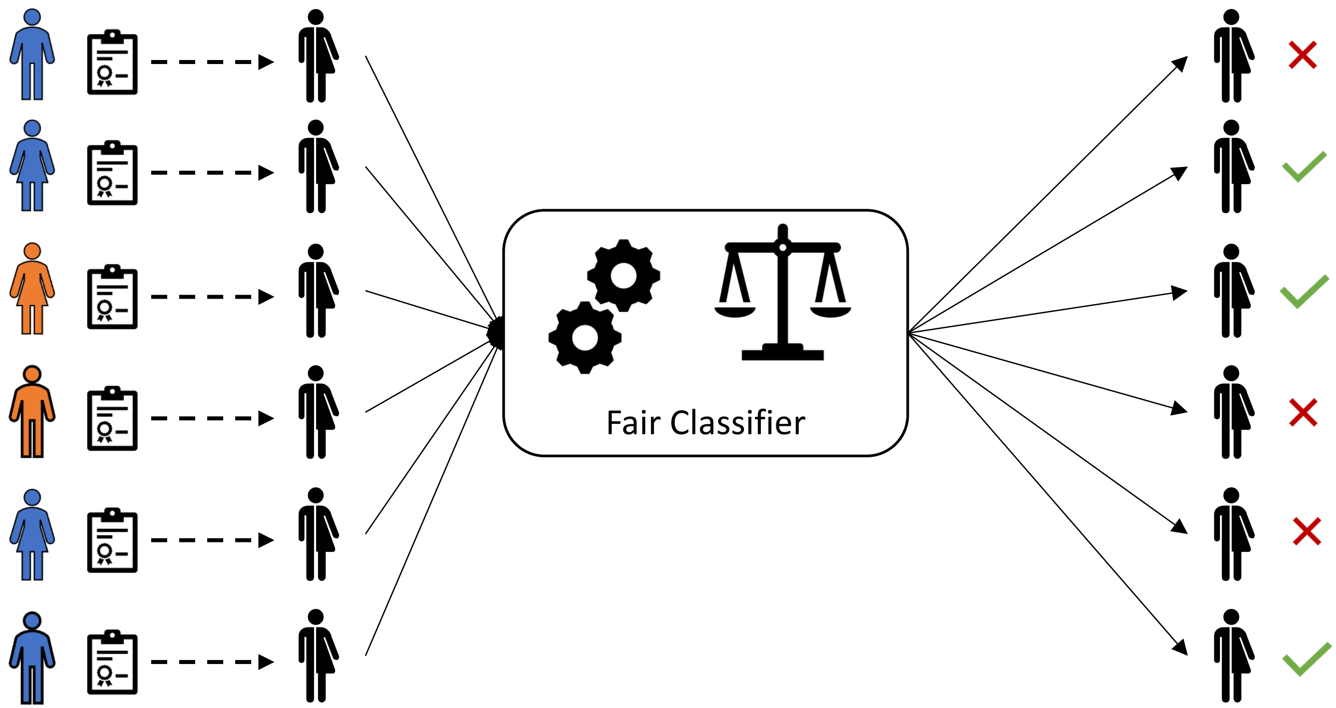
Fair Classification with Noisy Protected Attributes: A Framework with Provable Guarantees

L. Elisa Celis ▪ Lingxiao Huang ▪ Vijay Keswani ▪ Nisheeth K. Vishnoi



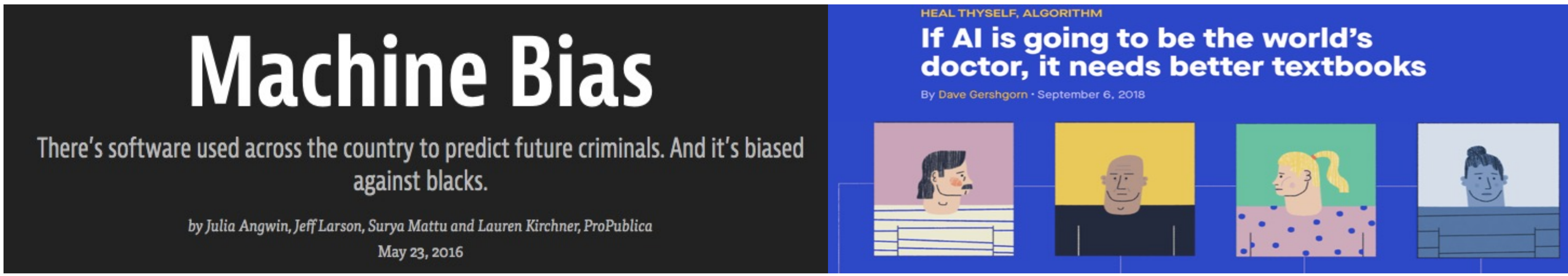
Yale

controlling-bias.github.io

ICML
International Conference
On Machine Learning

Fair Classification

Recent research in fair classification has proposed multiple solutions to address the disparate impact of automated prediction (Bellamy et al. 2018, Zafar et al 2017, Hardt et al. 2016)



Perturbations in protected attributes

- Data collection requires procedural and political decisions and can contain errors with respect to race, gender, or identity information (Saez et al., 2013, Nobles, 2000)
- Information about protected attributes may be missing entirely/prohibited from direct use (Data et al., 2004) and automated prediction can be biased (Muthukumar et al., 2018)

Existing fair classification methods do not always work with perturbed protected attributes

Can we do fair classification when protected attributes are perturbed?

