Constrained Fairness AI Reviews

July 2, 2020

Table of Contents

1. Fairness measure

2. Methodologies

   Fairness-aware Classifier with Prejudice Remover Regularizer (2012)

   Fairness Constraints: A Flexible Approach for Fair Classification (2019)

   Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (2019)

   Average Individual Fairness: Algorithms, Generalization and Experiments (2019)

Notations

- Consider a binary classification task.
- For $i = 1, \ldots, n$, we have a sample $(\mathbf{x}_i, z_i, y_i) \sim \mathcal{D}$ where
  - $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a feature vector;
  - $z_i \in \mathcal{Z}$ is a sensitive feature;
  - $y_i \in \{-1, 1\}$ is the corresponding class label;
  - $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Z} \times \{-1, 1\}$.
- For a mapping $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ parametrized $\boldsymbol{\theta}$, $\hat{y} = 1$ if $f_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0$ and $\hat{y} = -1$ otherwise.
- Here we consider $\mathcal{Z} = \{0, 1\}$.

Fairness Measure

- No disparate treatment (no direct discrimination):
  $P(\hat{y}|\mathbf{x}, z=0) = P(\hat{y}|\mathbf{x}, z=1)$
  (Note that, if $z \notin \mathbf{x}$, the resulting classifier does not suffer from disparate treatment since z is not used during test.)
- No disparate impact (statistical parity or demographic parity):
  $P(\hat{y}=1|z=0) = P(\hat{y}=1|z=1)$
- Equalized Odds: $P(\hat{y}=1|y, z=1) = P(\hat{y}=1|y, z=0), \ \forall y \in \{-1, 1\}$
- No disparate mistreatment:

$$P(\hat{y} \neq y|z=0) = P(\hat{y} \neq y|z=1) \qquad \text{(Error rate)}$$

$$P(\hat{y} \neq y|y=-1, z=0) = P(\hat{y} \neq y|y=-1, z=1) \qquad \text{(False positive rate)}$$

$$P(\hat{y} \neq y|y=1, z=0) = P(\hat{y} \neq y|y=1, z=1) \qquad \text{(False negative rate)}$$

Fairness Constrainted Classification

- For fair classification,

$$\text{minimize } L(\boldsymbol{\theta}) \qquad \} \text{ Classificer loss function}$$

$$\text{subject to } P_{\boldsymbol{\theta}}(\cdot|z=0) = P_{\boldsymbol{\theta}}(\cdot|z=1) \qquad \} \text{ Fairness constraints,}$$

where

- $\boldsymbol{\theta}$: a set of parameters for a classifier;
- $L(\boldsymbol{\theta})$: a loss function

Table of Contents

1. Fairness measure

2. Methodologies

Fairness-aware Classifier with Prejudice Remover Regularizer (2012)

Fairness Constraints: A Flexible Approach for Fair Classification (2019)

Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (2019)

Average Individual Fairness: Algorithms, Generalization and Experiments (2019)

Prejudice Remover Regularizer (2012)

Introduction

- **Prejudice** means a statistical dependence between a sensitive variable Z and target variable Y or a non-sensitive variable $\mathbf{X}$.
  - Direct prejudice: the use of a sensitive variable in a prediction model (equivalent to 'direct discrimination')
  - Indirect prejudice: statistical dependence between Y and Z
  - Latent prejudice: statistical dependence between $\mathbf{X}$ and Z
- In this paper, authors focus on 'indirect prejudice' and develop a technique to reduce it.

- Main idea: Use a approximation of the mutual information between Y and Z called by Prejudice Index (PI) as the Fairness regularizer

$$\mathrm{PI} = \sum_{y \in \{-1,1\}} \sum_{z \in \mathcal{Z}} P(y,z) \log \frac{P(y,z)}{P(y)P(z)} \qquad \text{(Prejudice Index)}$$

- Suppose that we have a prediction model $M_{\boldsymbol{\theta}}(y|\mathbf{x}, z)$. For example, in case of logistic regression, we used

$$M_{\boldsymbol{\theta}}(y|\mathbf{x}, z) = y\sigma(\mathbf{x}^{\top}\mathbf{w}_z) + (1-y)(1-\sigma(\mathbf{x}^{\top}\mathbf{w}_z)),$$

where $\sigma(\cdot)$ is a sigmoid function, and $\boldsymbol{\theta} = \{\mathbf{w}_z\}_{z \in \mathcal{Z}}$.

