Individual Fairness AI Reviews

Sarah Kim

August 4, 2020

Table of Contents

Metric Based

Fairness Through Awareness (2012)

Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

Counterfactual Fairness (2017)

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

Average Individual Fairness (2019)

Table of Contents

Metric Based

- Fairness Through Awareness (2012)
- Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

- Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

- Counterfactual Fairness (2017)
- When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

- Average Individual Fairness (2019)

Introduction

- An algorithm is **individual fair** if it gives similar predictions to similar individuals, i.e.,

$$\left|P(\hat{Y}_i = y|X_i) - P(\hat{Y}_j = y|X_j)\right| \leq \epsilon; \text{ if } d(X_i, X_j) \approx 0$$

  where i, j denote two individuals.

- $d(i, j)$ is a distance metric between two individuals, and here we assume $d(i, j)$ is given for specific task.

Treating similar individuals similarly

- In a binary classification problem, we consider randomized mappings $M : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$ from individuals to probability distribution over outcomes. To classify $x \in \mathcal{X}$ choose an outcome y according to the distribution $M(Y = y|x)$.
- Find a mapping from individuals to distribution over outcomes that minimizes expected loss **subject to the** (D, d)-**Lipschitz condition**,

$$D(M(\cdot|x), M(\cdot|x')) \leq d(x, x'), \ \forall x, x' \in \mathcal{X}$$

where D is a measure of similarity of distributions.

Treating similar individuals similarly

- In a binary classification problem, we consider randomized mappings $M : \mathcal{X} \to \Delta(\mathcal{Y})$ from individuals to probability distribution over outcomes. To classify $x \in \mathcal{X}$ choose an outcome y according to the distribution $M(Y = y|x)$.
- Find a mapping from individuals to distribution over outcomes that minimizes expected loss **subject to the** (D, d)-**Lipschitz condition**,

$$D(M(\cdot|x), M(\cdot|x')) \leq d(x, x'), \ \forall x, x' \in \mathcal{X}$$

where D is a measure of similarity of distributions.

Treating similar individuals similarly

- Denote $\mathcal{I}$ as an instance, and L as a loss function.
- The Fairness LP:

$$\mathrm{opt}(\mathcal{I}) \stackrel{\text{def}}{=} \min_{M} \int_{x} \sum_{y \in \{0,1\}} \sum_{\hat{y} \in \{0,1\}} L(y, \hat{y}) M(Y = \hat{y} | x) P(x, Y = y) dx \quad (1)$$
$$\text{subject to } \forall x, x' \in \mathcal{X} : \; D(M(\cdot|x), M(\cdot|x')) \le d(x, x')$$

Probability metrics

- Let P, Q denote probability measures on a finite domian A.
- Total variation norm between P and Q:

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

- Relative ℓ_∞ norm between P and Q:

$$D_\infty(P, Q) = \sup_{a \in A} \log \left(\max \left\{ \frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)} \right\} \right)$$

- **Lemma.** Let $D \in \{D_{\text{tv}}, D_\infty\}$. Given an instance $\mathcal{I}$, we can compute $\text{opt}(\mathcal{I})$ with a linear program of size $\text{poly}(|\mathcal{X}|, |\mathcal{Y}|)$ (w.r.t. M).

Table of Contents

Metric Based

Fairness Through Awareness (2012)

Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

Counterfactual Fairness (2017)

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

Average Individual Fairness (2019)

Introduction

- In the previous work, the individual-fair learning algorithm is computationally intractable (even for simple fair-learning tasks).
- Suppose a similariy metric d is given.
- The author proposed a approximately individual-fair condition which is a relaxed version of the previous individaul fairness.

