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

$$t = f_t(\epsilon_t),$$
$$y = f_y(t, \epsilon_y).$$

Read as “$y$ is generated by a function of $t$ and noise”

Noise terms: capture unobserved information.
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 \( \{x_i, t_i, y_i\}_{i=1}^{N} \)

- \( x_i \) - feature vector of an instance
- \( t_i \) - binary observed treatment of an instance
- \( y_i \) - an observed factual outcome of an instance

\( (x_4, t_4, y_4) \)

\( (x_3, t_3, y_3) \)

\( (x_2, t_2, y_2) \)

\( (x_1, t_1, y_1) \)

- \( t = 1 \): take medicine
- \( t = 0 \): take no medicine
- \( y = 1 \): good health outcome
- \( y = 0 \): bad health outcome
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) $E[\tau | x]$
  - Average treatment effect (ATE) $E[\tau]$
  - Not directly estimable from data
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

\[ y^1, y^0 \perp t | x \]

- In a causal graph

- How it works in identifying CATE/ITE

\[
E[\tau|x] = E[y^1 - y^0|x] = E[y^1|x] - E[y^0|x] = [y^1|x, t = 1] - [y^0|x, t = 0] = E[y|x, t = 1] - E[y|x, t = 0]
\]
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

$$y^1, y^0 \perp t \mid z$$

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.

---

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

Neural network methods learning latent confounders
- CEVAE [2]

Ensembles of trees that also rely on strong ignorability

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

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]

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

Causal Identification

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

Causality-aware ML

• Learning causal representation (invariant risk minimization)
• Unbiased interactive 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:
  \[ X_1 \leftarrow \text{Gaussian}(0, \sigma^2), \]
  \[ Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2), \]
  \[ X_2 \leftarrow Y + \text{Gaussian}(0, 1). \]

• Data from two environments:
\[ \mathcal{E}_{tr} = \{\text{replace } \sigma^2 \text{ by 10, replace } \sigma^2 \text{ by 20}\}. \]

Causality-aware ML

What if we do not know the SCM?

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 \( h = \emptyset \)

\[\begin{align*}
X_1 &\leftarrow \text{Gaussian}(0, \sigma^2), \\
Y &\leftarrow X_1 + \text{Gaussian}(0, \sigma^2), \\
X_2 &\leftarrow Y + \text{Gaussian}(0, 1).
\end{align*}\]


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
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]

Conclusion

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

Survey paper:

Repository:
https://github.com/rguo12/awesome-causality-algorithms
Q & A

[QUESTIONS]

[ANSWERS]