AAAI 2023 Tutorial

Machine Learning for Causal Inference

https://aaai23causalinference.github.io/

Zhixuan Chu[1], Jing Ma[2], Jundong Li[2], Sheng Li[2]
[1] Ant Group, Hangzhou, China
[2] University of Virginia, Charlottesville, USA

Tuesday, February 7, 2023

Agenda

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causal Inference-aided Machine Learning
5. Causal Inference Applications
6. Future Directions

1. **Background on Causal Inference**
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causal Inference-aided Machine Learning
5. Causal Inference Applications
6. Future Directions

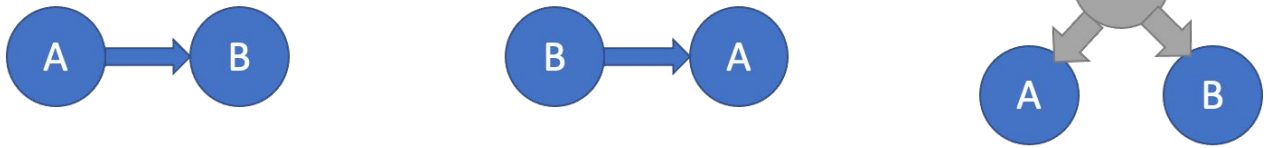
Causality

- *Causality* is also referred to as “causation”, or “cause and effect”
- Causality has been extensively discussed in many fields, such as statistics, philosophy, psychology, economics, education, and health care.

Figure: https://www.pinterest.com/pin/195906652514487365/

Correlation and Causation

- **Correlation does not imply causation**
- For two correlated events A and B, the possible relations might be: (1) A causes B, (2) B causes A, (3) A and B are consequences of a common cause, but do not cause each other, etc.



- Example of (3): As ice cream sales increase, the rate of drowning deaths increases sharply. The two events are correlated. However, increasing ice cream consumption and drowning deaths may not have causal relationships.

Causal Inference

- **Causal inference** is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
- Two major tasks in causal inference
  - **Treatment Effect Estimation (This Tutorial)**: estimate the causal effects of an intervention on subjects, e.g., the effects of medication
  - **Causal Discovery**: infer causal structure from data, i.e., finding causal relations among variables

Experimental Study vs. Observational Study

- **Experimental Study**
  - Randomized Controlled Trial (RCT)
  - Assignment of control/treated is random
  - Gold-standard for studying causal relationships
  - Expensive and time-consuming, e.g., A/B testing

- **Observational Study**
  - Assignment is NOT random
  - Approaches: structural causal models, potential outcome framework
  - Simple, efficient and interpretable, e.g., nearest neighbor matching

About This Tutorial

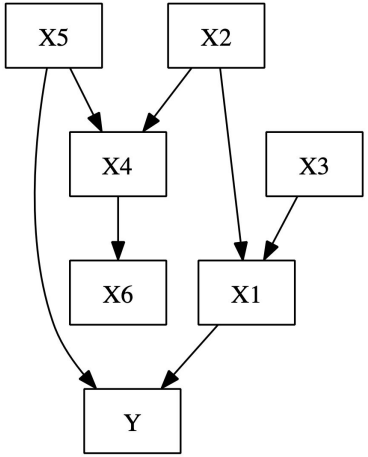
- Causal inference is an active research area with many research topics, this tutorial mainly focuses on the potential outcome framework in observational study
- Machine learning could assist causal inference at different stages. In this tutorial, we focus on how to design representation learning methods and graph neural networks for causal inference. Moreover, we will discuss causality-aided machine learning.

About This Tutorial

- **Schedule**
  - 8:30 AM - 9:15 AM: Background on Causal Inference (S. Li)
  - 9:15 AM - 10:00 AM: Representation Learning based Methods (Chu / S. Li)
  - 10:00 AM - 10:30 AM: Coffee Break
  - 10:30 AM - 11:15 AM: Graph Neural Networks based Methods (J. Li / Ma)
  - 11:15 AM - 11:45 AM: Causality-aided Machine Learning (Chu / Ma)
  - 11:45 AM - 12:10 AM: Causal Inference Applications (TBD)
  - 12:10 AM - 12:30 AM: Future Directions and Closing Remarks (S. Li / J. Li)
- **Website:** https://aaai23causalinference.github.io/

Causal Inference Paradigms

- **Graphical Causal Models**
  - Causal graphs are probabilistic graphical models to encode assumptions about data-generating process *[Pearl, 2009]*
  - Related approach: structural equation modeling (SEM)
- **Potential Outcome Framework**
  - An approach to the statistical analysis of cause and effect based on the potential outcomes *[Rubin, 2005]*
  - Also known as Rubin causal model (RCM), or Neyman–Rubin causal model



An Example of Causal Graph

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.
[Rubin, 2005] Donald Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2005.

