

$$\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\mathbf{x}}[\mathcal{F}], \mathbf{x}) \doteq \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i (f(\mathbf{x}_i) - \hat{\mathbb{E}}_{\mathbf{x}}[f]) \right| \right]$$

An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins

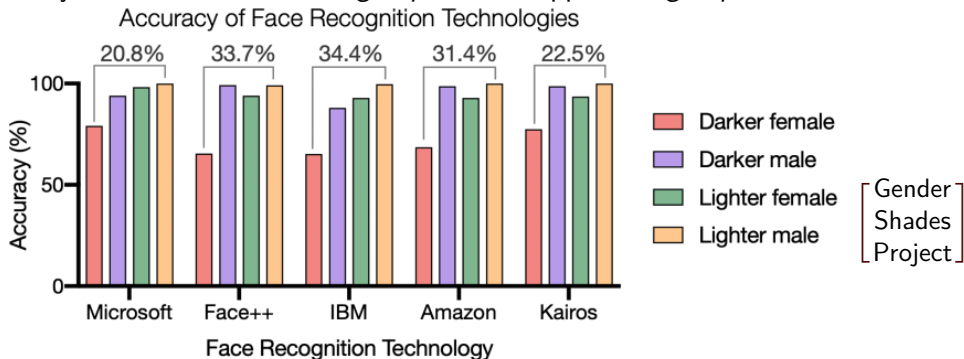
Brown University Department of Computer Science

July 2021

The Twenty-Second ACM Conference on Economics and Computation

Fairness in Machine Learning (or Lack Thereof)

- ML systems often trained on *group A*, then applied to *group B*

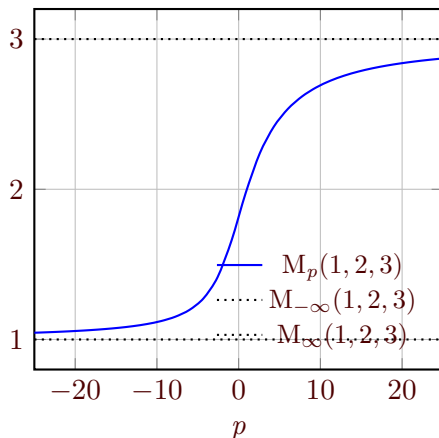


- Differential performance \Rightarrow algorithmic discrimination
 - Facial recognition and policing
 - Speech recognition and accessibility
 - Many more examples

The Power Mean

Suppose vector $\ell = (\ell_1, \dots, \ell_g)$ representing *utility* or *loss* across a population

$$M_p(\ell) \doteq \begin{cases} p = -\infty & \min_{i \in \{1, \dots, g\}} \ell_i \\ p = 0 & \sqrt[n]{\prod_{i=1}^n \ell_i} \\ p = \infty & \max_{i \in \{1, \dots, g\}} \ell_i \\ \text{else} & \sqrt[p]{\frac{1}{n} \sum_{i=1}^n \ell_i^p} \end{cases}$$



- Smooth interpolation between *min*, *arithmetic mean*, and *max*
- Other special cases: *geometric*, *harmonic*, and *quadratic* means

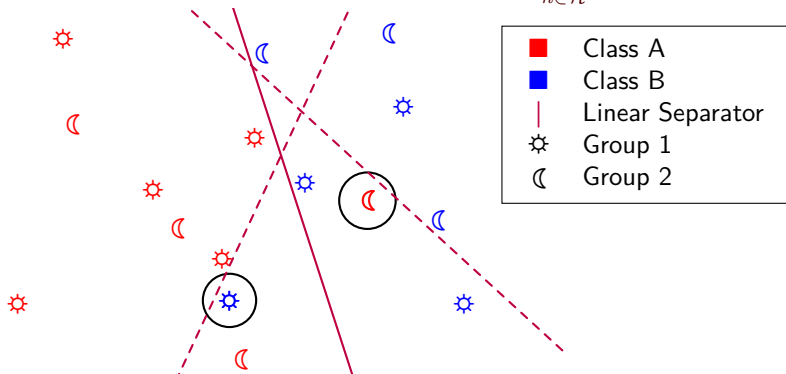
Learning Fair Linear Classifiers

- We can handle each group individually:

$$\hat{R}(h; \mathbf{x}_i, \mathbf{y}_i) \doteq \frac{1}{m} \sum_{j=1}^m \ell(h(\mathbf{x}_{i,j}), \mathbf{y}_{i,j}); \quad \forall i: \hat{h}_i \doteq \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}(h; \mathbf{x}_i, \mathbf{y}_i)$$

- What is the best classifier overall?

- Empirical welfare minimization $\hat{h} \doteq \operatorname{argmin}_{h \in \mathcal{H}} \Lambda(\hat{R}(h; \mathbf{x}_1, \mathbf{y}_1), \hat{R}(h; \mathbf{x}_2, \mathbf{y}_2))$



Visit my poster session for more information!

