Connections between Causal Inference and Machine Learning

@ 因果科学与CausalAI读书会

郭若城

数据挖掘与机器学 习实验室(Data Mining and Machine Learning Lab)

亚利桑那州立大学（Arizona State University）

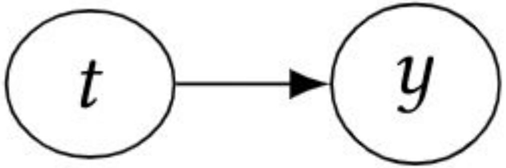
rguo12@asu.edu

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"

The Myth of Causal ML

Causal inference: studies the data generating process (DGP).

ML: design and understand "*curve fitting*" models.

Causal ML:

1. Using ML models to answer questions in the DGP
    a. Bayesian additive regression tree [1]
    b. RL for causal discovery [2]
2. Using the causal model of DGP to design ML models
    a. Invariant Risk Minimization [3]
    b. Unbiased Learning to Rank [4]

[1] Hill, Jennifer L. "Bayesian nonparametric modeling for causal inference." Journal of Computational and Graphical Statistics 20, no. 1 (2011): 217-240.
[2] Zhu, Shengyu, Ignavier Ng, and Zhitang Chen. "Causal Discovery with Reinforcement Learning." In International Conference on Learning Representations. 2019.
[3] Arjovsky, Martin, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. "Invariant risk minimization." arXiv preprint arXiv:1907.02893 (2019).
[4] Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased learning-to-rank with biased feedback." In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 781-789. 2017.

Causal Inference and Machine Learning

Machine learning methods for causal inference

- **Causal effect estimation**

Causality-aware Machine Learning

- **Learning causal features for out-of-distribution generalization**
- **Unbiased interactive ML**

Machine learning methods for causal inference

- **Causal effect estimation**

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

In academia, why do we study observational data?

- They are convenient to collect
- In ML/DM, most datasets are observational.
  - Rich auxiliary information: network, text and image etc.

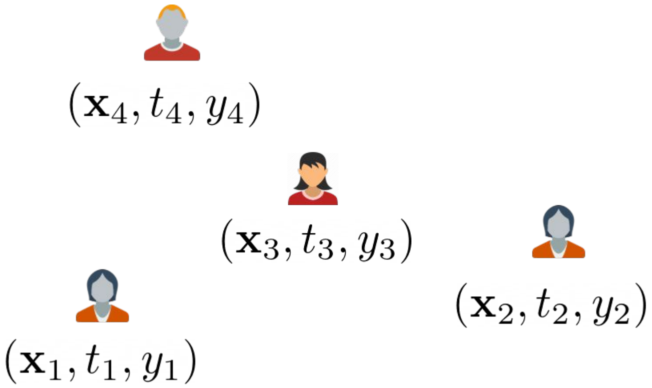
Introduction

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 can not be written as a function of probabilistic quantities directly

- 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 Effects — Identification — Probabilistic Quantities

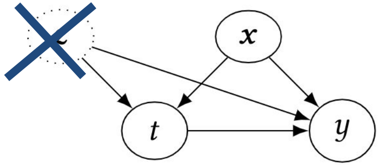
Causal Identification

Unconfoundedness

- It assumes that
  - all the **confounders** have been measured as the observed features **x**,
  - 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

Can be estimated!

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

Causal Identification

Unconfoundedness can be untenable given observational data

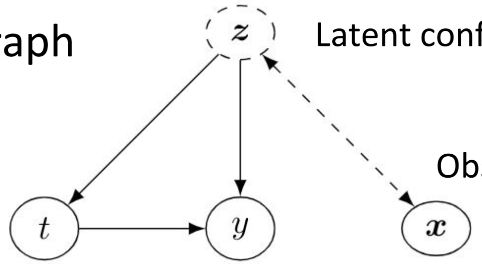
- There can exist hidden confounders (e.g., socio-economic status)
- Using Unconfoundedness can lead to **confounding bias**.

Relax the 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 approximately 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.

