Multiaccurate Proxies for Downstream Fairness

Emily Diana

ediana@wharton.upenn.edu

THANK YOU TO MY COLLABORATORS

Wesley Gill

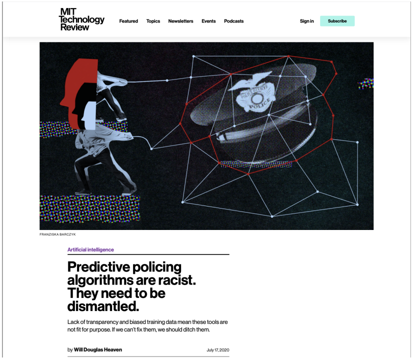
Michael Kearns

Krishnaram Kenthapadi

Aaron Roth

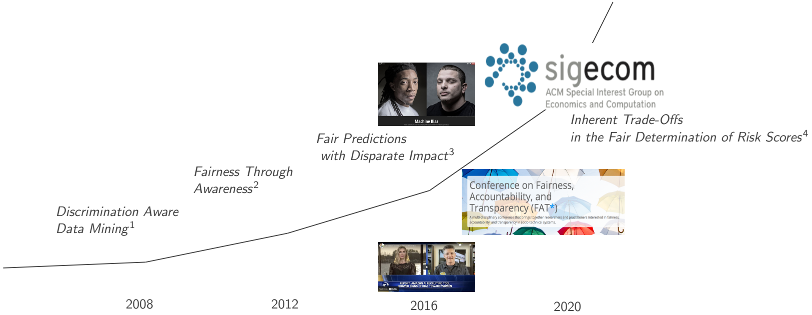
Saeed Sharifi-Malvajerdi

ALGORITHMIC FAIRNESS IN THE NEWS



Algorithmic fairness aims to understand and prevent bias in machine learning models.

ALGORITHMIC FAIRNESS IN THE LITERATURE



Challenges: How do we decide which definitions to use? How do we decide what constitutes harm? When and how do we intervene? How do we balance trade-offs?

[1] Pedreshi, Ruggieri, and Turini. Discrimination-Aware Data Mining. KDD '08

[2] Dwork, Hardt, Pitassi, Reingold, and Zemel. Fairness Through Awareness. ITCS '12.

[3] Chouldehova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data '17

[4] Kleinberg, Mullainathan, and Raghavan. Inherent Trade-offs in the Fair Determination of Risk Scores. '18

CHALLENGE

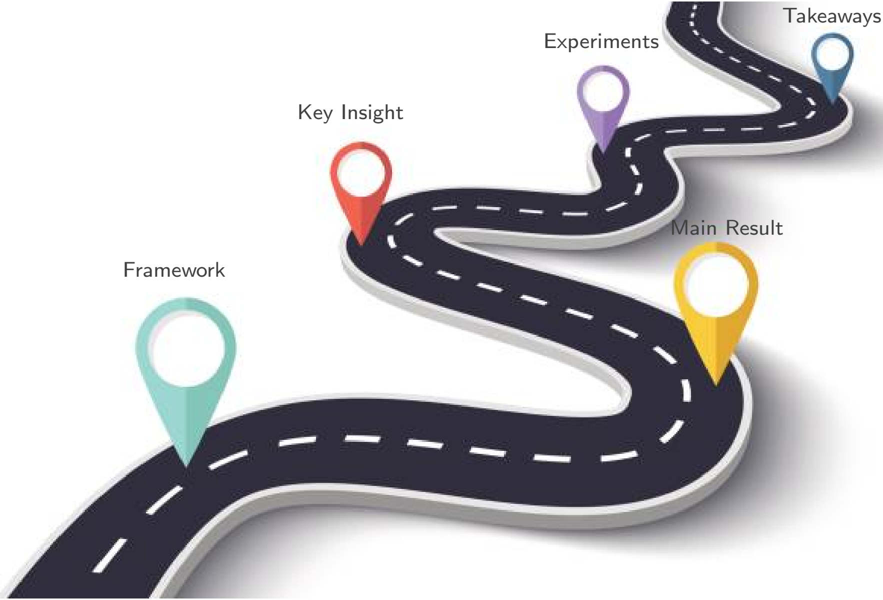
- Often one wants to train a model that is fair with respect to a sensitive feature that has been redacted from training data.
- Could be for legal[5] or policy reasons[6].

