From Parity to Preference-based Notions of Fairness in Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, Adrian Weller
Max Planck Institute for Software Systems, Max Planck Institute for Intelligent Systems, University of Cambridge

1. Data driven decision making

- Classifiers automate human decision making
  - Learn from past decisions made by humans
- Outcomes 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 as unfairness?
  - Wrongful relative disadvantage [Altman'16]

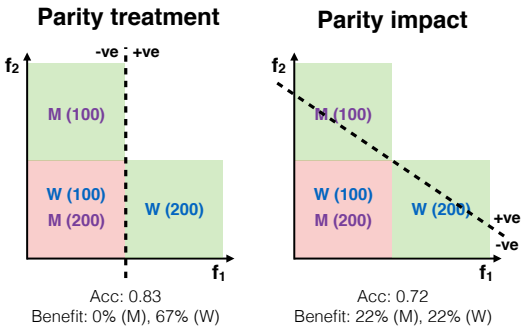
2. Existing notions of fairness

- Based on parity, *i.e.*, equality of treatment or impact
- Sensitive feature value $z \in \mathcal{Z}$, classifier $\boldsymbol{\theta}_z$
- Group benefit: Exp. getting beneficial outcome

$$\mathcal{B}_z(\boldsymbol{\theta}_z) = \mathbb{E}_{\boldsymbol{x}|z}[\mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z(\boldsymbol{x})) = 1\}]$$

- Parity treatment: $\mathcal{B}_z(\boldsymbol{\theta}_z) = \mathcal{B}_z(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z'$
- Parity impact: $\mathcal{B}_z(\boldsymbol{\theta}_z) = \mathcal{B}_{z'}(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z'$
- No parity → Wrongful relative disadvantage
  - Is parity the only criterion of fairness?

3. Parity can be a stringent criterion



Few benefits
Low accuracy

4. New notions of fairness

Preferred treatment

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z'$$

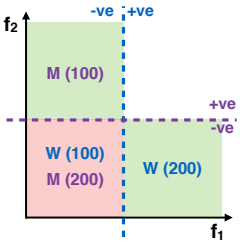
(Inspired by **Envy-freeness**)

Preferred impact

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}'_z) \quad \text{for all } z$$

$\boldsymbol{\theta}'_z$: Parity impact classifier
(Inspired by **Bargaining Solution**)

Key Idea: All groups *prefer* their respective outcomes despite disparity



More benefits
High accuracy

5. Training preferentially fair classifiers

Goal: Maximize accuracy subject to preferred treatment criterion (similar procedure for preferred impact)

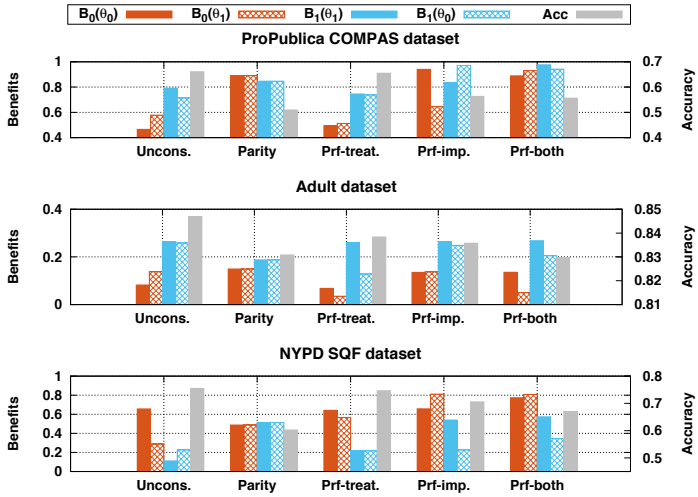
$$\begin{aligned} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} \quad & -\tfrac{1}{N}\textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T\boldsymbol{x}) = y\} \\ \text{subject to} \quad & \textstyle\sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_z^T\boldsymbol{x}) = 1\} \geq \sum_{\boldsymbol{x}\in\mathcal{D}_z} \mathbb{I}\{\text{sign}(\boldsymbol{\theta}_{z'}^T\boldsymbol{x}) = 1\} \quad \text{for all } z, z' \end{aligned} \longrightarrow \begin{aligned} \underset{\{\boldsymbol{\theta}_z\}}{\text{minimize}} \quad & -\tfrac{1}{N}\textstyle\sum_{(\boldsymbol{x},y,z)\in\mathcal{D}} \ell_{\boldsymbol{\theta}_z}(\boldsymbol{x}, y) + \sum_{z\in\mathcal{Z}} \lambda_z \Omega(\boldsymbol{\theta}_z) \\ \text{subject to} \quad & \textstyle\sum_{\boldsymbol{x}\in\mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T\boldsymbol{x}) \geq \sum_{\boldsymbol{x}\in\mathcal{D}_z} \max(0, \boldsymbol{\theta}_{z'}^T\boldsymbol{x}) \quad \text{for all } z, z' \end{aligned}$$

Both objects and constraints non-convex
Hard to solve efficiently

Convex objective, convex-concave constraints
Efficient solution procedures (DCCP) [Shen'16]

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

6. Evaluation and discussion



Datasets
- ProPublica COMPAS data: African-American (0) & White (1)
- Adult data: Female (0) & Male (1)
- NYPD SQF data: African-American (0) & White (1)

Evaluation takeaways
- Preferential fairness leads to higher accuracy
- Higher group benefits as compared to parity

Insight: Preferential fairness subsumes parity fairness
- Each parity treatment classifier also satisfies preferred treatment
- Each parity impact classifier also satisfies preferred impact
- Theoretically, preferential fairness allows for more accurate solutions

Moving forward
- Individual- vs. group-level preferences
- Beyond convex boundary-based classifiers

Paper and code at: fate-computing.mpi-sws.org