- To derive a approximation of PI, define

$$P_{\boldsymbol{\theta}}(Y, \mathbf{X}, Z) = M_{\boldsymbol{\theta}}(Y|\mathbf{X}, Z)P(\mathbf{X}, Z) \text{ and } \hat{P}_{\boldsymbol{\theta}}(Y, \mathbf{X}, Z) = M_{\boldsymbol{\theta}}(Y|\mathbf{X}, Z)\hat{P}(\mathbf{X}, Z)$$

  where $P(\mathbf{X}, Z)$ is the joint distribution of $(\mathbf{X}, Z)$ and $\hat{P}(\mathbf{X}, Z)$ is the sample distribution.

- Then,

$$\mathrm{PI} \approx \mathrm{PI}_{\boldsymbol{\theta}} = \sum_{y \in \{-1,1\}} \sum_{z \in \mathcal{Z}} \sum_{\mathbf{x} \in \mathcal{X}} P_{\boldsymbol{\theta}}(y, \mathbf{x}, z) \log \frac{P_{\boldsymbol{\theta}}(y|z)}{P_{\boldsymbol{\theta}}(y)}$$

- Using sample distribution over $\mathbf{x}$ and z, $\mathrm{PI}_{\boldsymbol{\theta}}$ can be calculated by

$$\mathrm{PI}_{\boldsymbol{\theta}} \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \{-1,1\}} M_{\boldsymbol{\theta}}(y|\mathbf{x}_i, z_i) \log \frac{\hat{P}_{\boldsymbol{\theta}}(y|z_i)}{\hat{P}_{\boldsymbol{\theta}}(y)} =: \mathrm{R}_{PR}(\boldsymbol{\theta}),$$

where

$$\hat{P}_{\boldsymbol{\theta}}(y|z) = \frac{\sum_{\{(\mathbf{x}_i, z_i) \in \mathcal{D} \text{ s.t. } z_i = z\}} M_{\boldsymbol{\theta}}(y|\mathbf{x}_i, z)}{|\{(\mathbf{x}_i, z_i) \in \mathcal{D} \text{ s.t. } z_i = z\}|}, \ \hat{P}_{\boldsymbol{\theta}}(y) = \frac{\sum_{(\mathbf{x}_i, z_i) \in \mathcal{D}} M_{\boldsymbol{\theta}}(y|\mathbf{x}_i, z_i)}{|\mathcal{D}|}.$$

Prejudice Remover Regularizer (2012)

Method

- Objective function:

$$\text{Minimize} \quad -\sum_{i=1}^{n} \log M_{\boldsymbol{\theta}}(y_i|\mathbf{x}_i, z_i) + \eta \mathrm{R}_{PR}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2, \qquad (1)$$

where λ and η are positive regularization parameters.

✓ non-convex regularizer

Table of Contents

1. Fairness measure

2. Methodologies

   Fairness-aware Classifier with Prejudice Remover Regularizer (2012)

   Fairness Constraints: A Flexible Approach for Fair Classification (2019)

   Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (2019)

   Average Individual Fairness: Algorithms, Generalization and Experiments (2019)

- Most of fairness constraints are a non-convex function of $\boldsymbol{\theta}$, hence leading to non-convex optimization.
- Main idea: Use proposed covariance measure of decision boundary unfairness, which serve as a tractable **proxy** to several of definitions of unfairness, into fairness constraints.

Decision Boundaray Fairness (2019)

- To design a fair convex boundary-based classifier, they defines a measure of decision boundary fairness:

$$\text{Covariance}(Z, d_{\boldsymbol{\theta}}(\mathbf{X})),$$

where $d_{\boldsymbol{\theta}}(\mathbf{x})$ is the signed distance from the feature vector $\mathbf{x}$ to the decision boundary.

