Causal Fairness

Sheridan Grant

University of Washington

slgstats@uw.edu

March 18, 2021

Causal Inference Crash Course

Causal Diagrams & Terminology

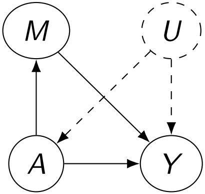


Figure: Mediation with an unobserved confounder

- Arrows represent direct causal effects, but are inherently abstracted from real-world data-generating process
- A is the "treatment" (in fairness, "sensitive attribute")
- Y is the outcome
- M *mediates* the effect of A on Y
- U is an unobserved *confounder*

Causal Diagrams & Terminology

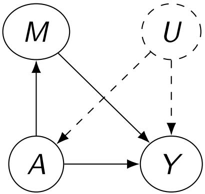


Figure: Mediation with an unobserved confounder

$Y(a)$: the outcome had A been intervened upon to take value a. *A may have taken on value a naturally, anyway.* Let a' denote the "control" level, a the "treatment" (or level of interest). E.g. when assessing racial discrimination, often a' represents white people and a represents Black people.

Causal Diagrams & Terminology

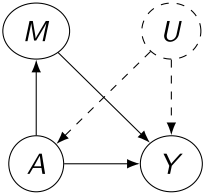


Figure: Mediation with an unobserved confounder

- Average treatment effect (ATE): $E[Y(a) - Y(a')]$.
- Average treatment effect *on the treated* (ATT): $E[Y(a) - Y(a')|A = a]$.
- If A is randomized, then $\text{ATE} = \text{ATT}$.

Types of causal effects

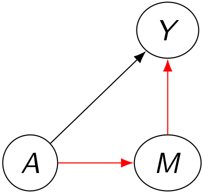


Figure: Mediation with no confounders

Classical effect decomposition:

- Direct effect: $E[Y(a, M(a')) - Y(a')]$
- Indirect (mediation) effect: $E[Y(a) - Y(a, M(a'))]$
- Total effect (ATE or ATT): sum of direct and indirect effects

Types of causal effects

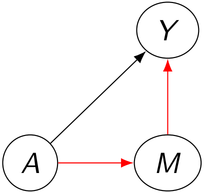


Figure: Mediation with no confounders

Classical effect decomposition (linear model):

- Fit $Y = \beta_0 + \beta_A A + \beta M + \epsilon$; β_A is direct effect
- Fit $Y = \beta'_0 + \beta'_A A + \epsilon$; β'_A is total effect
- $\beta'_A - \beta_A$ is indirect effect

Types of causal effects

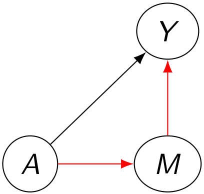


Figure: Mediation with no confounders

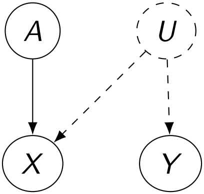
In more complex graphs:

- All of this generalizes to complex diagrams, multiple mediators/paths, confounders, etc.
- Modern causal (often semiparametric) inference studies this
- Nabi and Shpitser 2017 points you to many such papers

Motivating Causal Fairness

When does Fairness through Unawareness fail?

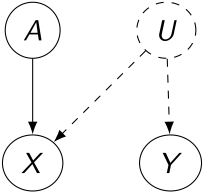
- Fairness through *Unawareness* is a naive but initially appealing approach: simply don't consider race when making decisions (human OR algorithmic).
- In fact, sometimes the only fair thing to do is to explicitly consider race: Fairness through *Awareness*.
- Example: Y = accident rate; X = color of car (1 if red); A = race (1 if black).



When does Fairness through Unawareness fail?

Unfair approaches:

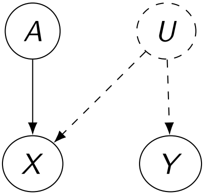
- Model relationship between car color and accidents, charge black people more (because Y and A are d-connected/dependent given X) even though race doesn't affect accident risk



When does Fairness through Unawareness fail?

Fair(?) approaches:

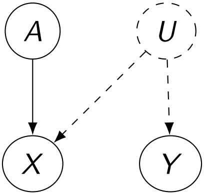
- Randomly price insurance (go out of business)



When does Fairness through Unawareness fail?

Fair(?) approaches:

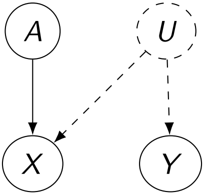
- Randomly price insurance (go out of business)
- Model relationship between accidents and race, find none (because A and Y are not d-connected/marginally independent), randomly price, go out of business



When does Fairness through Unawareness fail?