Approximate Metric-Fariness

- **Def.** A predictor h is (α, γ)-approximately metric-fair (MF) w.r.t. a similarity metric d and a data distribution $\mathcal{D}$ if

$$\mathcal{L}_{\gamma}^{F} := P_{x,x' \sim \mathcal{D}}[|h(x) - h(x')| > d(x, x') + \gamma] \leq \alpha.$$

- If $\alpha = 0$, then $\mathcal{L}_{\gamma}^{F}$ means perfect MF.
- Notation
  - $H^{\alpha,\gamma}$: the set of functions which satisfy (α, γ)-approximate MF on $\mathcal{D}$
  - $\widehat{H}^{\alpha,\gamma}$: the set of functions which satisfy (α, γ)-approximate MF on the training set.

Fair learning

- Objective for fair learning:

$$\text{minimize}_h \ err_S(h) \text{ subject to } h \in \widehat{H}^{\alpha,\gamma}$$

where S is a training set, $err_S(h)$ denotes the expected ℓ_1 error of h.

- Since $\widehat{H}^{\alpha,\gamma}$ in the contraint induces 0/1 loss, they use empirical ℓ_1 MF violation $\xi_S(h)$ given by

$$\xi_S(h) = \sum_{x,x' \in S} \max(0, |h(x) - h(x')| - d(x,x')).$$

- For some $\tau \in [0,1]$,

$$\text{minimize}_h \ err_S(h) \text{ subject to } \xi_S(h) \le \tau.$$

Main contributions

- (Generalization) This fair learning guaranteeing fairness not just for the training sets but also for the underlying population distribution, under some conditions.
- (Efficiency) This algorithm guarantees to contruct polynomial-time learning algorithm which satisfies approximate MF and best-possible accuracy (for classes of linear and logistic predictors).

Table of Contents

Metric Based

Fairness Through Awareness (2012)

Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

Counterfactual Fairness (2017)

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

Average Individual Fairness (2019)

Introduction

- We consider the linear contextual bandits problem with strong individual fairness constraints.
- In this paper, a class of distance functions is specified by Mahalanobis distance (i.e., for some matrix A, $d(x_1, x_2) = \|Ax_1 - Ax_2\|_2$).

Linear Contextual Bandits

- Notation
  - t: round, $t \in [T]$
  - k: number of multi-arms
  - $x_i^t \in \mathbb{R}^d$: contexts vector of an arm i in round t
  - i^t: chosen arm at round t after observing contexts
  - $r_{i^t}^t$: after choosing an arm, observed some stochastic reward s.t. $r_{i^t}^t$ is sub-gaussian and $\mathbb{E}[r_{i^t}^t] =< x_{i^t}^t, \theta >$ where $\theta \in \mathbb{R}^d$ is a coefficient vector
  - $h^t = ((\mathbf{x}^1, i^1, r^1), \dots, (\mathbf{x}^{t-1}, i^{t-1}, r^{t-1}))$: a history at round t
  - $\pi^t = \pi^t(h^t, \mathbf{x}^t) \in \Delta[k]$: the probability distribution over actions that the algorithm plays action i at round t
- Note that the algorithm does not observe the reward for the actions not chosen.

Fairness Constraints and Feedback

- **Def 1.** Algorithm L is Lipschitz-fair on round t w.r.t. d if for all i, j:

$$|\pi_i^t - \pi_j^t| \leq d(x_i^t, x_j^t).$$

- **Def 2 (Fairness Oracle).** Given d, a fairness oracle O_d defined as follows:

$$O_d(x^t, \pi^t) = \{(i, j) : |\pi_i^t - \pi_j^t| > d(x_i^t, x_j^t)\}$$

- Assumption: algorithm L have access to a fairness oracle, use this feedback to learn d