Decision Boundaray Fairness (2019)

Method

- For free of disparate impact
  - Define

$$\begin{aligned}\text{Cov}_{DI}(Z, d_{\boldsymbol{\theta}}(\mathbf{X})) &= \mathbb{E}[(Z - \bar{Z})d_{\boldsymbol{\theta}}(\mathbf{X})] - \mathbb{E}[(Z - \bar{Z})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{X}) \\ &= \mathbb{E}[(Z - \bar{Z})d_{\boldsymbol{\theta}}(\mathbf{X})] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\boldsymbol{\theta}}(\mathbf{x}_i)\end{aligned} \tag{2}$$

  - If a decision boundary has no disparate impact, i.e., $P(d_{\boldsymbol{\theta}}(\mathbf{X}) \geq 0 | Z = 0) = P(d_{\boldsymbol{\theta}}(\mathbf{X}) \geq 0 | Z = 1)$, then $\text{Cov}_{DI}(Z, d_{\boldsymbol{\theta}}(\mathbf{X})) = 0$.
  - Note that the converse is not true, hence we call this covariance measure a proxy.

- To train a classifier free of disparate impact,

$$\begin{aligned} &\text{minimize } L(\boldsymbol{\theta}) \\ &\text{subject to } \left| \frac{1}{N} \sum_i (z_i - \bar{z}) d_{\boldsymbol{\theta}}(\mathbf{x}_i) \right| \leq c, \end{aligned} \tag{3}$$

whre $c > 0$ is a given threshold.

- For free of disparate mistreatment
  - Consider overall misclassification rate:

$$\begin{aligned}\text{Cov}_{OMR}(Z, g_{\boldsymbol{\theta}}(Y, \mathbf{X})) &= \mathbb{E}[(Z - \bar{Z})(g_{\boldsymbol{\theta}}(Y, \mathbf{X}) - \bar{g}_{\boldsymbol{\theta}}(Y, \mathbf{X}))] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})g_{\boldsymbol{\theta}}(y_i, \mathbf{x}_i),\end{aligned} \tag{4}$$

    where $g_{\boldsymbol{\theta}}(y, \mathbf{x}) = \min(0, y d_{\boldsymbol{\theta}}(\mathbf{x}))$.
  - If a decision boundary has no disparate mistreatment w.r.t. OMR, then $\text{Cov}_{OMR}(Z, d_{\boldsymbol{\theta}}(\mathbf{X})) = 0$.

Decision Boundaray Fairness (2019)

Method

- In contrast to the covariance measure for disparate impact, Cov_{OMR} is not convex.
- Fortunately, it can be easily converted into convex-concave constraints, and then apply a Disciplined Convex-Concave Programe (DCCP).

✓ convex optimization, proxy contraints, restrictions on other fairness measure

Table of Contents

1. Fairness measure

2. Methodologies

Fairness-aware Classifier with Prejudice Remover Regularizer (2012)

Fairness Constraints: A Flexible Approach for Fair Classification (2019)

Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (2019)

Average Individual Fairness: Algorithms, Generalization and Experiments (2019)

- To suggest fair classifier with provable guarantees
  1. convert various fairness measures to linear-fractional group performance functions;
  2. based on (1), derive an optimal solution for the classification problem with fairness constraints and its is a form of $\mathbf{I}(s_{\lambda^\star}(\mathbf{x}) > 0)$ where s_λ and $\lambda^\star$ will be specified.

Meta algorithm for Fairness Constraints (2019)

Definitions

- Suppose that $Z \in \{1, \dots, p\}$.
- Fixing different values of Z partitions the domain $D = \mathcal{X} \times \mathcal{Z} \times \{-1, 1\}$ into p groups
$$G_i := \{(\mathbf{x}, i, y) \in D\}.$$
- **Definition 2.2** (Group performance function)
For any $f \in \mathcal{F}$, define a group performance function $q : \mathcal{F} \to [0, 1]^p$ as $q(f) = (q_1(f), \dots, q_p(f))$ where $q_i(f) = P[\mathcal{E} | G_i, \mathcal{E}']$ for some events $\mathcal{E}, \mathcal{E}'$.
- **Example.** For false positive rate with $\mathcal{E} := (f = 1)$ and $\mathcal{E}' := (Y = 0)$, $q_i(f) = P(f = 1 | G_i, Y = 0)$.