Machine learning for causal inference

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

Previous work on i.i.d. observational data:

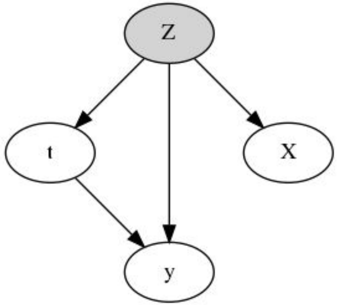
Neural network methods learning latent confounders

- Causal Effect Variational Autoencoder (CEVAE) [1]

When X is text features

- We can extract latent confounders using pretrained language models with **properly designed loss function** [2]

CEVAE



[1] Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." In NeurIPS, 2017.
[2] Veitch, Victor, Dhanya Sridhar, and David M. Blei. "Using text embeddings for causal inference." arXiv preprint arXiv:1905.12741 (2019).

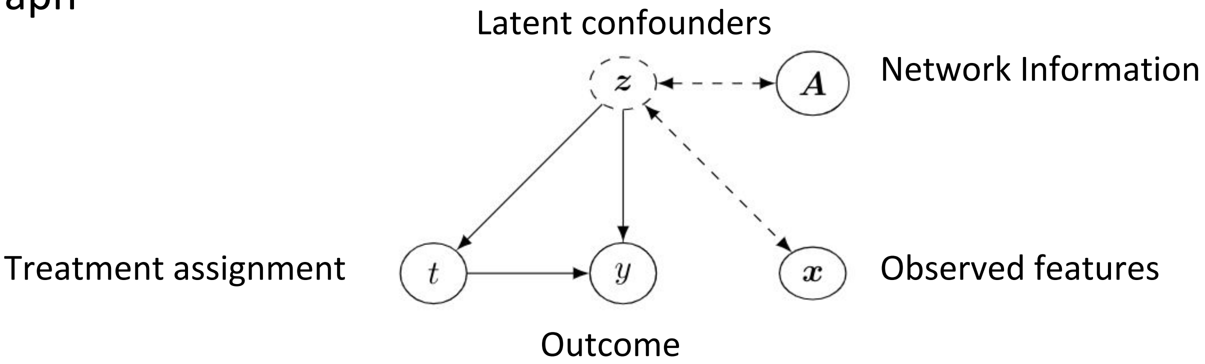
Causal Identification

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

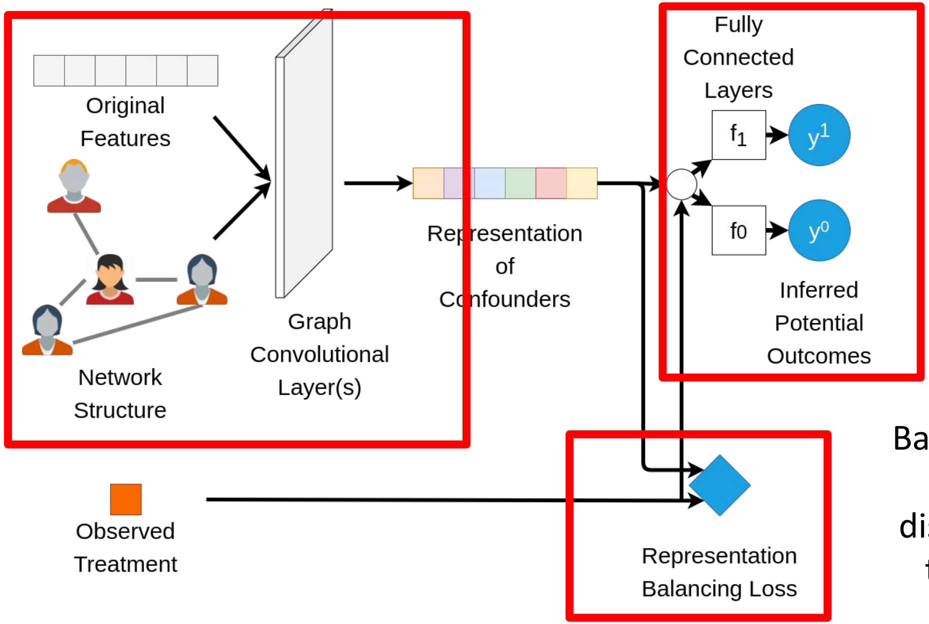


An example of "All models are **wrong**, but some are **useful**."