Best Fair Policy

In round $t = 1, \ldots, T$,

1. Parameter estimation: $\hat{\theta}^t = (X^{t\top} X^t + \lambda I)^{-1} X^{t\top} R^t$ where $X^t = [x_{i_1}^1, \ldots, x_{i_{t-1}}^{t-1}]$ and $R^t = [r_{i_1}^1, \ldots, r_{i_{t-1}}^{t-1}]$
2. Reward estimation and UCB (upper confidence bound): $\tilde{r}_i^t =< \hat{\theta}^t, x_i^t >$ and $\hat{r}_i^t = \tilde{r}_i^t + B_i^t$ with $P(|r_i^t - \tilde{r}_i^t| \leq B_i^t) = 1 - \delta$.
3. Policy estimation: given $\hat{\mathbf{r}}^t = (\hat{r}_1^t, \ldots, \hat{r}_k^t)$, $\hat{\mathbf{d}}^t = (\hat{d}(x_i^t, x_j^t))_{i<j}$,

$$\pi^t(\hat{\mathbf{r}}^t, \hat{\mathbf{d}}^t) = \operatorname*{argmax}_{\pi \in \Delta[k]} \sum_{i=1}^{k} \pi_i \hat{r}_i^t \quad \text{subject to } |\pi_i - \pi_j| \leq \hat{d}(x_i^t, x_j^t), \forall (i,j) \tag{2}$$

Estimation for d

$\pi^t = \pi(\bar{r}^t, \hat{d}^t)$
Pull an arm i^t according to π^t and receive a reward $r^t_{i^t}$
$S = \mathbf{O}_d(x^t, \pi^t)$
$R = \{(i,j) | (i,j) \notin S \wedge |p^t_i - p^t_j| = \hat{d}^t_{ij}\}$
for $(i,j) \in S$ **do**
    DistanceEstimator$_{ij}.feedback(\perp)$
    $v^t_{ij} = 1$
end
for $(i,j) \in R$ **do**
    DistanceEstimator$_{ij}.feedback(\top)$
    $v^t_{ij} = 1$
end

Table of Contents

Metric Based

- Fairness Through Awareness (2012)
- Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

- Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

- Counterfactual Fairness (2017)
- When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

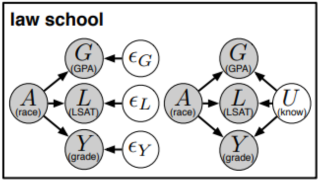
- Average Individual Fairness (2019)

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

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)

```
procedure FAIRLEARNING(𝒟, ℳ)                                    ▷ Learned parameters θ̂
    For each data point i ∈ 𝒟, sample m MCMC samples U_1^(i), ..., U_m^(i) ~ P_ℳ(U | x^(i), a^(i)).
    Let 𝒟' be the augmented dataset where each point (a^(i), x^(i), y^(i)) in 𝒟 is replaced with the
corresponding m points {(a^(i), x^(i), y^(i), u_j^(i))}.
    θ̂ ← argmin_θ Σ_{i'∈𝒟'} l(y^(i'), g_θ(U^(i'), x_{⊁A}^(i'))).
end procedure
```

Table of Contents

Metric Based

Fairness Through Awareness (2012)

Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

Counterfactual Fairness (2017)

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

Average Individual Fairness (2019)

Introduction

- 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

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

  for all $a' \neq a$

- (ϵ, δ)-ACF

  f satisfies (ϵ, δ)-ACF if

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

Algorithm

- objective function :

$$\min_f \frac{1}{n}\sum_{i=1}^{n} l(f(x_i, a_i), y_i) + \lambda \sum_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n}\sum_{a' \neq a_i} \mu_j(f, x_i, a_i, a') \quad (7)$$

where $\mu_j(f, x_i, a_i, a') := \mathbb{I}[|f(x_{i,A\leftarrow a_i}, a_i) - f(x_{i,A\leftarrow a'}), a')| > \epsilon]$

- surrogated version :

$\mu_j(f, x_i, a_i, a') := \max\{0, |f(x_{i,A\leftarrow a_i}, a_i) - f(x_{i,A\leftarrow a'}), a')| - \epsilon\}$

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.

