Loss-Aversively Fair Classification

Junaid Ali, Muhammad Bilal Zafar, Adish Singla, Krishna P. Gummadi

Max Planck Institute for Software Systems

1. Fairness in classification

- Classifiers applied in scenarios with social implications
  - Loan approval, hiring, bail decisions, *etc.*
  - Sensitive feature groups (men, women, *etc.*)
  - Beneficial outcomes (*e.g.,* getting loan)
- Potential for unfairness (many recent examples)
- What constitutes unfairness?
  - Wrongful relative disadvantage [Altman'16]

2. Existing notions: Nondiscrimination

- Parity of benefits between different salient social groups (*e.g.* gender)

$$\mathit{Benefits}_{\male}(\theta) = \mathit{Benefits}_{\female}(\theta),$$

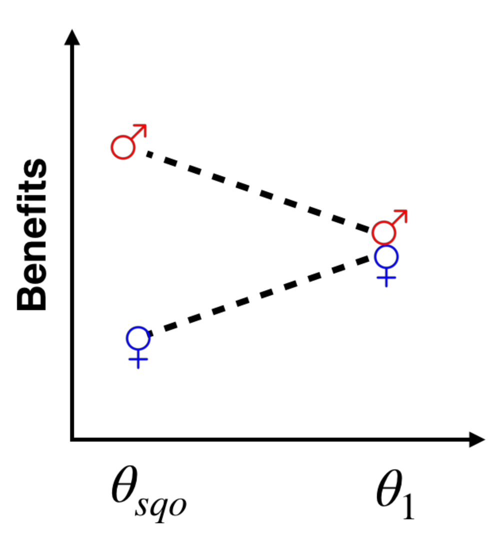
- Statistical parity (SP): Equal acceptance rate for men and women, i.e.,

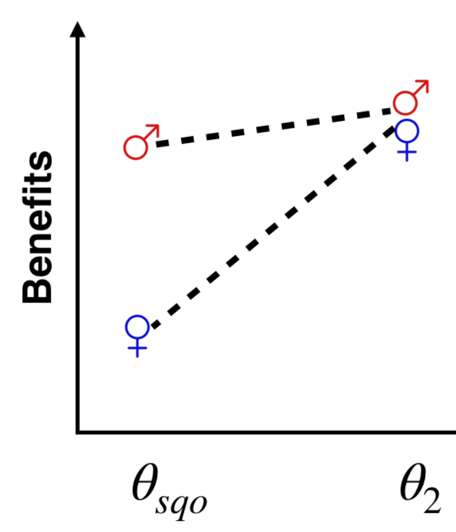
$$P(\hat{y} = 1 | \male) = P(\hat{y} = 1 | \female),$$

- Equality of Opportunity (EOP): Equal true positive rate for men and women, i.e.,

$$P(\hat{y} = 1 | y = 1, \male) = P(\hat{y} = 1 | y = 1, \female).$$

3. Several ways to achieve parity





θ_{sqo} is a discriminatory status quo classifier. θ_1 and θ_2 represent two ways of updating θ_{sqo} with a nondiscriminatory classifier.

- θ_1 achieves nondiscrimination by lowering benefits for men, which might be unacceptable.
- θ_2 equalizes benefits loss-aversively, *i.e.*, by increasing benefits for both the groups.

4. New notion: Loss-averse update

- Inspired by Endowment effect:
  - People ascribe more value to things merely because they own them. [Khaneman *et al* 1990]
- Loss-averse Update:

$$\mathit{Benefits}_{\male}(\theta) \geq \mathit{Benefits}_{\male}(\theta_{sqo}),$$
$$\mathit{Benefits}_{\female}(\theta) \geq \mathit{Benefits}_{\female}(\theta_{sqo}).$$

Key idea: All groups should be at least as well off as in the status quo system.

5. Loss-aversively removing discrimination in classification

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{x,y\in\mathcal{D}} \mathbb{I}(\text{sign}(\theta^T x) = y) && \longrightarrow \text{Accuracy} \\ \text{subject to} \quad & \mathcal{B}_{z=0}(\theta) = \mathcal{B}_{z=1}(\theta) && \longrightarrow \text{Nondiscrimination constraint} \\ & \mathcal{B}_{z=k}(\theta) \geq \mathcal{B}_{z=k}(\theta_{sqo}), \forall k \in \{0,1\}. && \longrightarrow \text{Loss-averse constraint} \end{aligned}$$

SP: Replacing nonconvex objective and constraints with convex proxies.

