Bridging the Gap:
Connect Causal Inference to Machine Learning

Virtually @THU via Zoom

Ruocheng Guo（郭若城）

Data Mining and Machine Learning Laboratory

Arizona State University

rguo12@asu.edu

Outline

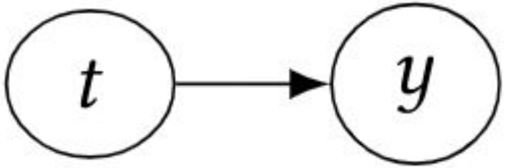
- Introduction to causal inference
- Connection to machine learning and data mining
  - Using machine learning for causal inference
  - Causality-aware machine learning

Introduction

What is causality?

- A definition with random variables
  - Given two random variables t (treatment) and y (outcome), we say t causes y iff changing the value of t would cause a change in the value of y.
- In Pearl's structural causal models (SCMs)

The causal graph



The structural equations

$$t = f_t(\boxed{\epsilon_t}),$$

$$y = f_y(t, \boxed{\epsilon_y}).$$

Noise terms: capture unobserved information.

Read as "y is generated by a function of t and noise"

Causal Inference and Machine Learning

Using machine learning methods for causal inference

- **Problems related to causal effects**
- Causal discovery

Causality-aware ML

- **Causal representation learning (e.g., invariant risk minimization)**
- **Unbiased interactive ML: learning to rank/recommendation/machine translation**

Using machine learning methods for causal inference

- **Problems related to causal effects**

Introduction

Why do we care about causal effects?

- They are crucial for decision making
  - A/B tests in tech companies
  - Clinical trials performed by FDA

Why do we study observational data?

- Convenient to collect
  - Do not need randomized controlled trials
- Rich auxiliary information: network, text and image etc.
- In ML/DM, most datasets are observational.

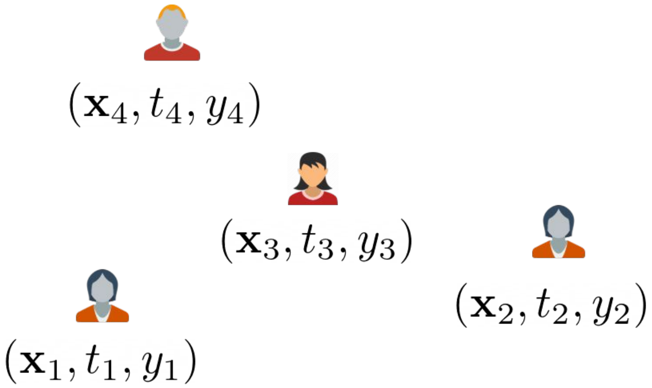
Introduction

Observational data $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^{N}$

$\mathbf{x}_i$ - feature vector of an instance

t_i - binary observed treatment of an instance

y_i - an observed factual outcome of an instance



$$
\begin{aligned}
t = 1 &: \text{ take medicine} \\
t = 0 &: \text{ take no medicine} \\
y = 1 &: \text{ good health outcome} \\
y = 0 &: \text{ bad health outcome}
\end{aligned}
$$

The Challenge

With observational data, what can we estimate?

Probabilistic quantities: joint, conditional and marginal distributions of observed variables.

Causal effect

- In potential outcome framework
  - Potential outcomes $y_i^t, t \in \{0, 1\}$
  - Individual treatment effect (ITE) $\tau_i = y_i^1 - y_i^0$
  - Conditional average treatment effect (CATE) $E[\tau|\mathbf{x}]$
  - Average treatment effect (ATE) $E[\tau]$
  - Not directly estimable from data

Causal Identification

Causal Identification

- With causal assumptions, we can identify causal effects by writing them as functions of probabilistic quantities.

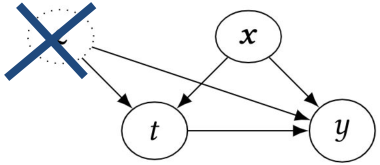
Causal Identification

Strong ignorability

- It assumes that
  - all the **confounders** have been measured as the observed features **x**,
  - each instance's probability to receive the treatment (the true propensity score) is between 0 and 1.
- In the potential outcome framework
  - As a conditional independence

$$y^1, y^0 \perp t|\mathbf{x}$$

- In a causal graph



- How it works in identifying CATE/ITE

$$\begin{aligned} E[\tau|\mathbf{x}] &= E[y^1 - y^0|\mathbf{x}] \\ &= E[y^1|\mathbf{x}] - E[y^0|\mathbf{x}] \\ &= [y^1|\mathbf{x}, t=1] - [y^0|\mathbf{x}, t=0] \\ &= \boxed{E[y|\mathbf{x}, t=1]} - \boxed{E[y|\mathbf{x}, t=0]}, \end{aligned}$$

Can be estimated!

