Counterfactual Fairness

20.07.02

Presenter : YC, Choi

Contents

Counterfactual Fairness

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

Contents

Counterfactual Fairness

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

Notation

- A : the set of protected attributes
- X : observable attributes
- U : latent attributes
- Y : the outcome to be predicted
- $\hat{Y}$: predictor, a random variable that depends on A, X, U

Other fairness measure

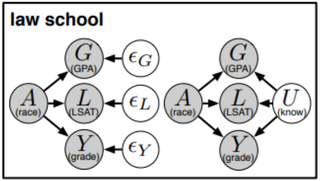
1. Fairness Through Unawareness(FTU)
   An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.
2. Individual Fairness(IF)
   If individual i and j are similar under a metric $d(\cdot, \cdot)$, then their predictions should be similar ($\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.

Other fairness measure

3. Demographic Parity(DP)
   A predictor $\hat{Y}$ satisfies demographic parity if $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$
4. Equality of Opportunity(EO)
   A predictor $\hat{Y}$ satisfies equality of opportunity if
   $P(\hat{Y}=1|A=0, Y=1) = P(\hat{Y}=1|A=1, Y=1)$

Causal Models and Counterfactuals

- Causal model is defined by (U, V, F)
- V : observable variables
- U : set of latent background variable, which are factors not caused by V
- F is a set of functions $\{f_1, \ldots, f_n\}$ such that $V_i = f_i(pa_i, U_{pa_i})$ where
  $pa_i \subseteq V \setminus \{V_i\}, U_{pa_i} \subseteq U$
  pa_i referes to the "parents" of V_i



Causal Models and Counterfactuals

- Intervention on variable V_i
  substitution of equation $V_i = f_i(pa_i, U_{pa_i})$ with the equation $V_i = v$
- Counterfactual
  - the value of Y if A had taken value a
  - solution for Y given $U = u$ where the equations for A are replace with $A = a$
  - $Y_{A \leftarrow a}(u)$ or Y_a

Counterfactual Fairness

- (Definition) Predictor Y is counterfactually fair if any context $X = x$ and $A = a$

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

for all y and for any value a' attainable by A

Counterfactual Fairness

- (Lemma) Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A is invariant with respect to the counterfactual values of A.

Algorithm

- $\hat{Y} \equiv g_\theta(U, X_{\nsucc A})$: predictor parameterized by θ
- $X_{\nsucc A} \subset X$: non-descendants of A
- $\mathcal{D} \equiv \{(A^{(i)}, X^{(i)}, Y^{(i)} : i = 1, \ldots, n\}$: training data
- $l(\cdot, \cdot)$: loss function(squared loss or log-likelihood)

1: **procedure** FAIRLEARNING($\mathcal{D}, \mathcal{M}$) ▷ Learned parameters $\hat{\theta}$
2: For each data point $i \in \mathcal{D}$, sample m MCMC samples $U_1^{(i)}, \ldots, U_m^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
3: Let $\mathcal{D}'$ be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in $\mathcal{D}$ is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}$.
4: $\hat{\theta} \leftarrow \text{argmin}_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(U^{(i')}, x_{\nsucc A}^{(i')}))$.
5: **end procedure**

Contents

Counterfactual Fairness

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness

- The previous method requires that one provides the causal model that generated the data at hand
- There are infinitely many structural equations compatible with the same observable distribution.
- It is desirable to integrate competing causal models to provide counterfactually fair decisions

Definition

- $(\epsilon, 0) - ACF$(Approximate Counterfactual Fairness)
  A predictor $f(\mathcal{X}, A)$ satisfies $(\epsilon, 0) - ACF$ if given the sensitivity attribute $A = a$ and any instantiation x of the other observed variable $\mathcal{X}$, we have that

$$|f(x_{A \leftarrow a}, a) - f(x_{A \leftarrow a'}, a')| < \epsilon \tag{5}$$

  for all $a' \neq a$
- $(\epsilon, \delta) - ACF$
  f satisfies $(\epsilon, \delta) - ACF$ if

$$\mathbb{P}_U(|f(\mathcal{X}_{A \leftarrow a}, a) - f(\mathcal{X}_{A \leftarrow a'}, a')| < \epsilon | \mathcal{X} = x, A = a) > 1 - \delta \tag{6}$$

Algorithm

- objective function :
$$min_f \frac{1}{n}\sum_{i=1}^{n} l(f(x_i, a_i), y_i) + \lambda \sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n}\sum_{a' \neq a_i} \mu_j(f, x_i, a_i, a') \tag{7}$$
where $\mu_j(f, x_i, a_i, a') := \mathbb{I}[|f(x_{A\leftarrow a}, a) - f(X_{A\leftarrow a'}), a')| > \epsilon]$
- surrogated version : $\mu_j(f, x_i, a_i, a') := \max\{0, |f(x_{A\leftarrow a}, a) - f(X_{A\leftarrow a'}, a')| - \epsilon\}$
- maximum risk version : $\lambda \sum_{j=1}^{m} \max_i \sum_{a' \neq a_i} \mu_j(f, x_i, a_i, a')$