Question: How do we make a model fair with respect to race if we don't have data about race?

[5] In the United States it is against the law to use race as an input to consumer lending models

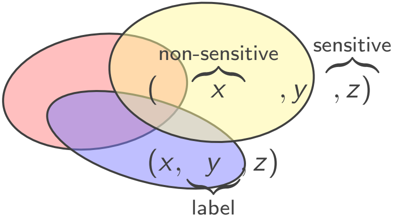
[6] Many large consumer-facing organizations choose not to ask their customers for such information.

PLAN



FRAMEWORK

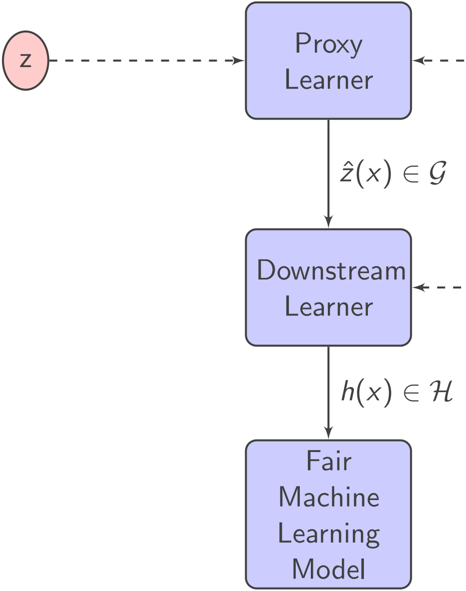
- Data domain $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ divided into K groups



- Proxy model class $\mathcal{G} : \mathcal{X} \to \mathbb{R}^K$
- Proxy $\hat{z} \in \mathcal{G}$ is a vector of K real numbers $(\hat{z}_1, ..., \hat{z}_K)$
- Downstream model class $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$

Proxy Learner aims to find proxy $\hat{z}$ such that if a Downstream Learner trains a model h that is fair with respect to $\hat{z}$, h is also fair with respect to z.

FRAMEWORK


z
Proxy Learner
x,y
$\hat{z}(x) \in \mathcal{G}$
Downstream Learner
$h(x) \in \mathcal{H}$
Fair Machine Learning Model

KEY INSIGHT: PROXY CAN BE REAL VALUED

We can write fairness constraints, usually defined with respect to binary valued group membership using a real valued proxy:

$$\begin{aligned}
\Pr\left[h(x) \neq y | z_k = 1\right] &= \frac{\Pr\left[z_k = 1, h(x) \neq y\right]}{\Pr\left[z_k = 1\right]} \\
&= \frac{\mathbb{E}\left[\mathbb{1}\left[z_k = 1\right]\mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}\left[\mathbb{1}\left[z_k = 1\right]\right]} \\
&= \frac{\mathbb{E}\left[z_k \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}\left[z_k\right]}
\end{aligned}$$

KEY INSIGHT: REPLACE Z WITH $\hat{Z}$

If the following holds:

$$\frac{\mathbb{E}\left[z_k \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}[z_k]} = \frac{\mathbb{E}\left[\hat{z}_k(x) \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}\left[\hat{z}_k(x)\right]}$$

Then if a model is fair with respect to $\hat{z}$

$$\frac{\mathbb{E}\left[\hat{z}_{k_i}(x) \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}[\hat{z}_{k_i}(x)]} = \frac{\mathbb{E}\left[\hat{z}_{k_j}(x) \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}\left[\hat{z}_{k_j}(x)\right]}$$

it also satisfies fairness constraints with respect to the true attribute z.