Causal Identification

Strong ignorability can be untenable given observational data

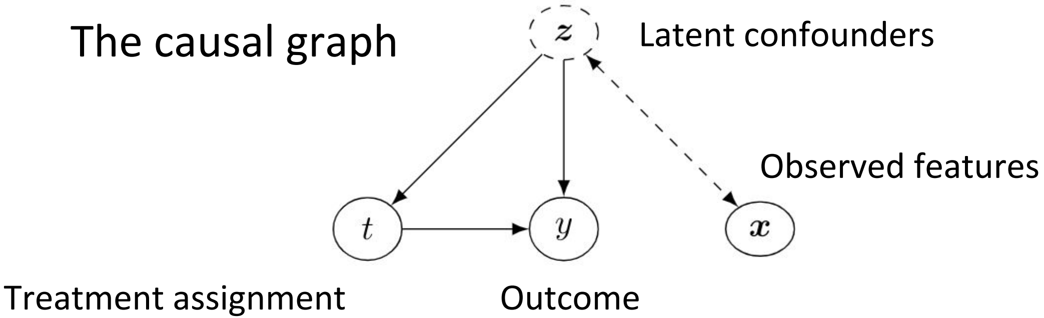
- There can exist hidden confounders (e.g., socio-economic status)
- Using strong ignorability can lead to **confounding bias**.

Relax the strong ignorability assumption with latent confounders **z** [1].

As conditional independence

$$y^1, y^0 \perp t | \mathbf{z}$$

The causal graph



- Latent confounders **z** are not observable(s), we only assume their existence.
- We can learn **z** from data via machine learning models.

[1] Kuroki, Manabu, and Judea Pearl. "Measurement bias and effect restoration in causal inference." Biometrika 101, no. 2 (2014): 423-437.

Using machine learning for causal inference

With causal effects identified, causal effect estimation is a regression problem.

Some recently published papers with i.i.d. observational data:

Neural network methods with strong ignorability

- CFRNet [1], SITE [5] and GANITE [6]

Neural network methods learning latent confounders

- CEVAE [2]

Ensembles of trees that also rely on strong ignorability

- BART [3] and Causal Forest [4]

More to find https://github.com/rguo12/awesome-causality-algorithms or in our survey [7]

[1] Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." In ICML, 2016.
[2] Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." In NeurIPS, 2017.
[3] Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics.* 2011.
[4] Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *JASA.* 2018.
[5] Yao, Liuyi, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. "Representation learning for treatment effect estimation from observational data." In NeurIPS, 2018.
[6] Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets." In ICLR, 2018.
[7] Guo, Ruocheng, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. "A survey of learning causality with data: Problems and methods." in ACM CSUR (2020).

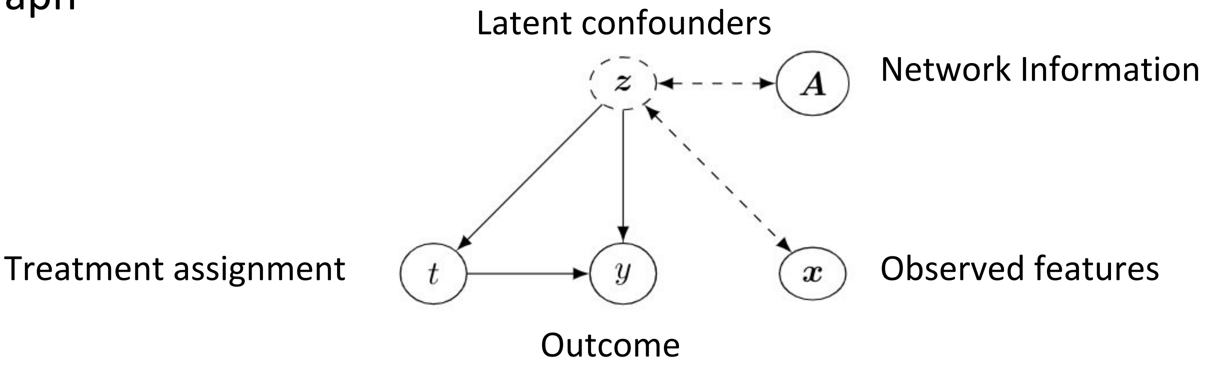
Causal Identification

What if there exists network information?

We propose to use **network information** along with observed features to improve the learned latent confounders.

- Network information can compensate for hidden confounders.
- *Homophily*: similar individuals are more likely to connect with each other.