Machine learning for causal inference

- Learning individual-level causal effects from networked observational data [1,3]

Learning latent confounders w. GCN



Outcome inference with fully connected layers

Balancing the latent confounder distributions 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, Yichuan Li, K. Selçuk Candan, Adrienne Raglin, and Huan Liu. "IGNITE: A Minimax Game Toward Learning Individual Treatment Effects from Networked Observational Data." IJCAI, 2020.

Take away

We can utilize ML models to learn latent confounders from multi-modal data for causal identification.

- Latent confounders relax the unconfoundedness assumption.
- Latent confounders compensate for unobserved confounders using multi-modal data.

Future work

- Can we have theoretical guarantees on (1) the latent variables are actually confounders and (2) the bias of estimates with latent confounders?
- Can we interpret latent confounders?

Causality-aware ML

- Learning causal features for out-of-distribution generalization
- Unbiased interactive ML

Causality-aware ML

- Learning causal features for out-of-distribution generalization
- Unbiased interactive ML

OOD Generalization

A truth generalizes to any data distribution.

- Laws of Physics hold across time and space.

However, discovery of a truth is challenging due to unobserved variables.

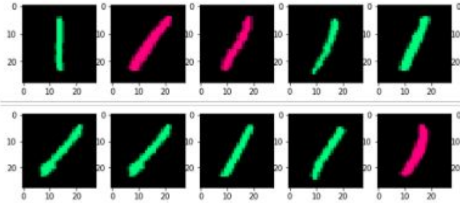
Therefore, practically, we aim to design ML models that are based on **incorrect causal model but can generalize to unseen distributions**.

OOD Generalization

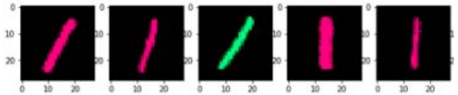
- Deep learning algorithms can fit spurious correlations.
  - Examples

Training environments:
camel -- sand
cow -- grass

Test environment:
camel – grass
cow -- sand



Training environments:
1 -- green



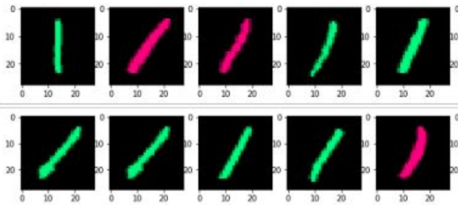
Test environment:
1 -- red

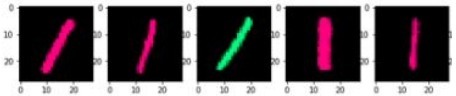
Figures from
Shuxi Zeng, Pengchuan Zhang, Denis Charles, Eren Manavoglu, Emre Kiciman – *Robust Neural Network for Causal Invariant Features Extraction. Neurips 2019 workshop*

OOD Generalization

To test whether a DNN is OOD generalizable, we consider

- Non-i.i.d. training data collected from multiple domains
- Test data domain is different from any training domain





Training environments:
1 -- green

Test environment:
1 -- red

OOD Generalization

- Spurious correlations vary with domains.
- But causal relationships are transferable across different domains.
- Based on them, invariant risk minimization (IRM) [1] is proposed to.
  - Learn representations of causal features
  - Capture the invariant relationships between causal features and labels

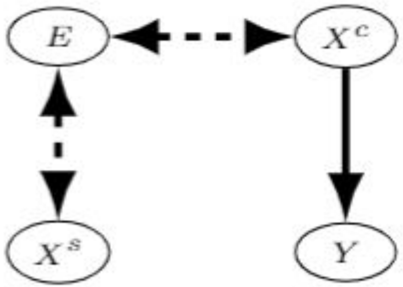


Figure 1: An example DAG showing the causal model in OOD prediction. $P(Y|X^c)$is invariant across domains. The spurious correlation $P(Y|X^s)$ may vary. The thick edges show strong Λ spuriousness.

