated and control units).

How can we address all of these difficult problems? First, we start with the idea that even though we cannot obtain the causal effects directly through observed outcomes ($y_1 - y_0$), we can **approximate** them. Modern Machine Learning (ML) methods are excellent estimators, and often handle inherent data issues (see the three examples above) reasonably well, too. Second, there is abundance of non-experimental data nowadays, opening up opportunities to answer causal questions that could not be answered otherwise with traditional RCTs. These two observations motivate this project — to use ML techniques and methods in order to approximate accurate causal effects from observational data. As the ML methods are at the core of this approach, it is in principle very similar to the classic supervised learning ML, with a few extra bits here and there to adjust to the specifics of inferring causal effects.

The general aims of this project are as follows:

- Explore common CI benchmark datasets, their properties, and relevant performance metrics.
- Get familiar with well-established types of CI estimators, including an understanding of their main characteristics.
- Develop an intuition about how CI problems relate to the usual machine learning tasks.

2 Tasks to be done for each assignment

- **Assignment 1.** Tasks 1, 2, 3a, 5a from Section 3. Think about 4 (understand the problem and plan next steps — no need to actually do this yet).

- **Assignment 2.** The rest of the tasks (excluding optional ones) from Section 3.
  - Tasks marked as **"optional"** are suggestions on how to progress your work beyond the main brief. These will be taken into account when marking in the "How much depth was achieved in the project?" section of the marking scheme for Assignment 2.

3 Tasks

1. Load, clean, and explore the datasets provided (IHDP and JOBS), which are described in Section 4.
   - Can you identify the causal questions for each dataset?
2. Select appropriate evaluation metrics for each dataset. Justify your choices. Things to consider:
   - Granularity of predictions and measured errors (average vs. individual).
   - Whether the dataset includes counterfactual outcomes (or true individual effects).
3. **Simple learners**
   (a) Choose at least one regression model that provides feature importances. Examples: *linear regression*, *decision tree*, *random forest*, *boosted trees*.
   - Taking into account the model and the dataset, preprocess the data if needed.
   - In your report, make sure that you justify your choice of regressor. All models have strengths and weaknesses — what makes the one you chose more appealing?

   (b) Train the model(s) of your choice on both datasets using all available features (including *treatment*).
   (c) Make effect predictions and compute relevant metrics.
   - Make predictions $\hat{y}_1$ and $\hat{y}_0$ by setting *treatment* to 1 and 0 respectively. Obtain effect estimates as $\hat{y}_1 - \hat{y}_0$.

   (d) Perform hyperparameter optimisation for your model(s) using grid search and appropriate modeling techniques.
   (e) Report the performance (chosen metric(s)) of your model(s) using 10-fold cross-validation or a training/validation/test split.
   (f) Plot feature importances and briefly comment on them.
4. **Propensity score re-weighting**
   (a) Train a classifier (of your choice) to predict *propensity scores* based on background features X. See section 5.1 for more details.
   (b) Create a function that calculates sample weights (w_i) based on their propensity scores ($e(x_i)$) you calculated in the previous step.
   (c) Use the weights returned by the function as the sample weights to train a weighted regressor (use the regressor(s) chosen in task 3).
   (d) Repeat steps 3d - 3f, but for weighted regression.
   (e) Things to reflect on:
   - Do you see any differences in the importance of *treatment* after introducing IPSW? Why do you think weighting is needed? Can you think of another way of dealing with this problem?
   - How about evaluation metrics? Do you observe any significant differences there when compared to simple learners from task 3?
5. **Advanced CATE estimators**
   (a) Choose at least one CATE estimator from EconML package (`https://econml.azurewebsites.net/reference.html#cate-estimators`). Comment on your choice(s).
   (b) Train the estimator(s) on the data.

(c) Predict effects and calculate relevant evaluation metrics.

(d) Report the metrics and comment on your results. How do they compare to previous results?

6. **(Optional) Custom estimator**

- In case you feel constrained about previously discussed methods, you are more than welcome to apply even more advanced methods, whatever you like! Alternatively, you could try to implement one of the EconML's CATE estimators from scratch on your own, or even modify them.
- Make sure you thoroughly describe what you are doing here (justify choices, design decisions, etc.).
- How is your estimator different or better than the ones from previous tasks?

4 Dataset descriptions

We are going to use two types of datasets in this project. IHDP is a popular benchmark example where all outcomes are simulated for the purpose of evaluating CI estimators. JOBS, on the other hand, is arguably closer to real life conditions as it is a mixture of experimental and observational data. The key difference between those two is access to counterfactuals, hence the need to handle them differently. The datasets can be downloaded from here: https://github.com/dmachlanski/CE888_2022/tree/main/project/data.

4.1 IHDP

The Infant Health Development Program (IHDP) dataset was collected to investigate the effect of high-quality childcare and home visits on the future cognitive test score of low-birth-weight, premature infants. The dataset contains 25 features, including measurements about the child (e.g., child-birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex...) and information about the mother at the time she gave birth (e.g., age, marital status, educational attainment, whether she worked) and her behaviours during the pregnancy (e.g., whether she smoked cigarettes, drank alcohol, took drugs...). These are background variables X. The *treatment* variable (t) indicates whether a family was part of the control (i.e., $t = 0$, no support was provided) or the treatment (i.e., $t = 1$, support was provided) group. The *outcome* column records the cognitive test score for the child. The dataset was introduced in [1] based on a clinical trial [2]. We use a semi-synthetic version of the data, where the outcomes (both factual and counterfactual) are simulated (with some added random noise) based on real pre-treatment covariates. For this reason, the dataset also includes true (noiseless) individualised effects per each data unit, which are better suited for performance evaluation than the outcomes due to lack of noise. The use of counterfactuals/true effects is forbidden in the training stage (evaluation only).