- **Definition 2.3** (Linear-fractional group performance functions, $\mathcal{Q}_{\text{linf}}$)

  A group performance function q is called linear-fractional if for any $f \in \mathcal{F}$ and $i \in [p]$, $q_i(f)$ can be rewritten as

$$q_i(f) = \frac{\alpha_0^{(i)} + \sum_{r=1}^{k} \alpha_r^{(i)} \cdot \Pr\left[f = 1 \mid G_i, \mathcal{A}_r^{(i)}\right]}{\beta_0^{(i)} + \sum_{r=1}^{l} \beta_r^{(i)} \cdot \Pr\left[f = 1 \mid G_i, \mathcal{B}_r^{(i)}\right]} \quad (5)$$

  for two integers $k, l \geq 0$, events $\mathcal{A}_1^{(i)}, \ldots, \mathcal{A}_k^{(i)}, \mathcal{B}_1^{(i)}, \ldots, \mathcal{B}_l^{(i)}$ that are independent of the choice of f, and parameters $\alpha_0^{(i)}, \ldots, \alpha_k^{(i)}, \beta_0^{(i)}, \ldots, \beta_l^{(i)} \in \mathbb{R}$ that are independent of the choice of f.
- If $l = 0$ and $\beta_0^{(i)} = 1$ for all i, q is said to be linear, denoted by $\mathcal{Q}_{\text{lin}}$.

Meta algorithm for Fairness Constraints (2019)

Definitions

		$q_i(f)$		L/LF
		$\mathcal{E}$	$\mathcal{E}'$	
fairness defn.	statistical	$f=1$	$\emptyset$	$\mathcal{Q}_{\text{lin}}$
	conditional statistical	$f=1$	$X \in S$	$\mathcal{Q}_{\text{lin}}$
	false positive	$f=1$	$Y=0$	$\mathcal{Q}_{\text{lin}}$
	false negative	$f=0$	$Y=1$	$\mathcal{Q}_{\text{lin}}$
	true positive	$f=1$	$Y=1$	$\mathcal{Q}_{\text{lin}}$
	true negative	$f=0$	$Y=0$	$\mathcal{Q}_{\text{lin}}$
	accuracy	$f=Y$	$\emptyset$	$\mathcal{Q}_{\text{lin}}$
	false discovery	$Y=0$	$f=1$	$\mathcal{Q}_{\text{linf}}$
	false omission	$Y=1$	$f=0$	$\mathcal{Q}_{\text{linf}}$
	positive predictive	$Y=1$	$f=1$	$\mathcal{Q}_{\text{linf}}$
	negative predictive	$Y=0$	$f=0$	$\mathcal{Q}_{\text{linf}}$

Figure 1 : Group perfromance functions for different fairness metrics

- **Definition 2.5** (Group-Fair)

  For some fairness constraint, set $\ell_i, u_i \geq 0$ for all $i \in [p]$. Then we consider the classification problem with some fairness constraint:

$$\min_{f \in \mathcal{F}} \Pr[f \neq Y] \\ s.t., \ \ell_i \leq q_i(f) \leq u_i, \ \forall i \in [p]. \tag{Group-Fair}$$

- **Theorem 3.2** (Solution characterization and computation for $q \in \mathcal{Q}_{\mathrm{lin}}$)

  Given any parameters $\ell, u \in [0,1]^p$, there exist optimal Lagrangian parameters $\lambda^\star \in \mathbb{R}^p$ such that $\mathbf{I}[s_{\lambda^\star}(\mathbf{x}) > 0]$ is an optimal fair classifier for Group-Fair.

  Here, $s_\lambda(\mathbf{x}) := \Pr\left[Y = 1 \mid X = \mathbf{x}\right] - 0.5 + \sum_{i \in [p]} \lambda_i \cdot \psi_i(\mathbf{x})$, and $\psi_i(\mathbf{x}) = \sum_{r=1}^{k} \frac{\alpha_r^{(i)}}{\Pr\left[G_i, \mathcal{A}_r^{(i)}\right]} \cdot \Pr\left[G_i, \mathcal{A}_r^{(i)} \mid X = \mathbf{x}\right].$

- **Theorem 3.2** (Solution characterization and computation for $q \in \mathcal{Q}_{\mathrm{lin}}$)

  Given any parameters $\ell, u \in [0,1]^p$, there exist optimal Lagrangian parameters $\lambda^\star \in \mathbb{R}^p$ such that $\mathbf{I}[s_{\lambda^\star}(\mathbf{x}) > 0]$ is an optimal fair classifier for Group-Fair.