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

Invariant Risk Minimization

- Initial formulation: difficult to solve

$$\begin{aligned} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \quad & \sum_{e\in\mathcal{E}_{tr}} R^e(w\circ\Phi) \\ \text{subject to} \quad & w\in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi),\ \text{for all } e\in\mathcal{E}_{tr}. \end{aligned} \qquad \text{(IRM)}$$

w Classifier

Φ Feature Representation

- Simplified version:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda\cdot\left\|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\right\|^2, \qquad \text{(IRMv1)}$$

w Dummy Classifier

Φ Logit(s)

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

Invariant Risk Minimization

- Initial formulation: difficult to solve

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{tr}} R^e(w\circ\Phi)$$
$$\text{subject to} \quad w\in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi), \text{ for all } e\in\mathcal{E}_{tr}. \qquad \text{(IRM)}$$

w Classifier

Φ Feature Representation

- Simplified version:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda\cdot \|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\|^2, \qquad \text{(IRMv1)}$$

w Dummy Classifier

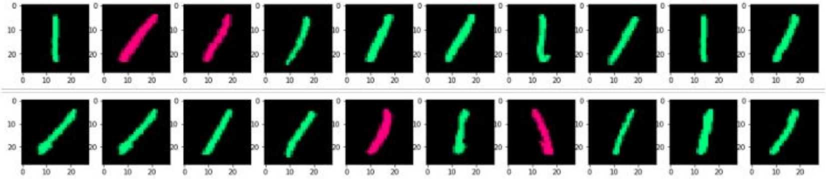
Φ Logit(s)

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

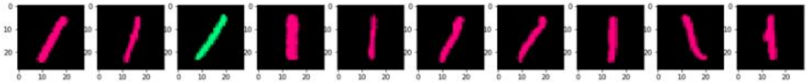
Colored MNIST Experiments

- Labels: Class 1 digits 1-4, Class 2 digits 5-9, 25% random label flipping.
- Environments:

	Class 1	Class 2
Env1 (training)	90% green	10% green
Env2 (training)	80% green	20% green
Env3 (test)	10% green	90% green



Training



Test

Figures taken from
Shuxi Zeng, Pengchuan Zhang, Denis Charles, Eren Manavoglu, Emre Kiciman – *Robust Neural Network for Causal Invariant Features Extraction. Neurips 2019 workshop*

Colored MNIST Experiments

Algorithm	Acc. train envs.	Acc. test env.
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

Table 1: Accuracy (%) of different algorithms on the Colored MNIST synthetic task. ERM fails in the test environment because it relies on spurious color correlations to classify digits. IRM detects that the color has a spurious correlation with the label and thus uses only the digit to predict, obtaining better generalization to the new unseen test environment.

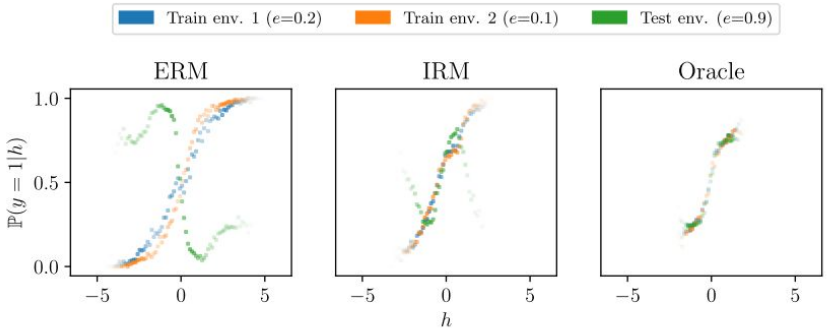


Figure 5: $P(y = 1|h)$ as a function of h for different models trained on Colored MNIST: (left) an ERM-trained model, (center) an IRM-trained model, and (right) an ERM-trained model which only sees grayscale images and therefore is perfectly invariant by construction. IRM learns approximate invariance from data alone and generalizes well to the test environment.