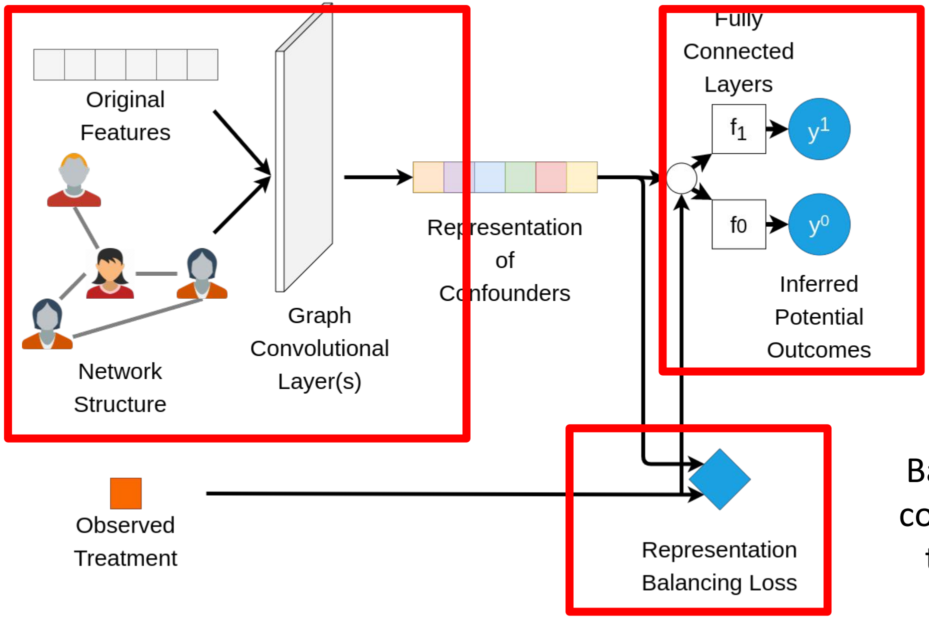
The causal graph



Using machine learning for causal inference

- Learning individual-level causal effects from networked observational data [1]

Learning latent confounders w. GCN



Outcome inference with fully connected layers

Balance the latent confounders of the treated and the controlled

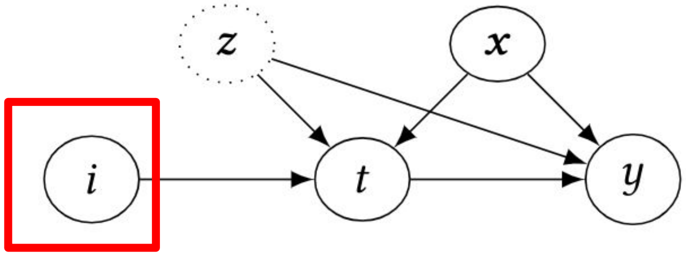
- Extend to counterfactual evaluation of treatment assignments in network data [2]

[1] Guo, Ruocheng, Jundong Li, and Huan Liu. "Learning Individual Causal Effects from Networked Observational Data." WSDM 2020.
[2] Guo, Ruocheng, Jundong Li, and Huan Liu. "Counterfactual Evaluation of Treatment Assignment Functions with Networked Observational Data." SDM 2020.

Causal Identification

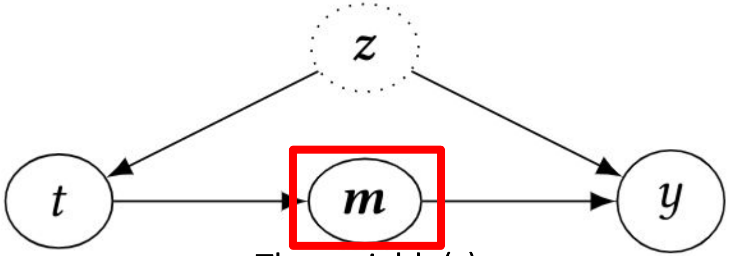
- Causal graphs are powerful tools to represent causal assumptions (conditional independence) for identification.
- When we observe some special variables:

Identification with Instrumental variables



Instrumental variable

Identification with Front-door criterion



The variable(s) satisfying the front-door criterion

- Automatic discovery of identification strategies? Still an open problem.

Causality-aware ML

- Learning causal representation (invariant risk minimization)
- Unbiased interactive ML

Causality-aware ML

When we have prior causal knowledge of the data:
We can impose various causal constraints in the objective of ML algorithms [1].

Data: from multiple (n_e) training environments $D_e := \{(x_i^e, y_i^e)\}_{i=1}^{n_e}$
Task: predict y from the two features (x1,x2), generalize to different environments.

- Suppose data generated by the SCM:

$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

- Fit linear regressions

regress from X_1^e, to obtain $\boxed{\hat{\alpha}_1 = 1}$ and $\hat{\alpha}_2 = 0$,
regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\boxed{\hat{\alpha}_2 = \sigma(e)/(\sigma(e) + \frac{1}{2})}$
regress from (X_1^e, X_2^e), to obtain $\boxed{\hat{\alpha}_1 = 1/(\sigma(e) + 1) \text{ and } \hat{\alpha}_2 = \sigma(e)/(\sigma(e) + 1).}$

Causal relation. Invariant to environment.