  Here, $s_\lambda(\mathbf{x}) := \Pr[Y = 1 \mid X = \mathbf{x}] - 0.5 + \sum_{i \in [p]} \lambda_i \cdot \psi_i(\mathbf{x})$, and $\psi_i(\mathbf{x}) = \sum_{r=1}^{k} \frac{\alpha_r^{(i)}}{\Pr\left[G_i, \mathcal{A}_r^{(i)}\right]} \cdot \Pr\left[G_i, \mathcal{A}_r^{(i)} \mid X = \mathbf{x}\right]$.

  Moreover, $\lambda^\star$ can be computed in polynomial time as a solution to the following convex program:

$$\lambda^\star = \arg\min_{\lambda \in \mathbb{R}^p} \mathbb{E}_X\left[(s_\lambda(X))_+\right] + \sum_{i \in [p]} \left(\alpha_0^{(i)} - u_i\right)\lambda_i + \sum_{i \in [p]} (u_i - \ell_i) \cdot (\lambda_i)_+ . \tag{6}$$

- Note that, $\lambda^\star$ estimated by the stochastic subgradient method.
- Also, for estimates of $\Pr\left[Y = 1 \mid X = \mathbf{x}\right]$, $\Pr\left[G_i, \mathcal{A}_r^{(i)} \mid X = \mathbf{x}\right]$, authors used logistic regression or Gaussian Naivs Bayes.
- Further, authors provide the solution for Group-Fair with $\mathcal{Q}_{\text{linf}}$ and they expanded the algorithms given multiple fairness constraints.

✓ most of fairness measure are contained in $\mathcal{Q}_{\text{linf}}$, rough estimates of $s_\lambda(\mathbf{x})$

Table of Contents

1. Fairness measure

2. Methodologies

Fairness-aware Classifier with Prejudice Remover Regularizer (2012)

Fairness Constraints: A Flexible Approach for Fair Classification (2019)

Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees (2019)

Average Individual Fairness: Algorithms, Generalization and Experiments (2019)

- 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.

Average Individual Fairness (2019)

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: $D = \left\{\mathbf{x}_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$ 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.

- **Definition 2.1** (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:

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

- **Definition 2.2** (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.:

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

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

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

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

Average Individual Fairness (2019)

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**

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

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

Average Individual Fairness (2019)

Method-Empirical version

- We use the dual perspective of constrained optimization: reduce the fair learning task to a two-player game
- First, rewirte the constraints as follows:

$$\mathbf{r}(\mathbf{p}, \gamma; \hat{\mathcal{Q}}) = \begin{bmatrix} \mathcal{E}(\mathbf{x}_i, \mathbf{p}; \hat{\mathcal{Q}}) - \gamma - 2\alpha \\ \gamma - \mathcal{E}(\mathbf{x}_i, \mathbf{p}; \hat{\mathcal{Q}}) - 2\alpha \end{bmatrix}_{i=1}^{n} \in \mathbb{R}^{2n} \tag{7}$$

- Let the corresponding dual variables $\boldsymbol{\lambda} \in \Lambda$, where $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{2n} | \|\boldsymbol{\lambda}\|_1 \leq B\}$ for some $B > 0$.
- To solve fair learning problem, consider the following minimax problem:

$$\min_{\mathbf{p}, \gamma \in [0,1]} \max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}(\mathbf{p}, \gamma, \boldsymbol{\lambda}) \tag{8}$$

Reference

- Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012).
  Fairness-aware classifier with prejudice remover regularizer.
  Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.
- Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019).
  Fairness Constraints: A Flexible Approach for Fair Classification.
  J. Mach. Learn. Res., 20(75), 1-42.
- Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019).
  Classification with fairness constraints: A meta-algorithm with provable guarantees.
  Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 319-328).
- Kearns, M., Roth, A., & Sharifi-Malvajerdi, S. (2019).
  Average individual fairness: Algorithms, generalization and experiments.
  arXiv preprint arXiv:1905.10607.