Potential Outcome Framework (1)

- **Unit**: A unit is the atomic research object in the causal study
- **Treatment**: An action that applies to a unit

  In the binary treatment case (i.e., $W = 0$ or 1), *treated* group contains units received treatment $W = 1$, while *control* group contains units received treatment $W = 0$
- **Outcome**: response of units after treatment/control, denoted as Y
- **Treatment Effect**: The change of outcome when applying the different treatments on the units

An Illustrative Example

- **Task:** Evaluate the treatment effects of several different medications for one disease, by exploiting the observational data, such as the electronic health records (EHR)
- **Observational data** may include: (1) demographic information of patients, (2) specific medication with the specific dosage taken by patients, and (3) the outcome of medical tests
- **Units**: patients
- **Treatments**: different medications
- **Outcome**: recovery, blood test results, or others

Potential Outcome Framework (2)

- **Potential Outcome**: For each unit-treatment pair, the outcome of that treatment when applied on that unit is the potential outcome. $Y(W=w)$
- **Observed Outcome**: Outcome of treatment that is actually applied. In binary case,
- **Counterfactual Outcome**: Potential outcome of the treatments that the unit had not taken. In binary case,
- A unit can only take one treatment. Thus, counterfactual outcomes are not observed, leading to the well-known "*missing data*" problem

Potential Outcome Framework (3)

- **Treatment Effects** can be defined at the population, treated group, subgroup and individual levels
- *Population Level*: Average Treatment Effect (**ATE**)

$$\text{ATE} = \mathbb{E}[\mathbf{Y}(W=1) - \mathbf{Y}(W=0)]$$

- *Treated group*: Average Treatment Effect on the Treated Group (**ATT**)

$$\text{ATT} = \mathbb{E}[\mathbf{Y}(W=1)|\mathbf{W}=1] - \mathbb{E}[\mathbf{Y}(W=0)|\mathbf{W}=1]$$

- *Subgroup*: Conditional Average Treatment Effect (**CATE**)

$$\text{CATE} = \mathbb{E}[\mathbf{Y}(W=1)|X=x] - \mathbb{E}[\mathbf{Y}(W=0)|X=x]$$

- *Individual*: Individual Treatment Effect (**ITE**)

$$\text{ITE}_i = Y_i(W=1) - Y_i(W=0)$$

Assumptions

- **Assumption 1**: Stable Unit Treatment Value Assumption (**SUTVA**)

  The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

- This assumption emphasizes that:
  - **Independence of each unit**, i.e., there are no interactions between units. In our example, one patient's outcome will not affect other patients' outcomes
  - **Single version for each treatment**. For instance, one medicine with different dosages are different treatments under the SUTVA assumption

Assumptions

- **<u>Assumption 2</u>**: **Ignorability**

  *Given the background variable, **X**, treatment assignment W is independent of the potential outcomes, i.e.,*

- Following our example, this assumption implies that:
  - If two patients have the same background variable ***X***, their potential outcome should be the **same** whatever the treatment assignment is.
  - Analogously, if two patients have the same background variable value, their ***treatment assignment mechanism*** should be same whatever the value of potential outcomes they have

Assumptions

- **<u>Assumption 3</u>: Positivity**

  *For any set of values of **X**, treatment assignment is not deterministic:*

  $$P(W = w|X = X) > 0 \quad \forall w \text{ and } x.$$

- If treatment assignment for some values of X is deterministic, the outcomes of at least one treatment could never be observed. It would be unable and meaningless to estimate causal effects
- It implies “common support” or “overlap” of treated and control groups
- The ignorability and the positivity assumptions together are also called ***Strong Ignorability*** or ***Strongly Ignorable Treatment Assignment***

A Naive Solution

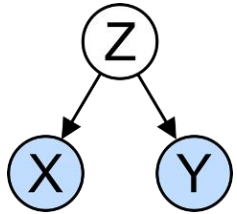
- The core problem in causal inference is: how to estimate the average potential treated/control outcomes over a specific group?
- One naive solution is to calculate the difference between the average treated and control outcomes, i.e.,

$$\hat{\mathrm{ATE}} = \frac{1}{N_T}\sum_{i=1}^{N_T} Y_i^F - \frac{1}{N_C}\sum_{i=1}^{N_C} Y_j^F$$

- However, this solution is not reasonable due to the existence of confounders

Confounders

- **Confounders**: Variables that affect both treatment assignment and outcome
- In the medical example, **age** is a confounder
  - Age would affect the recovery rate
  - Age would also affect the treatment choice