Fair(?) approaches:

- Randomly price insurance (go out of business)
- Model relationship between accidents and race, find none (because A and Y are not d-connected/marginally independent), randomly price, go out of business
- Model relationship between accidents vs. race AND car color, charge red cars more, give black people fair "discount" that accounts for association with accident-prone (but not accident-*causing*) trait



Counterfactual Fairness (Kusner et al. 2017)

Computing Counterfactual Predictions

Consider an outcome Y, sensitive attribute A, covariates X (which may contain descendants anbd/or ancestors of A), latent variables U *that are non-descendants of* A, and a predictor $\hat{Y}$ that is a function of X, possibly A, and possibly U. We wish to compute the counterfactual distribution of

$$\hat{Y}_{A \leftarrow a'}(U) | Y = y, X = x, A = a$$

Computing Counterfactual Predictions

Algorithm for computing

$$\hat{Y}_{A \leftarrow a'}(U) | Y = y, X = x, A = a$$

1. (Only if $\hat{Y}$ is a function of U:) Compute $\mathrm{P}(U|Y = y, X = x, A = a)$. The paper (Kusner et al. 2017) crucially omits Y, which is often needed to learn this posterior distribution (we are interested in latent variables that are informative about Y, after all).
2. Intervene by setting $A = a'$, and use the SEM associated with the causal diagram to also change the values of all descendants of A in X to $X(A = a')$.
3. Compute $\hat{Y}$ from the new $A = a'$, $X(A = a')$, and possibly by averaging over $\mathrm{P}(U|Y = y, X = x, A = a)$

Counterfactual Fairness

Definition

$\hat{Y}$ is counterfactually fair if

$$P(\hat{Y}_{A\leftarrow a}(U)|Y=y,X=x,A=a) = P(\hat{Y}_{A\leftarrow a'}(U)|Y=y,X=x,A=a)$$

- A sufficient condition for counterfactual fairness is if $\hat{Y}$ is not a function of A *or any of its descendants*
- Proposition: sufficient condition for counterfactual fairness: 1) no direct effect of A on Y, 2) model covariates d-separate A from Y, 3) model is correctly specified
- Authors admit that allowing for race to affect $\hat{Y}$ along some paths (Nabi and Shpitser 2017, next section) is desirable

Counterfactually Fair Estimation

1. Write down causal model for latent variables U that are non-descendants of A
2. Generate synthetic latent variables from $\mathrm{P}(U|X, A)$
3. Minimize $L(Y, f(U, X \setminus \mathrm{desc}(A)))$ empirically over the observed data and synthetic latent variables.

The learned $\hat{f}$ trivially satisfies counterfactual fairness because it satisfies the sufficient condition from previous slide.

Pathwise Fairness (Nabi and Shpitser 2017)

Hypothetically Fair Worlds

Causal models seek to reconstruct a hypothetical world in which the treatment was randomly assigned. Nabi and Shpitser 2017 do this with fairness: estimate a “fair” world that is KL-close to the observed world.

- Assume linearity, standardized variables for now
- “fair”: PSE strengths restricted to $[\epsilon_l, \epsilon_u]$
- Divide covariates into X and Z, and condition on the Z covariates—that is, assume they come from a “fair world.”
- Estimate parameters of p^* subject to PSE constraints.
- For future predictions: 1) use $\tilde{X}_i \equiv \mathrm{E}^*[X|Z_i]$ in place of X_i, 2) use $p^*(Y_i, \tilde{X}_i, Z_i)$ to make predictions
- Example: BART

COMPAS Results

Use BART (Chipman, George, and McCulloch 2010) as outcome model, but in MCMC reject any step yielding a PSE outside constrained range.

Model	Accuracy	NDE ($1 =$ fair)
Unconstrained	67.8%	1.3
Constrained	66.4%	1.05
Race-unaware	64%	2.1

Table: Accuracies and race NDE for various BART models of COMPAS data.

Challenges for Future Work

- In general, constraining PSEs introduces nonconvex constraints: assuming a linear SEM, a 1-length path needs only convex constraints, but a 2-length path (e.g. $A \to M \to Y$) require a nonconvex constraint ($\epsilon_l < \beta_{A \to M} \cdot \beta_{M \to Y} < \epsilon_u$). This is clearly a serious problem and one of the main gaps in the paper.
- Choice of X and Z. Authors discuss "tradeoffs" but it appears to me that the more variables in Z the better (judging from the developments in "Fair Inference From Finite Samples," the authors seem to agree).

References I

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees". EN. In: The Annals of Applied Statistics 4.1 (Mar. 2010), pp. 266–298. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/09-AOAS285. URL: https://projecteuclid.org/euclid.aoas/1273584455 (visited on 03/20/2019).

Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: arXiv:1610.02413 [cs] (Oct. 2016). arXiv: 1610.02413. URL: http://arxiv.org/abs/1610.02413 (visited on 10/16/2018).

References II

Matt J Kusner et al. "Counterfactual Fairness". In: Advances in Neural Information Processing Systems 30. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4066–4076. URL: http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf (visited on 10/16/2018).

Razieh Nabi and Ilya Shpitser. "Fair Inference On Outcomes". In: arXiv:1705.10378 [stat] (May 2017). arXiv: 1705.10378. URL: http://arxiv.org/abs/1705.10378 (visited on 08/21/2018).

Appendix of Slides that are Partially Wrong

When do associative fairness metrics fail?

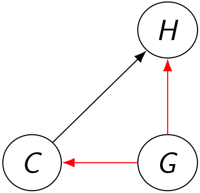


Figure: Prior conviction C, hiring H, and gender G

p(H=1| G,C)	G value	C value	p(C=1 |G)
0.06	1	1	0.99
0.01	0	1	0.01
0.2	1	0	
0.05	0	0	

Figure: Rates of hiring H for different genders G and prior conviction status C.

This distribution actually displays equality of opportunity! (Hardt, Price, and Srebro 2016)

Discussion Points

- Top of p. 6: "counterfactual fairness makes impossibility result regarding calibration and equalized odds irrelevant"
- Latent variables U ("distribution of background variables [U] as given by a... model... that is available by assumption")—how to model them? Tradeoff between assumptions and informativeness?