Take away

- IRM learns causal features that have invariant relationships with labels.
- IRM exploits the difference between training domains to learn causal features.

The Limitation of IRM

We find an edge case called **strong Lambda spuriousness** where IRM can be tricked to learn spurious features.

Strong Lambda spuriousness
The spurious correlations among X^s, E and Y are strong

How does IRM fail?
IRM is imposing the conditional independence

$$Y \perp\!\!\!\perp E | F(X)$$

Under **strong Lambda spuriousness**, we find that F(X) = E is a good solution to the regularized optimization problem of IRM since it leads to

- small ERM loss
- satisfies the conditional independence.

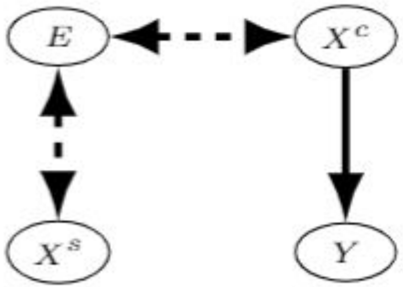


Figure 1: An example DAG showing the causal model in OOD prediction. $P(Y|X^c)$is invariant across domains. The spurious correlation $P(Y|X^s)$ may vary. The thick edges show strong Λ spuriousness.

A Simple and Effective Fix to IRM

The finding in the last slide implies that the IRM constraint is too general.

- Therefore, we should add more constraints.

We find that conditional distribution matching $P(F(X)|Y,E) = P(F(X)|Y)$ can compensate for IRM from two aspects

- It explicitly makes it more difficult to fit F(X) to E.
- If the predictor is an invertible function f, then it is reasonable to push F(X) from the same class to be similar, regardless of E.

Empirical Results

CMNIST+ [1]: a semi-synthetic new dataset with strong Lambda spuriousness.

Table 3: Test accuracy on CMNIST+: by combining CDM with IRM, performance is improved significantly and consistently under strong Λ spuriousness ($\rho \geq 0.8$) compared to IRM.

Method	$\rho = 0.8$	$\rho = 0.85$	$\rho = 0.9$
IRM-MMD (ours)	**52.91%**	**40.83%**	**37.96%**
IRM-ACDM (ours)	**57.23%**	**45.47%**	**42.85%**
IRM	47.01%	37.81%	36.31%
MMD	23.04%	25.22%	24.22%
ACDM	30.41%	29.48%	25.53%
ERM	30.16%	27.83%	24.61%
Oracle	73.10%	73.49%	73.58%

[1] Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix (Under review)

Causality-aware ML

- Learning causal features for out-of-distribution generalization
- Unbiased interactive ML

Unbiased Interactive ML

An overview of how the businesses work using interactive ML algorithms (e.g, recommender system and search ranking systems etc.).

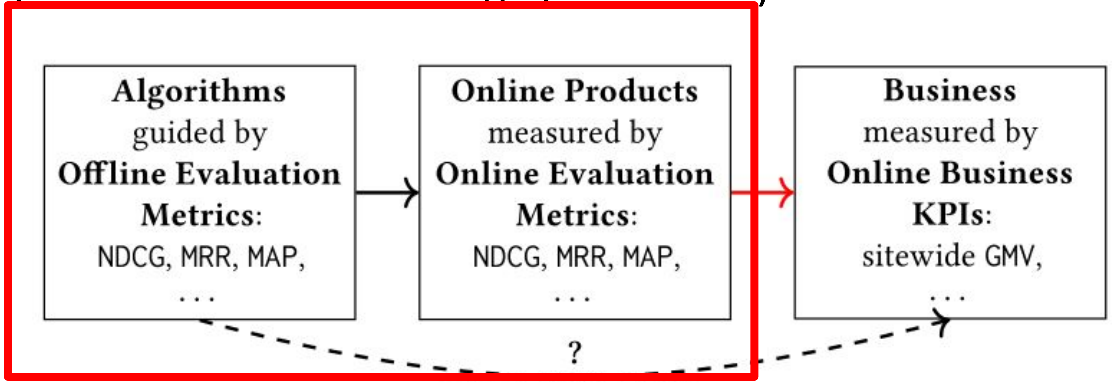


Figure 1: The Causal Path from Algorithms to Business

Figure from
Wang, Zenan, Xuan Yin, Tianbo Li, and Liangjie Hong. "Causal Meta-Mediation Analysis: Inferring Dose-Response Function From Summary Statistics of Many Randomized Experiments." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2625-2635. 2020.