Algorithm

Algorithm 1 Multi-World Fairness

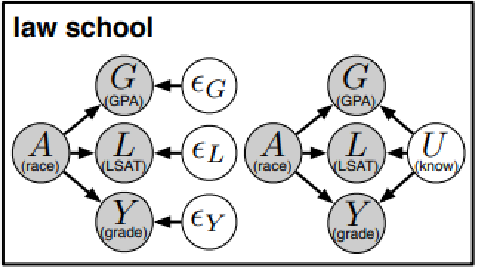
Input: features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, labels $\mathbf{y} = [y_1, \ldots, y_n]$, sensitive attributes $\mathbf{a} = [a_1, \ldots, a_n]$, privacy parameters (ϵ, δ), trade-off parameters $\mathcal{L} = [\lambda_1, \ldots, \lambda_l]$.
Fit causal models: $\mathbf{M}_1, \ldots, \mathbf{M}_m$ using $\mathbf{X}, \mathbf{a}$ (and possibly $\mathbf{y}$).
Sample counterfactuals: $\mathcal{X}_{A^1 \leftarrow a'}, \ldots, \mathcal{X}_{A^m \leftarrow a'}$ for all unobserved values a'.
for $\lambda \in \mathcal{L}$ **do**
    Initialize classifier f_λ.
    while loop until convergence **do**
        Select random batches $\mathbf{X}_b$ of inputs and batch of counterfactuals $\mathbf{X}_{A^1 \leftarrow a'}, \ldots, \mathbf{X}_{A^m \leftarrow a'}$.
        Compute the gradient of equation (7).
        Update f_λ using any stochastic gradient optimization method.
    end while
end for
Select model f_λ: For deterministic models select the smallest λ such that equation (5) using f_λ holds. For non-deterministic models select the λ that corresponds to δ given f_λ.

Experiments

- Law School Admission Council data.
- A : race(protected attribute)
- G : GPA, L : law school entrance exam score
- Y : First year average grade.
- Law school may be interested in predicting Y for all applicants to law school in order to decide whether to accept or deny them entrance.

Experiments

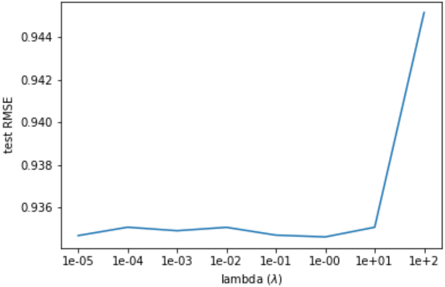
- Causal model



$$G = b_G + w_G^A A + \epsilon_G \qquad G \sim \mathcal{N}(b_G + w_G^A A + w_G^U U, \sigma_G)$$
$$L = b_L + w_L^A A + \epsilon_L \qquad L \sim \text{Poisson}(\exp(b_L + w_L^A A + w_L^U U))$$
$$Y = b_Y + w_Y^A A + \epsilon_Y \qquad Y \sim \mathcal{N}(w_Y^A A + w_Y^U U, 1)$$
$$\epsilon_G, \epsilon_L, \epsilon_Y \sim \mathcal{N}(0,1) \qquad U \sim \mathcal{N}(0,1)$$

Experiments

- ablation study about λ
- $\lambda \in \{10^{-5}, 10^{-4}, \ldots, 10^{10}\}$



Experiments

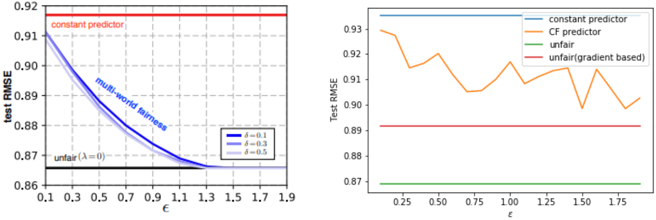


Figure 1: (LEFT)paper simulation (Right) Ours

Experiments

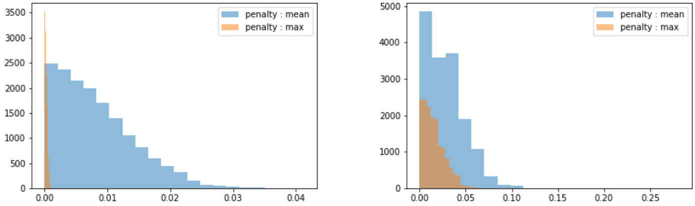


Figure 2: (LEFT)λ : 0.01 (Right) λ : 0.1