Table of Contents

Metric Based

Fairness Through Awareness (2012)

Probably Approximately Metric-Fair Learning (2018)

Bandit Problem Based

Online Learning with an Unknown Fairness Metric (2018)

Counterfactual Based

Counterfactual Fairness (2017)

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (2017)

Multiple classifications based

Average Individual Fairness (2019)

Introduction

- In this paper, consider multiple classification tasks.
  (ex. ads for internet users, public school admissions)
- Average Individual Fairness constraints: standard statistics (such as error or FP/FN rates) should be approximately equalized across all individuals
- Here, 'rate' is defined as the average over classification tasks.
- Given a sample of individuals and classification problems, authors design an algorithm for the fair empirical risk minimization task.

Notations

- $i \in [n]$: index for a individual, $j \in [m]$: index for a classification task
- $\mathcal{P}$: probability measure over $\mathcal{X}$, $\mathcal{Q}$: probability measure over the space of problems $\mathcal{F}$
- Dataset: $D = \{\mathbf{x}_i, (f_j(x_i))_{j=1}^m\}_{i=1}^n$ where $f_j(x_i) \in \{0, 1\}$ is the label corresponding to $\mathbf{x}_i$ for the jth classification task.
- Denote $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ as learning m randomized classifieres, where p_j is the learned classifier for the jth classification task.

Definitions

- **Def 1.** (Individual and Overall Error Rates)

  The individual error rate of $\mathbf{x}$ incurred by $\mathbf{p}$ is defined as follows:

  $$\mathcal{E}(\mathbf{x}, \mathbf{p}; \mathcal{Q}) = \mathbb{E}_{f \sim \mathcal{Q}} \left[\mathbb{P}_{h \sim \mathbf{p}_f}[h(\mathbf{x}) \neq f(\mathbf{x})] \right]$$

  The overall error rate of $\mathbf{p}$ is defined as follows:

  $$err(\mathbf{p}; \mathcal{P}, \mathcal{Q}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{P}} \left[\mathcal{E}(\mathbf{x}, \mathbf{p}; \mathcal{Q}) \right]$$

- **Def 2.** (Average Individual Fairness, AIF)

  We say $\mathbf{p}$ satisfies "(α, β)-AIF" w.r.t. $(\mathcal{P}, \mathcal{Q})$ if there exists $\gamma \geq 0$ s.t.:

  $$\mathbb{P}_{\mathbf{x} \sim \mathcal{P}} \left(|\mathcal{E}(\mathbf{x}, \mathbf{p}; \mathcal{Q}) - \gamma| > \alpha \right) \leq \beta$$

Method

- **Fair Learning Problem subject to $(\alpha, 0)$-AIF**

$$\min_{\mathbf{p}, \gamma \in [0,1]} \quad \mathit{err}(\mathbf{p}; \mathcal{P}, \mathcal{Q})$$

$$\text{s.t. } \forall \mathbf{x} \in \mathcal{X}: \quad |\mathcal{E}(\mathbf{x}, \mathbf{p}; \mathcal{Q}) - \gamma| \leq \alpha$$

Method-Empirical version

- The empirical versions of the overall error rate and the individual error rates can be expressed as:

$$err(\mathbf{p}; \hat{\mathcal{P}}, \hat{\mathcal{Q}}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(\mathbf{x}_i, \mathbf{p}; \hat{\mathcal{Q}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \mathbb{P}_{h_j \sim p_j}[h_j(\mathbf{x}_i) \neq f_j(\mathbf{x}_i)]$$

- **Empirical Fair Learning Problem**

$$\min_{\mathbf{p}, \gamma \in [0,1]} \quad err(\mathbf{p}; \hat{\mathcal{P}}, \hat{\mathcal{Q}})$$

$$\text{s.t. } \forall \mathbf{x} \in \mathcal{X} : \; |\mathcal{E}(\mathbf{x}, \mathbf{p}; \hat{\mathcal{Q}}) - \gamma| \leq \underbrace{2\alpha}_{\text{slightly relaxed}}$$