An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

Cyrus Cousins

Monday, July 19 @ 2:00pm - 2:30pm EST

&

Monday, July 19 @ 9:00pm - 9:30pm EST

An Axiomatic Theory of Provably-Fair Welfare-Centric Machine Learning

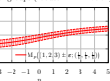
Cyrus Cousins of Brown University

Welfare, Malfare, and the Power Mean

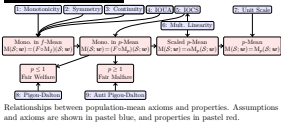
- The power-mean for $p \in \mathbb{R} \setminus \{0\}$ summarizes g values S with weights w :

$$M_p(S; w) \doteq \sqrt[p]{\sum_{i \in I} w_i S_i^p}$$

- Fair welfare: $p \geq 1, p = \infty$ is *maximix* over groups (egalitarianism)
- Fair malfare: $p \geq 1, p = \infty$ is *minimiz* over groups (robust minimization)
- Power-means are
 - Axiomatically Justified
 - Interpretable
 - Statistically Stable



Axioms of Cardinal Welfare



Estimating Malfare Values

- Assuming only *monotonicity*:
Suppose $\forall \omega \in \Omega: \hat{S}(\omega) - \epsilon \leq S(\omega) \leq \hat{S}(\omega) + \epsilon$. Then
$$M_p(\hat{0} \vee (\hat{S} - \epsilon); w) \leq M_p(S; w) \leq M_p(\hat{S} + \epsilon; w)$$
 where $a \vee b$ denotes the (elementwise) maximum.
- Suppose range r . Then with probability at least $1 - \delta$ over choice of \mathcal{X} :

Empirical Malfare Minimization

Empirical risk and risk of hypothesis h given loss ℓ :

$$\hat{R}(h; \ell, \mathcal{Z}) \doteq \frac{1}{|\mathcal{Z}|} \sum_{(x,y) \in \mathcal{Z}} [\ell(y, h(x))] \quad \& \quad R(h; \ell, \mathcal{D}) \doteq \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(y, h(x))]$$

We define *empirical malfare minimization* (EMM), given $\mathcal{M}(\cdot; \mathcal{D}_1, \mathcal{D}_2)$ and $\mathbf{z}_{1,2}$, with proxy and ideal models

$$\hat{h} \doteq \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{M}(h \mapsto \hat{R}(h; \ell, \mathbf{z}_1); \mathbf{w}) \quad \& \quad h^* \doteq \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{M}(h \mapsto R(h; \ell, \mathcal{D}_1); \mathbf{w})$$

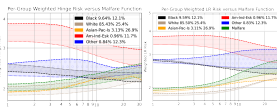
Under what conditions is \hat{h} a good proxy for h^* ?

Theorem 1 (Generalization Guarantees for Malfare Estimation). Suppose fair power-mean malfare $M_p(\cdot; \cdot)$ (i.e., $p \geq 1$), probability vector $\mathbf{w} \in \mathbb{R}_+^I$, loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$, samples $\mathbf{z}_1 \sim \mathcal{D}_1^N$, and hypothesis class $\mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}$. Then with probability at least $1 - \delta$ over choice of \mathcal{Z} ,

$$\sup_{h \in \mathcal{H}} |M_p(h \mapsto R(h; \ell, \mathcal{D}_1); \mathbf{w}) - M_p(h \mapsto R(h; \ell, \mathbf{z}_1); \mathbf{w})| \leq M_p(h \mapsto 2\mathbb{E}_{\mathcal{D}_1}(\ell \circ H, \mathbf{z}_1) + 3\sqrt{\frac{\ln \frac{1}{\delta}}{N}})$$

Experiments

- Training *linear models* on *adult* (census data) dataset
 - Support vector machine (hinge loss)
 - Logistic regression (cross entropy loss)
 - Losses weighted by group-conditional label frequencies
- Predict whether income is \leq or $>$ 50,000\$ per annum
- Minimize malfare over 5 ethnic groups



Fair PAC Learning

Definition 2 (Fair-PAC (FPAC) Learnability). Suppose hypothesis class sequence $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots \subseteq \mathcal{X} \rightarrow \mathcal{Y}$, and loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{0+}$.

We say \mathcal{H} is *fair PAC-learnable* w.r.t. loss function ℓ if \exists a (randomized) algorithm \mathcal{A} , such that for all:

- sequence indices d ,
- malfare \mathcal{M} satisfying axioms 1-7 & 9,
- additive approx. errors $\epsilon >$, and
- failure probabilities $\delta \in (0, 1)$;

\mathcal{A} can identify a hypothesis $h \in \mathcal{H}_d$, i.e., $h \leftarrow \mathcal{A}(\mathcal{D}_{1,2}, \mathbf{w}, \mathcal{M}, \epsilon, \delta, d)$, such that

- there exists some *sample complexity* function $m(\epsilon, \delta, d, g): (\mathbb{R}_+ \times (0, 1) \times \mathbb{N} \times \mathbb{N}) \rightarrow \mathbb{N}$ s.t. $\mathcal{A}(\mathcal{D}_{1,2}, \mathbf{w}, \mathcal{M}, \epsilon, \delta, d)$ consumes no more than $m(\epsilon, \delta, d, g)$ samples (finite sample complexity); and
- with probability at least $1 - \delta$ (over randomness of \mathcal{A}), h obeys

$$\mathcal{M}(h \mapsto R(h; \ell, \mathcal{D}_1); \mathbf{w}) \leq \inf_{h' \in \mathcal{H}} \mathcal{M}(h' \mapsto R(h'; \ell, \mathcal{D}_1); \mathbf{w}) + \epsilon$$

The class of such fair-learning problems is FPAC.

Finally, if for all d , the space of \mathcal{D} is restricted such that $\exists h \in \mathcal{H}_d$ s.t. $\max_{h \in \mathcal{H}_d} R(h; \ell, \mathcal{D}_1) = 0$, then (\mathcal{H}, ℓ) is *realizable-FPAC-learnable*.

Fair PAC Learnability