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} l_\theta(x,y) + \lambda||\theta||^2 \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}|}\left|\sum_{(x,z)\in\mathcal{D}} (z-\bar{z})d_\theta(x_i)\right| < c, \end{aligned}$$
$$\frac{1}{|\mathcal{D}_{z=k}|}\sum_{x\in\mathcal{D}_{z=k}} d_\theta(x) \geq \frac{1}{|\mathcal{D}_{z=k}|}\sum_{x\in\mathcal{D}_{z=k}} d_{\theta_{sqo}}(x) + \gamma,$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+$$

Can accommodate any convex boundary-based classifier (*e.g.*, logistic regression, linear / non-linear SVM)

EOP: Replacing nonconvex objective and constraints with convex proxies.

$$\begin{aligned} \text{minimize} \quad & -\frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} l_\theta(x,y) + \lambda||\theta||^2 \\ \text{subject to} \quad & \frac{1}{|\mathcal{D}_+|}\left|\sum_{(x,z)\in\mathcal{D}_+} (z-\bar{z})d_\theta(x_i)\right| < c, \end{aligned}$$
$$\frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{x\in\mathcal{D}^+_{z=k}} d_\theta(x) \geq \frac{1}{|\mathcal{D}^+_{z=k}|}\sum_{x\in\mathcal{D}^+_{z=k}} d_{\theta_{sqo}}(x) + \gamma,$$
$$\text{for all } k \in \{0,1\}, \gamma \in \mathbb{R}^+$$

6. Evaluation

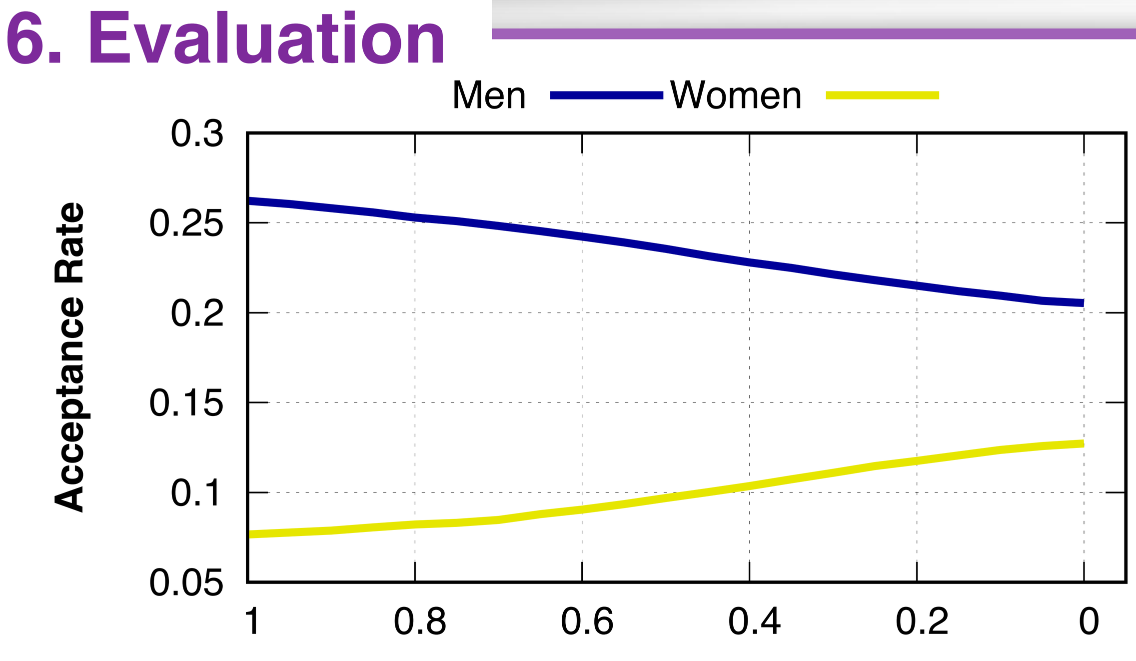


Figure 1: Statistical parity

Maximizing accuracy subject to nondiscrimination constraint lowers benefits for men.