Spurious correlations. Sensitive to environment.

- Two environments:

$$\mathcal{E}_{tr} = \{\text{replace } \sigma^2 \text{ by } 10, \ \text{replace } \sigma^2 \text{ by } 20\}.$$

[1] Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. "Invariant risk minimization." arXiv preprint arXiv:1907.02893 (2019).

Causality-aware ML

What if we do not know the SCM?

$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

Several heuristic hypotheses:

- Learning domain-invariant representations (domain adaptation)
  - $P(\Phi(X^e)) = P(\Phi(X^{e'}))$
  - Fails when $P(Y^e) \neq P(Y^{e'})$
- Causal prediction [1]
  - Representations s.t. residual distributions across environments are the same.
  - Fails when noise variance in Y changes with environment.
- Invariant risk minimization [2]
  - $E[Y^e|\Phi(X^e) = h] = E[Y^{e'}|\Phi(X^{e'}) = h], \forall h \in \{\Phi(X^e)\} \cap \{\Phi(X^{e'})\}$
  - Fails when $h = \emptyset$

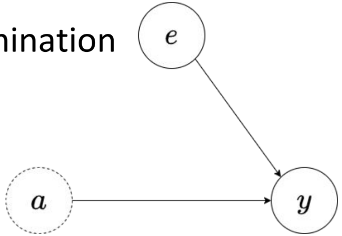
[1] Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78, no. 5 (2016): 947-1012.
[2] Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. "Invariant risk minimization." arXiv preprint arXiv:1907.02893 (2019).

Unbiased Interactive ML

Use Implicit feedback (click/purchase) as labels

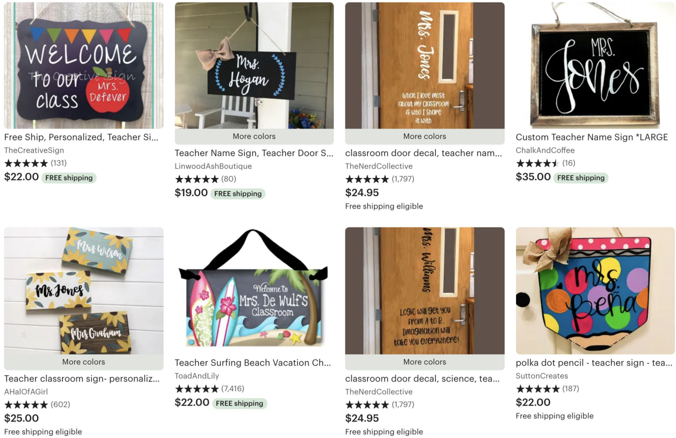
- User has to examine the prediction of the logging ML algorithm to provide label
- E.g., Product search in e-commerce

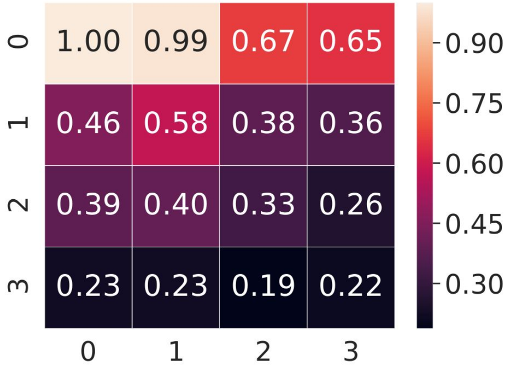
Examination



The user likes it

Positive feedback





Normalized feedback over positions

Unbiased Interactive ML

A general problem in interactive ML systems

- Search (position bias) [1]
- Recommendation (popularity bias, sampling bias of the logging policy) [2,3]
- Machine translation (sampling bias of the logging policy)[4]

[1] Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased learning-to-rank with biased feedback." In WSDM, 2017.
[2] Yang, Longqi, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. "Unbiased offline recommender evaluation for missing-not-at-random implicit feedback." In Recsys, 2018.
[3] Chen, Minmin, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. "Top-k off-policy correction for a REINFORCE recommender system." In WSDM, 2019.
[3] Lawrence, Carolin, Artem Sokolov, and Stefan Riezler. "Counterfactual Learning from Bandit Feedback under Deterministic Logging: A Case Study in Statistical Machine Translation." In EMNLP, 2017.

Conclusion

- ML can help causal inference.
- Causal knowledge can help ML algorithms.

Survey paper:

Guo, Ruocheng, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. "A survey of learning causality with data: Problems and methods." in ACM CSUR (2020)

Repository:

https://github.com/rguo12/awesome-causality-algorithms

Q & A