MAIN RESULT: PROXY DEFINITION

We say $\hat{z}$ is an α-proxy for z if for all classifiers $h \in \mathcal{H}$, and all groups $k \in [K]$,

$$\left| \frac{\mathbb{E}_{(x,z)}\left[z_k \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}_{(x,z)}\left[z_k\right]} - \frac{\mathbb{E}_{(x,z)}\left[\hat{z}_k(x) \mathbb{1}\left[h(x) \neq y\right]\right]}{\mathbb{E}_{(x,z)}\left[\hat{z}_k(x)\right]} \right| \leq \alpha$$

KEY INSIGHT: MULTIACCURACY

Then to learn a proxy, we can solve the linear program:

$$\begin{aligned}
\underset{\hat{z} \in \mathcal{G}}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n} (z_i - \hat{z}(x_i))^2 \\
\text{subject to} \quad & \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} \hat{z}(x_i), \\
& \sum_{i=1}^{n} z_i \mathbb{1}\left[h(x_i) \neq y_i\right] = \sum_{i=1}^{n} \hat{z}(x_i)\mathbb{1}\left[h(x_i) \neq y_i\right], \ \forall h \in \mathcal{H}
\end{aligned} \tag{1}$$

These constraints are multiaccuracy constraints – we want $\hat{z}$ to be an unbiased estimator for z on the set of points where h errs.

STRONG DUALITY AND LOW-REGRET DYNAMICS

Linear Program

$\Updownarrow$

Dual Problem
(min max Lagrangian)

$\Downarrow$

Solution to Game

$$\overbrace{L(\hat{z}, \lambda)}^{\text{Lagrangian}} = \sum_{i=1}^{n} \left[\left(z_i - \underbrace{\overset{\text{primal variable}}{\hat{z}(x_i)}} \right)^2 \right] + \underbrace{\lambda_0}_{\text{dual variable}} \left(\frac{\sum_{i=1}^{n} \hat{z}(x_i)}{\sum_{i=1}^{n} z_i} - 1 \right) + \sum_{h \in \mathcal{H}} \underbrace{\lambda_h}_{\text{dual}} \sum_{i=1}^{n} (z_i - \hat{z}(x_i)) \mathbb{1} [h(x_i) \neq y_i]$$

Under appropriate conditions[7], Solving Program (1) is equivalent to solving:

$$\min_{\hat{z} \in \mathcal{G}} \max_{\lambda} L(\hat{z}, \lambda) = \max_{\lambda} \min_{\hat{z} \in \mathcal{G}} L(\hat{z}, \lambda)$$

[7] Primal variable space is convex and compact, dual variable space is convex, and Lagrangian is convex-concave in primal and dual variables respectively.

ALGORITHM OVERVIEW: NO-REGRET DYNAMICS

Can cast problem as zero-sum game between Learner and Auditor [8]

- Proxy Learner uses Online Projected Gradient Descent to select $\hat{z}$ minimizing $L(\hat{z}, \lambda)$
- Auditor best responds, appealing to an oracle over downstream model class $\mathcal{H}$ to select λ maximizing $L(\hat{z}, \lambda)$

Freund and Schapire show that if a sequence of actions for the two players jointly has low regret, then the uniform distribution over each player's actions forms an approximate equilibrium.

[8] Here we consider the simpler case in which $\hat{z}$ is a linear function in its parameter space, so both $\hat{z}$ and its negation are convex. More details on the non-convex case are provided in the paper.

EXPERIMENTS: OVERVIEW

Simulating a downstream learner, we train a model to be fair with respect to four representations of the sensitive feature and evaluate its performance:

- True Labels: Z
- Baseline Proxy: Logistic regression of Z on X
- **$\mathcal{H}$-Proxy: Solution to Program (1) without squared error objective**
- **MSE Proxy: Solution to Program (1) with squared error objective**

EXPERIMENTS: ACS DATA

Conducted experiments on American Community Survey (ACS) datasets and tasks[9]

Dataset	Samples	$\mathcal{X}$ Dim	Label
ACSEmployment	196104	12	Employment
ACSIncome	**101270**	**4**	**Income** $>$ **\$50K**
ACSIncomePovertyRatio	196104	15	Income-Poverty Ratio $<$ 250%
ACSMobility	39828	17	Same address one year ago
ACSPublicCoverage	71379	15	Health Insurance
ACSTravelTime	89145	8	Commute $>$ 20 minutes

[9] Ding, Hardt, Miller, and Smith. Retiring Adult: New Datasets for Fair Machine. NeuRIPS 2021.