Unbiased Interactive ML

Goal: estimate online metrics with offline datasets

Here, bias describes the difference between offline metrics and online metrics.

Various types of bias in interactive ML systems

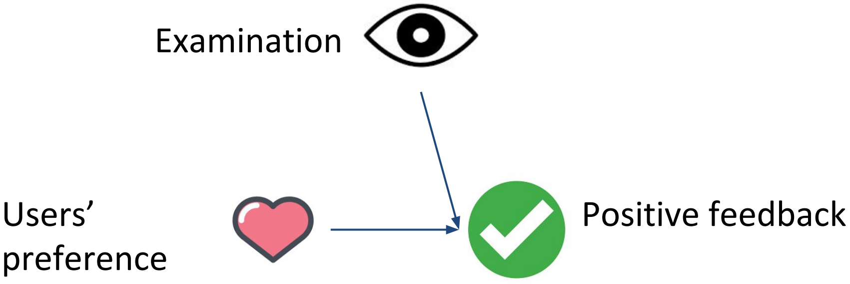
- Learning to rank (position bias) [1]
- Recommendation (popularity bias) [2,3]
- Interactive Machine Translation (sampling bias of the logging policy) [4]

[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.
[4] 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.

Unbiased Learning to Rank

In e-commerce and web search, we use user feedback (click/purchase) as labels.

- Examination Bias
  - A user has to *examine* items to provide labels.
  - Items are displayed based on some existing algorithms (logging policy).
- We need unbiased learning to rank in e-commerce product search and web search:

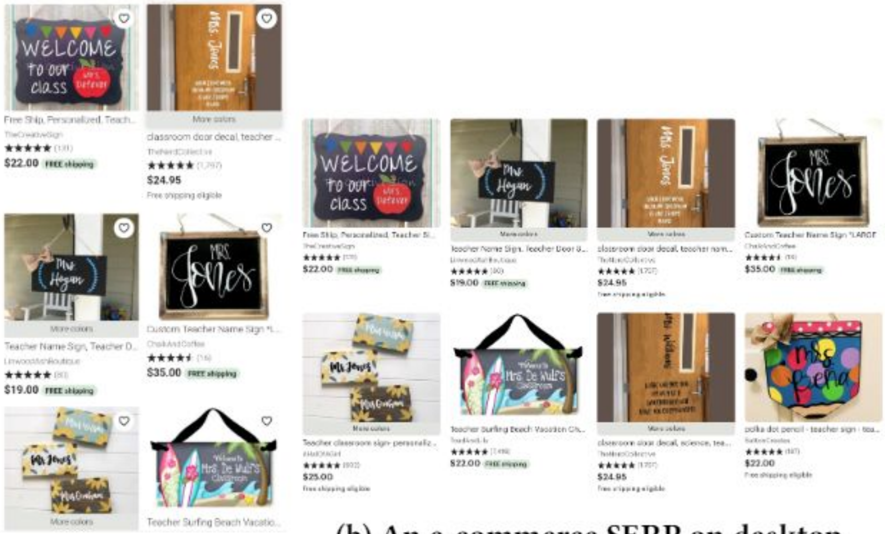


Grid-based Product Search

Previous work focuses on 1-D Web Search.

Compared to 1-D Web Search,
E-commerce Product Search

- Products are shown in 2-D grids.
- With different devices, SERPs are arranged differently.
- They can influence the design of propensity model in unbiased learning to rank.



