Generalizing Group Fairness via Utilities

UIC COMPUTER SCIENCE

Jack Blandin, Ian Kash
University of Illinois at Chicago

Motivation

Symptom

Numerous bespoke interpretations of group fairness definitions exist as attempts to extend them to specific applications.

Problem

Group fairness definitions assume a classification setting.

Solution

Use utility functions to define group fairness.

- Utility functions generalize better than classification variables.
- In addition to the decision-maker's utility function, make use of a *benefit function* that represents the individual's utility from encountering a given decision-maker policy.
- Generalize "qualification" as the existence of a mutually beneficial outcome for both the decision-maker and the individual.

Fairness in Classification

Demographic Parity

$$P(\hat{Y}=1 \mid Z=0) = P(\hat{Y}=1 \mid Z=1)$$

Equal Opportunity

$$P(\hat{Y}=1 \mid Y=1, Z=0) = P(\hat{Y}=1 \mid Y=1, Z=1)$$

Limiting Assumptions

Classification group fairness definitions usually make the following limiting assumptions:

1. **Equal predictions have equal outcomes.** Counter example: loan applications.
2. **Observed values of the target variable are independent of predictions.** Counter example: recidivism prediction for prison sentencing.
3. **The objective is to predict some unobserved target variable.** Counter example: reinforcement learning or clustering applications.
4. **Decisions for one individual do not impact other individuals.** Counter example: Drawing congressional district boundaries (via clustering).

Classification vs Utility Fairness Definitions

Classification Demographic Parity

Prediction; "Positive" Prediction

$$P(\hat{Y}=1 \mid Z=0) = P(\hat{Y}=1 \mid Z=1)$$

Probability of getting positive prediction

Benefit Demographic Parity

benefit function; min benefit for "positive" outcome

$$P(W \geq \tau \mid Z=0) = P(W \geq \tau \mid Z=1)$$

Probability of getting positive outcome

Classification Equal Opportunity

$$P(\hat{Y}=1 | Y=1, A=0) = P(\hat{Y}=1 | Y=1, A=1)$$

qualification indicator

Counterfactual Utility Equal Opportunity

$$P(W \geq \tau \mid \Gamma=1, Z=0) = P(W \geq \tau \mid \Gamma=1, Z=1)$$

mutually beneficial outcome indicator →

$$\Gamma = \begin{cases} 1 & \text{if } \exists \hat{Y}' : W_{\hat{Y}'} \geq \tau \wedge C_{\hat{Y}'} \leq \rho \\ 0 & \text{otherwise} \end{cases}$$

Generalizing Interpretation of "Qualified"

Classification Equal Opportunity

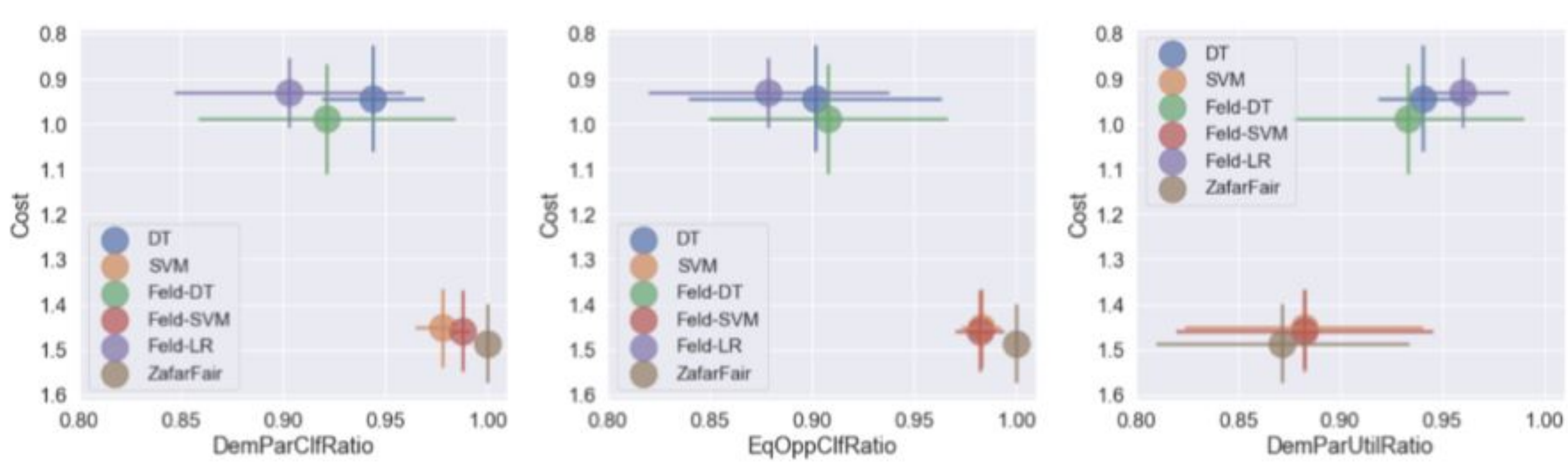
The probability that a qualified individual receives the beneficial outcome is independent of the individual's protected attribute.

Counterfactual Utility Equal Opportunity

For the subset of individuals where there exists a mutually beneficial outcome for both the individual and the decision-algorithm, the probability that a beneficial individual outcome occurring is independent of the individual's protected attribute.

Applications

Prediction-Outcome Disconnect for Loan Applications (German Credit Dataset)



Self-Fulfilling Prophecies with Recidivism Prediction

		$P(Y=1\|\hat{Y}=1)=0$	$P(Y=1\|\hat{Y}=1)=1$
		Dangerous	Backlash
$P(Y=1\|\hat{Y}=0)=0$	Detained	Unq	CfUtil
	Released	Unq	Clf, CfUtil
		Preventable	Safe
$P(Y=1\|\hat{Y}=0)=1$	Detained	Clf	Clf, CfUtil
	Released	Unq	Clf, CfUtil

References

- Blandin, J. and Kash, I. Fairness through counterfactual utilities. arXiv preprint arXiv:2108.05315, 2021.
- Imai, Kosuke, and Zhichao Jiang. "Principal Fairness for Human and Algorithmic Decision-Making." arXiv preprint arXiv:2005.10400, 2020.
- Dua, D., & Graff, C. (2017). Uci machine learning repository.