EXPERIMENTS: ACSIncome Race

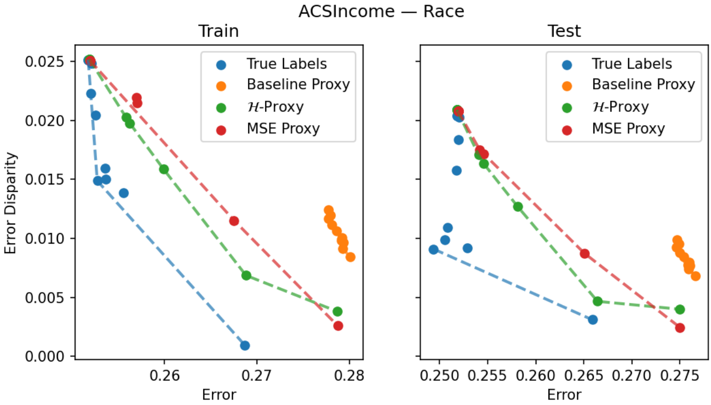


Figure: Proxy results on the ACSIncome dataset with race as sensitive feature

EXPERIMENTS: ACSIncome Age

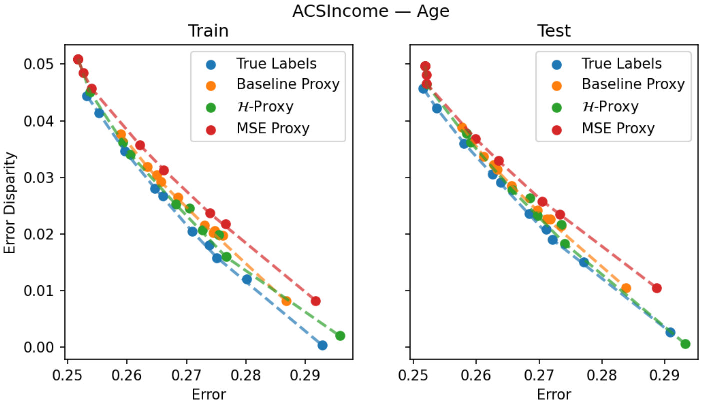


Figure: Proxy results on the ACSIncome dataset with age as sensitive feature

EXPERIMENTS: ACSIncome Sex

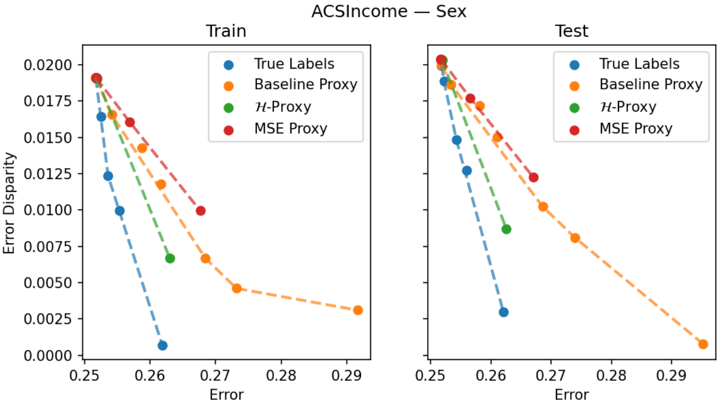


Figure: Proxy results on the ACSIncome dataset with sex as sensitive feature

TAKEAWAYS

- Possible to efficiently train proxies that can stand in for missing sensitive features to effectively train downstream classifiers subject to a variety of demographic fairness constraints.
- Results crucially depend on assumption that the data that the Proxy Learner uses to train its proxy is distributed identically to the data that the Downstream Learner uses.

THANK YOU!

QUESTIONS?

THANK YOU TO MY COLLABORATORS

Wesley Gill

Michael Kearns

Krishnaram Kenthapadi

Aaron Roth

Saeed Sharifi-Malvajerdi