(a) An e-commerce SERP on mobile device

(b) An e-commerce SERP on desktop

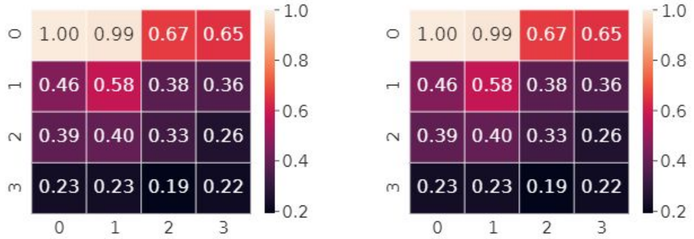
Figure 1: E-commerce SERPs on mobile devices and desktops: products are shown in two-column and four-column grids along with images and meta information.

Guo, Ruocheng, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. "Debiasing Grid-based Product Search in E-commerce." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2852-2860. 2020.

Data Analysis Results

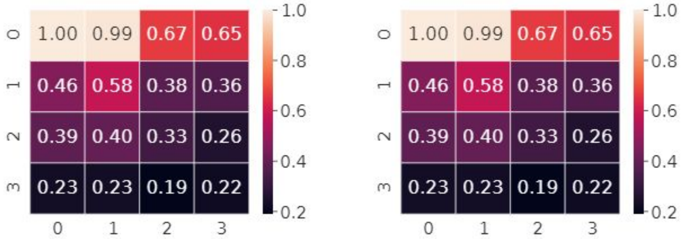
- We aim to verify the existence user behavior patterns observed in [1].
  - Row skipping
  - Slower decay
  - Middle bias
- NCTR and NPR are not monotonically decreasing from top to the bottom.
  - This motivates the usage of row skipping model.
- NCTR and NPR are dropping much slower than those in traditional web search [1].
  - This motivates the usage of slower decay model
- The middle bias pattern is not observed in our data.

[1] Xie, Xiaohui, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. "Grid-based Evaluation Metrics for Web Image Search." In The World Wide Web Conference, pp. 2103-2114. 2019.



(a) NCTR on mobile devices (b) NCTR on desktops

Figure 2: Normalized click through rate (NCTR) in the top 16 positions of the H&L dataset.



(a) NPR on mobile devices (b) NPR on desktops

Figure 3: Normalized purchase rate (NPR) in the top 16 positions of the H&L dataset.

Propensity Models

- Use propensity models based on data analysis results
  - Row skipping
  - Slower decay

- We adopt the IPS estimator to reweigh each positive feedback label when we train the ranker (e.g., lightGBM).

Guo, Ruocheng, Xiaoting Zhao, Adam Henderson, Liangjie Hong, and Huan Liu. "Debiasing Grid-based Product Search in E-commerce." In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2852-2860. 2020.

Learning to rank experiments

Without randomized experiments, we evaluate the unbiasedness of the ranking models by using training and test data collected from different time intervals.

- Baselines: unbiased LambdaMART, Dual Learning and traditional ranking models.
- Evaluation Metrics: purchase NDCG@K, revenue NDCG@K and purchase mean average precision (MAP)