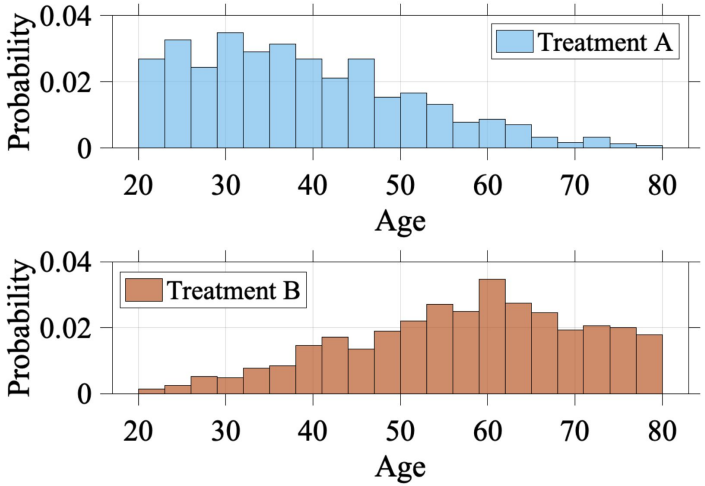
Recovery Rate / Treatment / Age	Treatment A	Treatment B
Young	234/270 = 87%	81/87 = **92%**
Older	55/80 = 69%	192/263 = **73%**
Overall	289/350 = **83%**	273/350 = 78%

Table 1. An example to show the spurious effect of confounder variable *Age*.

Simpson's paradox due to confounder

Selection Bias

- **Selection Bias**: The distribution of the observed group is not representative to the group we are interested in
- Confounder variables affect units' treatment choices, leading to selection bias
- Selection bias makes counterfactual outcome estimation more difficult



Classical Causal Inference Methods

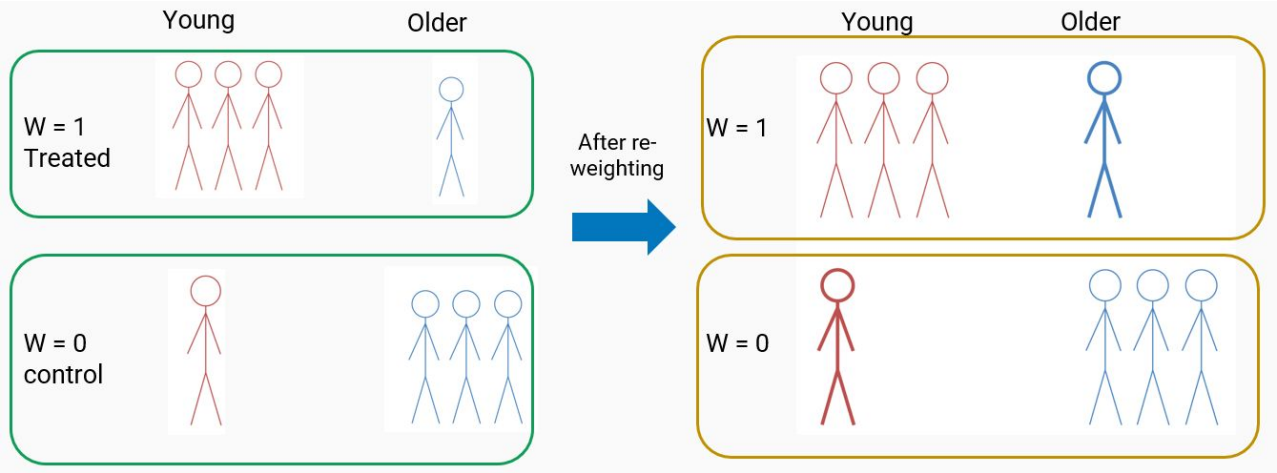
- Causal inference has been an active research area in statistics in the past several decades
- Categorization of Classical Methods
  - Re-weighting methods
  - Stratification methods
  - Matching methods
  - Tree-based methods

Re-weighting Methods

- **Challenge of Selection Bias**: due to different distributions of treated and control groups
- Sample re-weighting is a simple way to overcome the selection bias problem
- **Key Idea**: By *assigning appropriate weight to each sample* in the observation dataset, a ***pseudo-population*** is created on which the distributions of the treated group and control group are similar