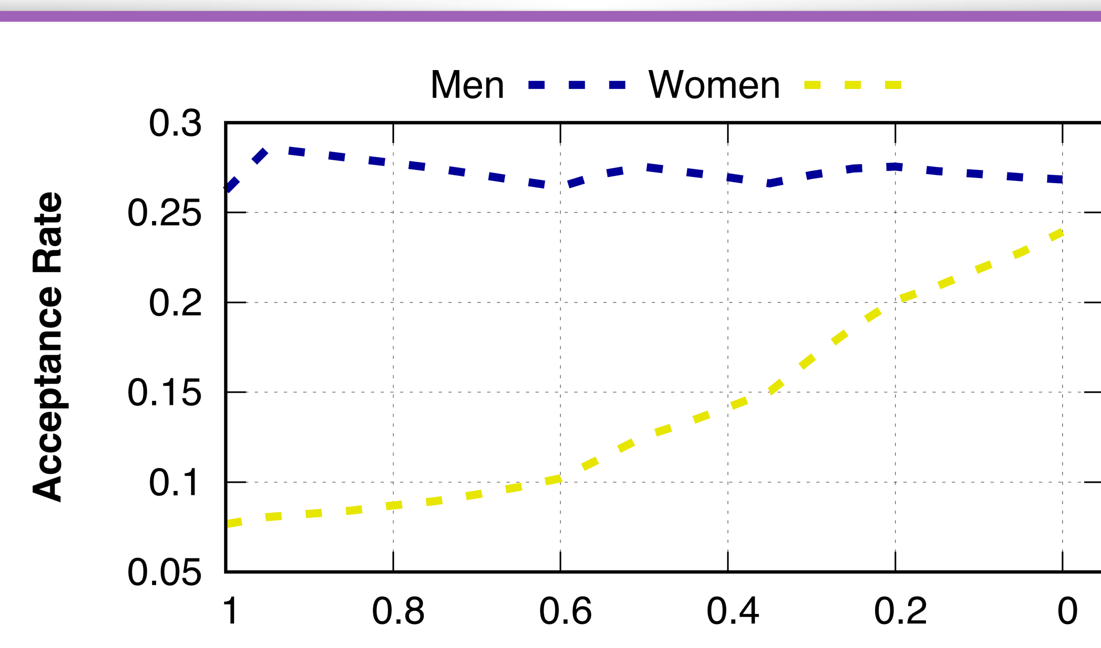


Figure 2: Statistical parity + loss-averse

Adding loss-averse constraint achieves nondiscrimination without lowering benefits for men.

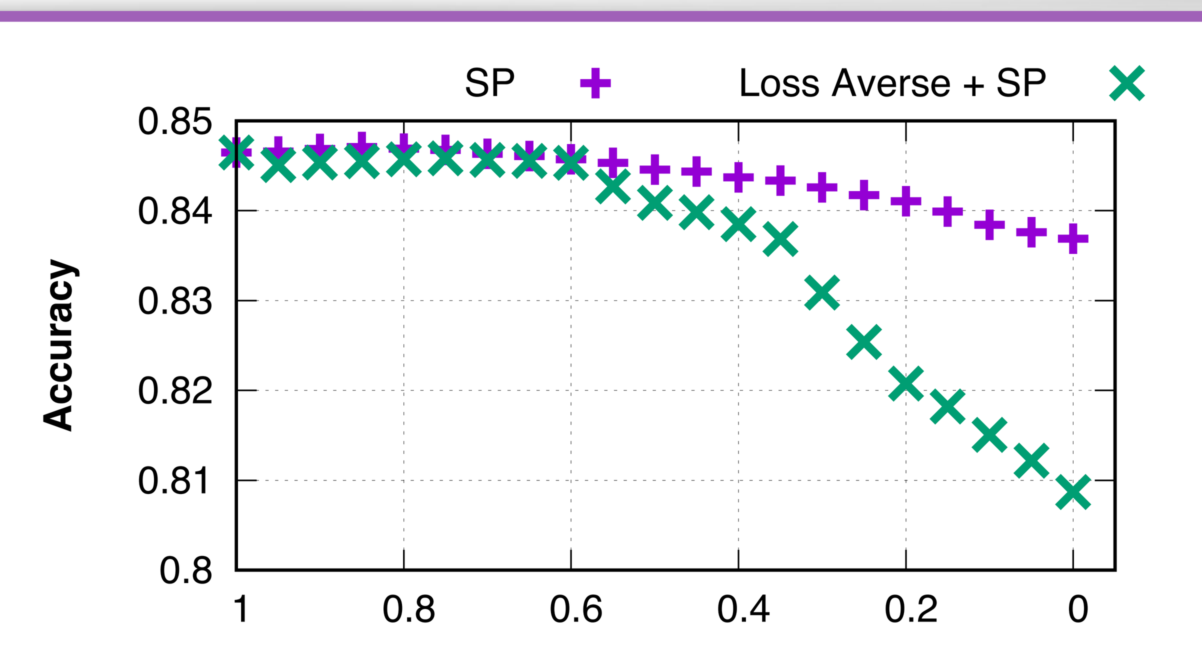


Figure 3: Accuracy fairness tradeoff

As expected, adding additional loss-averse constraint results in further loss in accuracy.

Dataset:

Adult data: UCI

- X-axis is the normalized covariance threshold between the sensitive attribute and the distance from decision boundary, which is used as a proxy for discrimination.
- Y-axis, in figures 1 and 2, shows acceptance rates, *i.e.*, fraction predicted to be in higher salary class.