Statistical Inference for Individual Fairness

[1]Subha Maity, [1]Songkai Xue, [2]Mikhail Yurochkin, [1]Yuekai Sun

[1]Department of Statistics, University of Michigan
[2]IBM Research, MIT-IBM Watson AI Lab

Introduction

ProPublica and Gender Shades studies show violations of group fairness by ML systems deployed in practice.

Goal: Assess individual fairness (IF) of ML systems.

Contributions:

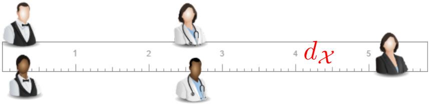
- a gradient flow algorithm to identify IF violations
- a statistically calibrated tool for detecting individual bias

Individual Fairness

In supervised learning it means **similar individuals (inputs to a model) should be treated (outputs of a model) similarly** (Dwork et al. 2012).

Similarity:

- fair metric $d_{\mathcal{X}}$ for individuals (input)
- prediction loss ℓ for outputs



IF Violation? Look at

$$\hat{\mu}_n = \mathbb{E}_n(\text{loss-ratio})_i = \mathbb{E}_n\left[\frac{\ell(f(x_i(T), y_i)}{\ell(f(x_i), y_i)}\right],$$

where $x_i(T)$ is IF-violated and similar to x_i.

Measuring Individual Bias

Idea: Measure individual bias with average loss ratio between pairs with IF violations, i.e.,

$$\ell(f(x'), y)/\ell(f(x), y).$$

Problem: Similar individuals with fairness violation are hard to come by in the data.

Solution: Generate fairness violated individual by finding maximal loss among similar individuals in terms of fair metric.

$$\max_{x' \in \mathcal{X}} \left\{ \ell\left(f(x'), y\right) - \lambda d_{\mathcal{X}}^{2}(x, x') \right\}$$

Measuring Individual Bias

Problems:

- difficult to solve for non-convex model f
- limiting distribution of the test statistics is difficult to characterize

Solution: Generate IF-violated individuals by early stopping with gradient ascent.

$$\partial_t x(t) = \nabla_{x(t)} \left\{ \ell\left(f(x(t)), y\right) - \lambda d_{\mathcal{X}}^2(x, x(t)) \right\} \text{ with } x(0) = x;$$

$$\text{IF-violated individual} \triangleq x(T)$$

Advantages: (1) computationally tractable; (2) $x \mapsto x(T)$ is smooth w.r.t. x.

- Finally, We measure IF violation with

$$\hat{\mu}_n = \mathbb{E}_n(\text{loss-ratio})_i = \mathbb{E}_n\left[\frac{\ell(f(x_i(T)), y_i)}{\ell(f(x_i), y_i)}\right]$$

Detecting Individual Bias

The (population) average loss ratio should not be much larger than one for an individually fair algorithm.

False alarm controlled tool? Statistically test

$$H_0 : \mathbb{E}\big[\,\text{loss-ratio}\,\big] \leq 1 + \varepsilon \quad \text{vs} \quad H_1 : \mathbb{E}\big[\,\text{loss-ratio}\,\big] > 1 + \varepsilon$$

Theorem (Asymptotic distribution) The central limit convergence holds for average loss ratio, i.e.,

$$\sqrt{n}\left(\frac{\hat{\mu}_n - \mathbb{E}\left[\text{loss-ratio}\right]}{\widehat{\text{sd}}(\text{loss-ratio})}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

Detection tool with ≈ 0.05 false alarm rate:

$$\text{individually biased if} \quad T_n = \hat{\mu}_n - 1.645 \times \frac{\widehat{\text{sd}}(\text{loss-ratio})}{\sqrt{n}} > 1 + \varepsilon.$$

Alternative Approach

Idea: measure individual bias by comparing (sample) prediction errors for fairness violated and original individuals.

$$\text{error-ratio}(\mathbb{P}_n) = \frac{\text{proportion of } \{\hat{f}(x_i(T)) \neq y_i\}}{\text{proportion of } \{\hat{f}(x_i) \neq y_i\}}$$

Pros: easy to interpret

Cons: harder to detect IF violation

Case Study: Adult

Task: predict if earning $\geq \$50$k with age, education, working hours per week, etc.

Sensitive attributes: sex and race

Table 1. Results over 10 iterations

	balanced acc	AOD_{gen}	AOD_{race}	Entropy loss T_n	Entropy loss reject prop	0-1 loss $\widetilde{T}_n$	0-1 loss reject prop
Baseline	0.817	-0.151	-0.061	3.676	1.0	2.262	1.0
Project	**0.825**	-0.147	-0.053	1.660	0.9	1.800	0.8
Reduction	0.800	**0.001**	**-0.027**	5.712	1.0	3.275	1.0
SenSR	0.765	-0.074	-0.048	**1.021**	**0.0**	**1.081**	**0.0**

Reduction enforces group fairness by sacrificing individual fairness. On the contrary SenSR shows improvement in both individual and group fairness.

Key takeaway: Our detection tool correctly identifies individual bias in an ML system.

Thank You!

Poster ID: #1665

Poster Session: #5 (May 4, 2021, 9:00 – 11:00 a.m. PDT)

Contact Information:

- Subha Maity, `smaity@umich.edu`
- Songkai Xue, `sxue@umich.edu`
- Mikhail Yurochkin, `mikhail.yurochkin@ibm.com`
- Yuekai Sun, `yuekai@umich.edu`