4.2 JOBS

This dataset, proposed by [3], is a combination of the experiment done by [4] as part of the National Supported Work Program (NSWP) and observational data from the Panel Study of Income Dynamics (PSID) [5]. Overall, the data captures people's basic characteristics via 17 variables about their background, whether they received job training from NSWP (*treatment*), and their employment status (*outcome*). Information about whether a sample comes from experimental or observational data is recorded under column *'e'*, though all records should be used for modelling.

5 Technical details

5.1 Inverse Propensity Score Weighting (IPSW)

The *propensity score* is the probability that an individual gets assigned to the treatment group, given their observable feature set, that is, $e(x) = P(t = 1|x)$. One way to implement inverse propensity score weighting is to weigh each sample by the inverse of the propensity score. More formally, a sample weight w_i for unit i can be obtained as:

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

5.2 Formal definitions

Let us start with a *potential outcome* defined as $\mathcal{Y}_t^{(i)}$, which is the observed outcome when individual i receives treatment t. Given this, the Individual Treatment Effect (ITE) can be written as:

$$ITE_i = \mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)} \tag{2}$$

Whereas the Average Treatment Effect (ATE) can be defined as:

$$ATE = \mathbb{E}\left[\mathcal{Y}_1 - \mathcal{Y}_0\right] \tag{3}$$

Where $\mathbb{E}[.]$ denotes mathematical expectation.

Given a set of treated subjects T that are part of sample E coming from an experimental study, and a set of control group C, we define the true Average Treatment effect on the Treated (ATT) as:

$$ATT = \frac{1}{|T|}\sum_{i \in T}\mathcal{Y}^{(i)} - \frac{1}{|C \cap E|}\sum_{i \in C \cap E}\mathcal{Y}^{(i)} \tag{4}$$

5.3 Metrics

The metrics presented here focus on measuring prediction errors. The assumed format is to denote them as ϵ_X, which translates to the amount of error made with respect to prediction type X (lower is better). In terms of treatment outcomes, $\mathcal{Y}_t^{(i)}$ and $\hat{y}_t^{(i)}$ denote true and predicted outcomes respectively for treatment t and individual i. Thus, following the definition of ITE (Eq. (2)), the difference $\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}$ gives a **true** effect, whereas $\hat{y}_1^{(i)} - \hat{y}_0^{(i)}$ a **predicted** one. Following this, we can define Precision in Estimation of Heterogeneous Effect (PEHE), which is the root mean squared error between predicted and true effects:

$$\epsilon_{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)} - (\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)}))^2} \tag{5}$$

Following the definition of ATE (Eq. (3)), we measure the error on ATE as the absolute difference between predicted and true average effects, formally written as:

$$\epsilon_{ATE} = \left|\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)}) - \frac{1}{n}\sum_{i=1}^{n}(\mathcal{Y}_1^{(i)} - \mathcal{Y}_0^{(i)})\right| \tag{6}$$

Following Eq. (4), the error on ATT is defined as the absolute difference between the true and predicted ATT:

$$\epsilon_{ATT} = \left|ATT - \frac{1}{|T|}\sum_{i \in T}(\hat{y}_1^{(i)} - \hat{y}_0^{(i)})\right| \tag{7}$$

Define policy risk as:

$$\mathcal{R}_{pol} = 1 - \left(\mathbb{E}\left[\mathcal{Y}_1|\pi(x) = 1\right]\mathcal{P}(\pi(x) = 1) + \mathbb{E}\left[\mathcal{Y}_0|\pi(x) = 0\right]\mathcal{P}(\pi(x) = 0)\right) \tag{8}$$

Where $\mathbb{E}[.]$ denotes mathematical expectation and policy π becomes $\pi(x) = 1$ if $\hat{y}_1 - \hat{y}_0 > 0$; $\pi(x) = 0$ otherwise. Note policy risk is often measured based on experimental samples only for better precision.

An example implementation of all the four metrics can be found here: https://github.com/dmachlanski/CE888_2022/blob/main/project/metrics.py.

6 Useful resources

- Recent causal inference surveys [6, 7].
- EconML package: https://econml.azurewebsites.net/index.html.
  - Specifically methods assuming no hidden confounders: https://econml.azurewebsites.net/spec/estimation.html.
- Popular books on causality: https://www.bradyneal.com/which-causal-inference-book.
- Introduction to Causal Inference online course: https://www.bradyneal.com/causal-inference-course.

References

[1] J. L. Hill, "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, vol. 20, pp. 217–240, Jan. 2011.

[2] J. Brooks-Gunn, F. R. Liaw, and P. K. Klebanov, "Effects of early intervention on cognitive function of low birth weight preterm infants," *The Journal of Pediatrics*, vol. 120, pp. 350–359, Mar. 1992.

[3] J. A. Smith and P. E. Todd, "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, vol. 125, no. 1-2, pp. 305–353, 2005.

[4] R. J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, vol. 76, no. 4, pp. 604–620, 1986.

[5] R. H. Dehejia and S. Wahba, "Propensity Score-Matching Methods For Nonexperimental Causal Studies," *The Review of Economics and Statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[6] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A Survey of Learning Causality with Data: Problems and Methods," *ACM Computing Surveys*, vol. 53, pp. 75:1–75:37, July 2020.

[7] 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.