Sample Re-weighting Methods

- Intuition example: **Age** (Yong/older) as the confounder
  - Young people: 75% chance of receiving treatment
  - Older people: only a 25% chance of receiving treatment



http://www.rebeccabarter.com/blog/2017-07-05-ip-weighting/

Stratification Methods

- **Stratification** adjusts the selection bias by splitting the entire group into subgroups, where within each subgroup, the treated group and the control group are similar under some measurements
- Stratification is also named as ***subclassification*** or ***blocking***
- ATE for stratification is estimated as

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j)\right]$$

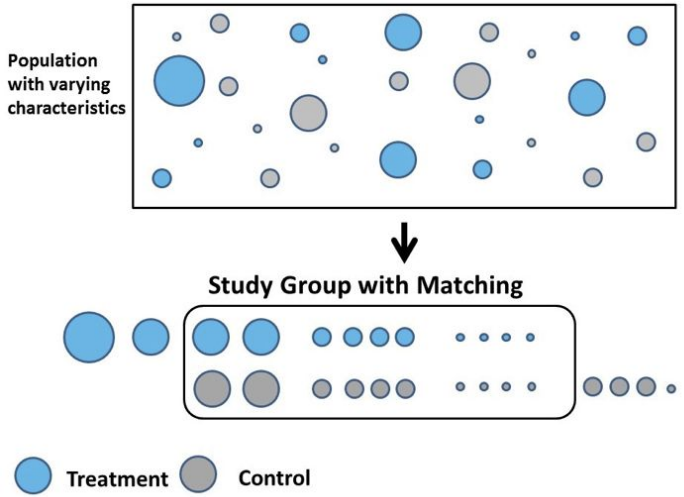
Matching Methods

- Matching methods estimate the counterfactuals and meanwhile reduce the estimation bias brought by the confounders
- Potential outcomes of the i-th unit estimated by matching are:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 1; \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1; \end{cases}$$

where *J(i)* is the matched neighbors of unit i in the opposite treatment group



Propensity Score Matching (PSM)

- Propensity scores denote conditional probability of assignment to a particular treatment given a vector of observed covariates.

$$e(x) = Pr(W = 1|X = x)$$

- Based on propensity scores, the distance between two units is

$$D(\mathbf{x}_i, \mathbf{x}_j) = |e_i - e_j|$$

- Alternatively, linear propensity score based distance metric

$$D(\mathbf{x}_i, \mathbf{x}_j) = |\text{logit}(e_i) - \text{logit}(e_j)|$$

P. Rosenbaum, and D. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

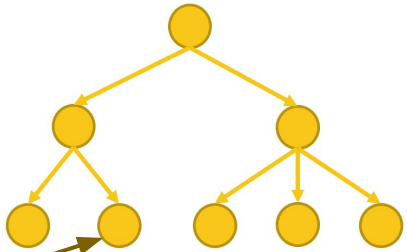
Tree-based Methods

- **Classification And Regression Trees (CART)**
  - Recursively partition the data space
  - Fit a simple prediction model for each partition
  - Represent every partitioning as a decision tree
- Leaf specific effect:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi) \Big]$$

$$\tau(x \mid \Pi) \equiv \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

A specific leaf node



P. Rosenbaum, and D. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

Causal Forest

- Single tree is noisy -> using forest
- Forests = nearest neighbor methods + adaptive neighborhood metric
  - k-nearest neighbors: seek the k closest points to x according to some prespecified distance measure
  - Tree-based methods: closeness is defined with respect to a decision tree, and the closest points to x are those that fall in the same leaf

S. Wager, and S. Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113.523 (2018): 1228-1242.

Why ML is Helpful for Causal Inference?

- **Machine Learning**
  - Various learning tasks, e.g., regression, classification, clustering
  - Various settings: multi-view, multi-task, transfer learning, etc.
  - Feature learning by shallow and deep models
- **Connections between Causal Inference and Machine Learning**
  - Matching in representation space
  - Covariate shift and group balancing
  - Counterfactual inference could be modeled as a regression problem

1. Background on Causal Inference
2. **Representation Learning based Methods**
3. Graph Neural Networks based Methods
4. Causal Inference-aided Machine Learning
5. Causal Inference Applications
6. Future Directions

1. Background on Causal Inference
2. Representation Learning based Methods
3. **Graph Neural Networks based Methods**
4. Causal Inference-aided Machine Learning
5. Causal Inference Applications
6. Future Directions

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. **Causal Inference-aided Machine Learning**
5. Causal Inference Applications
6. Future Directions

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. **Causal Inference-aided Machine Learning**
5. Causal Inference Applications
6. Future Directions

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causal Inference-aided Machine Learning
5. **Causal Inference Applications**
6. Future Directions

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causal Inference-aided Machine Learning
5. Causal Inference Applications
6. **Future Directions**

Resources

Thank you!