PC-fairness Note

Yuzhe Ou

Sep 2020

1 Overview

Fair machine learning becomes a important research field as to study how to develop predictive machine learning models that making decisions and fairly treat all groups of people. Pearl develops structural causal models (SCM), and based on it a number of causality-based fairness notions have been proposed for capturing different kind of fairness, including total effect, direct/indirect discrimination, counterfactual fairness and so on. In unidentifiable cases, measuring exact causal effect is unsolvable, thus some methods are proposed to bound these casual effects. The author of this paper proposed a new fairness notion called Path-specific Counterfactual Fairness (PC-fairness), which unifies previous representation of causal effects. Corresponding fairness constraint is then defined on the predictive model. To bound PC-fairness, a linear programming problem is formulated using response-function variable. The experimental results show that it achieves tighter bounds for truth value of fairness notion.

2 Identifiability

Definition of identifiability: Consider a class of models $\boldsymbol{M}$ with a description T, and objects ϕ and θ computable from each model. Then ϕ is θ-identified in T if ϕ is uniquely computable from θ in any $M \in \boldsymbol{M}$. In this case all models in $\boldsymbol{M}$ agree on θ will also agree on ϕ.

Let θ be the causal model parameters that identify the data generating process, ϕ be the causal effect that is computed from causal model, then, the identifiability of causal effects can be viewed as whether we can extract unique causal effect from the class of possible models (i.e. unidentifiable if there exist two causal models which exactly agree on the observational distribution).

3 SCM and causal effects

Definition 1: A structural causal model $\mathcal{M}$ is represented by a quadriple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where
1.$\mathbf{U}$ is a set of exogenous variables that are determined by factors outside the model.
2.$P(\mathbf{U})$ is a joint probability distribution defined over $\mathbf{U}$.
3.$\mathbf{V}$ is a set of endogenous variables that are determined by variables in $\mathbf{U} \cup \mathbf{V}$
4.$\mathbf{F}$ os a set of structural equations from $\mathbf{U} \cup \mathbf{V}$ to $\mathbf{V}$. Specifically, for each $V \in \mathbf{V}$, there is a function $f_V \in \mathbf{F}$ mapping from $\mathbf{U} \cup (\mathbf{V} \setminus V)$ to V, i.e., $v = f_V(pa_V, u_V)$, where pa_V is a realization

of a set of endogenous variables $\mathbf{PA}_V \in \mathbf{V} \setminus V$ that directly determines V, and u_V is a realization of a set of exogenous variables that directly determines V.

A notation: Y_x represents interventional variant while performing $do(X = x)$ or $do(x)$, denote $P(y_x) = P(Y_x = y)$.

Definition 2 (Total Causal Effect): The total causal effect if the value change of X from x_0 to x_1 on $Y = y$ is given by

$$\mathrm{TCE}(x_1, x_0) = P(y_{x_1}) - P(y_{x_0})$$

Definition 3 (Path-specifc Effect): Given a causal path set π, the π-specific effect of the value change of X from x_0 to x_1 on $Y = y$ through π (with reference x_0) is given by

$$\mathrm{PE}_\pi(x_1, x_0) = P(y_{x_1|\pi, x_0|\bar{\pi}}) - P(y_{x_0})$$

where $P(y_{x_1|\pi,x_0|\bar{\pi}})$ represents the post-intervention distribution of Y where the effect of intervention $do(x_1)$ is transmitted only along π while the effect of reference intervention $do(x_0)$ is transmitted along the other paths.

Definition 4 (Counterfactual Effect): Given a factual condition $\mathbf{O} = \mathbf{o}$, the counterfactual effect of the value change of X from x_0 to x_1 on $Y = y$ is given by

$$\mathrm{CE}(x_1, x_0|\mathbf{o}) = P(y_{x_1}|\mathbf{o}) - P(y_{x_0}|\mathbf{o})$$

4 PCE

This new fairness notion unifies following causal effects (and more) by taking different setting of $\mathbf{O}$ and π: Total effect, (System) Direct discrimination, (System) Indirect discrimination, Individual direct discrimination, Group direct discrimination, Counterfactual fairness, Counterfactual error rate.