Models	$NDCG_{pur}$					$NDCG_{rev}$					MAP_{pur}
	@1	@2	@5	@10	@20	@1	@2	@5	@10	@20	@20
Desktop PPS (Paper and Party Supplies)											
MART	0.082	0.121	0.181	0.232	0.291	0.078	0.126	0.184	0.234	0.289	0.079
RankBoost	0.087	0.117	0.184	0.243	0.303	0.087	0.110	0.182	0.241	0.303	0.084
LambdaMART	0.101	0.128	0.194	0.249	0.305	0.100	0.130	0.194	0.248	0.308	0.097
Random Forest	0.096	0.128	0.192	0.239	0.295	0.088	0.117	0.185	0.233	0.287	0.096
Unbiased LambdaMART	0.109	0.142	0.201	0.251	0.308	0.109	0.142	0.201	0.250	0.307	0.106
Dual Learning	0.098	0.136	0.211	0.277	0.327	0.098	0.136	0.211	0.277	0.327	0.094
Row Skipping	0.111	0.141	0.196	0.256	0.312	0.110	0.141	0.196	0.256	0.312	0.106
Slower Decay	**0.144**$^{+}$	**0.173**$^{+}$	**0.232**$^{+}$	**0.281**$^{+}$	**0.340**$^{+}$	**0.143**$^{+}$	**0.173**$^{+}$	**0.232**$^{+}$	**0.281**$^{+}$	**0.339**$^{+}$	**0.139**$^{+}$
Mobile PPS (Paper and Party Supplies)											
MART	0.154	0.197	0.236	0.294	0.347	0.148	0.184	0.227	0.289	0.343	0.154
RankBoost	0.067	0.116	0.181	0.232	0.286	0.085	0.135	0.201	0.252	0.300	0.067
LambdaMART	0.111	0.148	0.216	0.262	0.322	0.119	0.159	0.225	0.272	0.335	0.111
Random Forest	0.138	0.177	0.232	0.286	0.339	0.131	0.176	0.244	0.298	0.343	0.136
Unbiased LambdaMART	0.151	0.192	0.254	0.293	0.345	0.150	0.192	0.253	0.292	0.344	0.149
Dual Learning	0.102	0.144	0.235	0.291	0.340	0.100	0.143	0.235	0.290	0.339	0.101
Row Skipping	0.148	0.182	0.243	0.298	0.351	0.164^{+}	0.203^{+}	0.265^{+}	0.318^{+}	0.370^{+}	0.155
Slower Decay	**0.166**$^{+}$	**0.208**$^{+}$	**0.281**$^{+}$	**0.321**$^{+}$	**0.371**$^{+}$	**0.176**$^{+}$	**0.223**$^{+}$	**0.293**$^{+}$	**0.332**$^{+}$	**0.383**$^{+}$	**0.165**$^{+}$

Insights

- Our model outperforms Unbiased LambdaMART and Dual Learning.
  - Incorporating prior knowledge of user behavior patterns in 2-D display helps learning better propensity score models.
- Row skipping performs better in H&L.
  - We conjecture that this is because: (1) more specific intent of users (2) larger price variance.
- On mobile datasets, baselines can achieve similar results to the proposed method.
  - This is because mobile display (2 columns) is more similar to 1-D web search than desktops (4 columns).

Conclusion

The connections between causal inference and machine learning can be found in many applications across research directions. More connections remain to be explored.

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)

Causality Algorithm Repository:

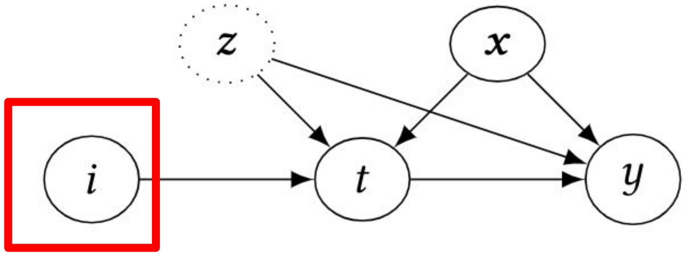
https://github.com/rguo12/awesome-causality-algorithms

Q & A

Causal Identification

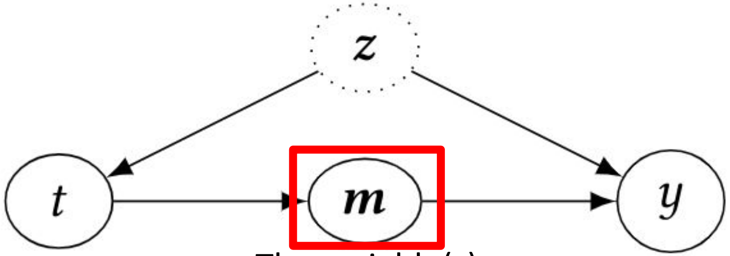
- Causal graphs are powerful tools to represent causal assumptions (e.g., conditional independence) for identification.
- When we observe some special variables:

Identification with Instrumental variables



Instrumental variable

Identification with Front-door criterion



The variable(s) satisfying the front-door criterion