Definition 5 (Path-specific Counterfactual Effect): Given a factual condition $\mathbf{O} = \mathbf{o}$ and a causal path set π, the path-specific counterffactual effect of the value change of X from x_0 to x_1 on $Y = y$ through π (with reference x_0) is given by

$$\mathrm{PCE}(x_1, x_0|\mathbf{o}) = P(y_{x_1|\pi, x_0|\bar{\pi}}|\mathbf{o}) - P(y_{x_0}|\mathbf{o})$$

Definition 6 (Path-specofic counterfactual Fairness (PC fairness)): Given a Given a factual condition $\mathbf{O} = \mathbf{o}$ where $\mathbf{O} \subseteq \{S, \mathbf{X}, Y\}$ and a causal path set π, predictor $\hat{Y}$ achieves the PC fairness if $\mathrm{PCE}(s_1, s_0|\mathbf{o}) = 0$ where $s_1, s_0 \in \{s^+, s^-\}$. $\hat{Y}$ achieves the τ-PC fairness if $|\mathrm{PCE}(s_1, s_0|\mathbf{o})| \leq \tau$.

5 Measuring PC-fairness

In order to measure PC-fairness, the author use response-function variable $R_V = \{0, ..., N_V - 1\}$ where $N_V = |V|^{PA_V}$ is the total number of different deterministic response functions mapping from PA_V to V ($N_V = |V|$ if V has no parent). Since we need to go through all possible models that agree on our observations in order to calculate the bound of causal effect, response-function variables are such equivalent encoding of possible structural equations set (i.e. capturing deterministic

relationship between a variable and its parents given particular unknown exogenous variables) that has nice property. In other words given parent and response-function variable, we'll know the value of child. Let $r_V = l_V(u_V)$ be the mapping from U_V to R_V. Then

$$f_V(pa_V, u_V) = f_V(pa_V, l_V^{-1}(r_V)) = g_V(pa_V, r_V)$$

where g_V is the composition of f_V and l_V^{-1}.

Then an indicator function is defined to handle the conditioning part:

$$\mathbb{I}(v; pa_V, r_V) = \begin{cases} 1 & \text{if } g_V(pa_V, r_V) = v, \\ 0 & \text{otherwise,} \end{cases}$$

Let $V(\mathbf{u})$ be the value that V would obtain if $\mathbf{U} = \mathbf{u}$ and $V(\mathbf{r}) = g_V(pa_V, r_V)$.Then we express the joint distribution in terms of response function variables:

$$\begin{aligned} P(v) &= \sum_{\mathbf{u}:\mathbf{V}(u)=\mathbf{v}} P(\mathbf{u}) \\ &= \sum_{\mathbf{r}:\mathbf{V}(r)=\mathbf{v}} P(\mathbf{r}) \\ &= \sum_{\mathbf{r}} P(r) \prod_{v \in \mathbf{V}} \mathbb{I}(v; pa_V, r_V) \end{aligned}$$

The left is to express PCE using $P(r)$. Some variables may take different value on π and $\bar{\pi}$, we need to divide non-protected attributes $\mathbf{X}$ into three disjoint sets and treat them separately: $\mathbf{W}$ denoting the set of witness variables (on both π and $\bar{\pi}$), $\mathbf{A}$ denoting the set of non-witness variables on π (but not $\hat{\pi}$), $\mathbf{B}$ denoting the set of non-witness variables on $\hat{\pi}$ (but not on π).

We also denote the values of $\mathbf{R}$ that satisfy $\mathbf{O}(\mathbf{r}) = \mathbf{o}$ by $\mathbf{r_o}$. Then we consider

$$\text{PCE}_\pi(s_1, s_0|\mathbf{o}) = P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o})$$

The first term of PCE writes:

$$\begin{aligned} P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o}) &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w_1},\mathbf{w_0}} P(\hat{Y}_{s_1|\pi,s_0|\bar{\pi}} = y, \mathbf{A}_{s_1|\pi} = \mathbf{a}, \mathbf{B}_{s_1|\bar{\pi}} = \mathbf{b}, \mathbf{W}_{s_1|\pi} = \mathbf{w_1}, \mathbf{W}_{s_0|\bar{\pi}} = \mathbf{w_0}) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w_1},\mathbf{w_0},\mathbf{r}\in\mathbf{r_o}} \frac{P(\mathbf{r})}{P(\mathbf{o})} \mathbb{I}(\hat{y}; pa^1_{\hat{Y}}, r_{\hat{y}}) \mathbb{I}(a; pa^1_A, r_A) \mathbb{I}(b; pa^0_B, r_B) \mathbb{I}(w_1; pa^1_W, r_W) \mathbb{I}(w_0; pa^0_W, r_W) \end{aligned}$$

Similarly, the second term writes:

$$\begin{aligned} P(\hat{y}_{s_0}|\mathbf{o}) &= \sum_{\mathbf{v}'} P(\hat{Y}_{s_0} = y, \mathbf{V}'_{s_0} = \mathbf{v}'|\mathbf{o}) \\ &= \sum_{\mathbf{v}',\mathbf{r}\in\mathbf{r_o}} \frac{P(\mathbf{r})}{P(\mathbf{o})} \mathbb{I}(\hat{y}; pa_{\hat{Y}}, r_{\hat{y}}) \prod_{V \in \mathbf{V}'} \mathbb{I}(v; pa_V, r_V) \end{aligned}$$

where $\mathbf{V}' = \mathbf{V} \setminus \{S, Y\}$

Therefore, given a set of response-function variables we are able to express PCE as a linear expression of $P(\mathbf{r})$.

6 Bounding PC fairness

Now we know that the PC fairness can be expressed using a linear expression of response-function variables whose distribution is not known. However we can consider all the possible situations to find out the max/min value of PCE of our predictive model.

Formally, the author formulate following linear programming problem:

$$\begin{aligned} \min/\max \;\; & P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o}), \\ \text{s.t.} \;\; & P(\mathbf{V}) = P(\mathcal{D}),\ \sum_{\mathbf{r}} P(\mathbf{r}) = 1,\ P(\mathbf{r}) \geq 0, \end{aligned}$$

This will traverse all the causal models (parameterized by $\mathbf{r}$) that agree on observational distribution $P(\mathcal{D})$, and find the lower and upper bound of PCE. Finally if the upper bound is less than $-\tau$ and lower bound is greater than τ, τ-PC fairness must be satisfied. Otherwise, it cannot be determined from the data. We can add this as a constraint and form a bilevel optimization framework

$$\begin{aligned} & \min \mathcal{L}(\mathcal{D};\theta) \\ & \text{s.t. } \max\{|\text{PCE}_\pi(s^+, s^-|\mathbf{o})|\} \leq \tau \\ & \text{w.r.t } P(\mathbf{V}) = P(\mathcal{D}),\ \sum_{\mathbf{r}} P(\mathbf{r}) = 1,\ P(\mathbf{r}) \geq 0 \end{aligned}$$

where $\mathcal{L}(\mathcal{D};\theta)$ is the loss function for our main model parameterized by θ, τ is predefined positive constant.

7 Discussion

The advantage of proposed framework is that the unified notion is more general for expressing fairness metric, and the solution of causal effect bound is obtained by linear programming which is straightforward and easy to implement. However, this may suffer from large causal graph while joint domain size for the response variables is exponentially grown with increase of in-degree. The author mentions that under certain circumstances, some response variables can be eliminated if not causing the unidentifiability, but identifying such cases could be a problem itself.

References

[1] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, September 2008.

[2] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3404–3414. Curran Associates, Inc., 2019.