Model and Main Result

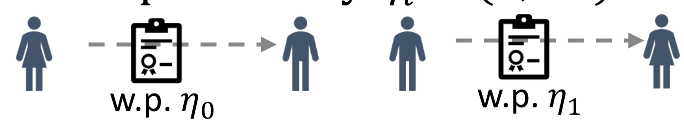
Target fair classification program

- N samples: $S = \{(x_j, z_j, y_j)\}_j \in$ (features)×(binary protected attribute)×(label)
- Loss function $L: \mathcal{F} \times S \to \mathbb{R}$
- Statistical rate $\Omega(f, S) = \frac{\min_{i \in \{0,1\}} \Pr[f=1|Z=i]}{\max_{i \in \{0,1\}} \Pr[f=1|Z=i]}$ and desired fairness guarantee $\tau \in [0,1]$

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{j \in [N]} L(f, s_j) \text{ such that } \Omega(f, S) \geq \tau$$

Perturbation model

- For protected attribute z $\in \{0,1\}$, $z = i \to \hat{z} = 1 - i$ with probability $\eta_i \in (0, 0.5)$
- Observed dataset: $\hat{S} = \{(x_j, \hat{z}_j, y_j)\}_j$



Such perturbations arise in important applications like randomized response models

λ-Assumption

$\exists \lambda \in (0, 0.5)$, s.t. $\max_i \Pr[f^\star = 1, Z = i] \geq \lambda$, where $f^\star$ is an optimal fair classifier

Main Result - Given an observed dataset $\hat{S}$, desired fairness guarantee $\tau \in [0,1]$, $\eta_0, \eta_1 \in (0, 0.5)$ and $\delta \geq 0$, suppose the λ-Assumption is satisfied for $\lambda \in (0, 0.5)$. We provide an optimization framework that outputs a classifier f s.t., with high probability,

- *(Accuracy guarantee)* empirical risk of f is less than or equal to empirical risk of $f^\star$
- *(Fairness guarantee)* statistical rate of f is atleast $\tau - 3\delta$

Our Framework

How do we estimate $\Pr[f = 1 \mid Z = i]$ using $\hat{Z}$ when η_0, η_1 are known?

First estimate $\Pr[\, f = 1 \mid Z = i \,]$ using $\hat{Z}$

$$\Gamma_i(f) := \frac{(1-\eta_{1-i})\Pr[f = 1, \hat{Z} = i] - \eta_{1-i}\Pr[f = 1, \hat{Z} = 1-i]}{(1-\eta_{1-i})\Pr[\hat{Z} = i] - \eta_{1-i}\Pr[\hat{Z} = 1-i]}$$

Estimated statistical rate $= \frac{\min_{i\in\{0,1\}} \Gamma_i(f)}{\max_{i\in\{0,1\}} \Gamma_i(f)}$

Need to guarantee (w.h.p.) we learn fair & accurate classifier even with large noise

Incorporate λ-Assumption as a constraint to obtain a classifier that is close to f^*

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{j\in[N]} L(f, s_j) \text{ such that}$$
$$\min_{i\in\{0,1\}} \Gamma_i(f) \geq (\tau - \delta)\cdot \max_{i\in\{0,1\}} \Gamma_i(f),$$
$$(1-\eta_{1-i})\Pr\left[f = 1, \hat{Z} = i\right] - \eta_{1-i}\Pr\left[f = 1, \hat{Z} = 1-i\right] \geq \lambda M - \delta, \text{for all } i \in \{0,1\}$$

$$\delta \geq 0 \text{ – relaxation parameter and } \mathrm{M} = (1-\eta_0-\eta_1)$$

λ can be estimated in applications, given estimates of $\Pr[Z = i]$ & $\Pr[Y = 1|Z = i]$

Framework and theoretical results can be extended to multiple protected attributes and other linear fairness metrics (e.g., equalized odds) and linear-fractional fairness metrics (e.g., false discovery rate, predictive parity) **- see paper for more details**

Empirical results

UCI Adult Income Dataset — Dataset Size ~40k, Protected attribute – sex, race (binary)

Noise model: $\eta_0 = 0.3$, $\eta_1 = 0.1$
(Minority group is more likely to contain errors in real-world applications - Nobles, 2000)

Metrics: Accuracy and statistical rate (with respect to true protected attributes – "SR")

Protected Attribute - ***sex***

	Acc	SR
Unconstrained	**.80** (0)	.31 (.01)
DLR-SR τ=.9	.76 (.01)	.85 (.15)
Lamy et al. '19	.78 (.02)	.69 (.09)
Awasthi et al.'20	.77 (0)	.66 (.05)
Wang et al. '20	.70 (.05)	.73 (.12)

Protected Attribute - ***race***

	Acc	SR
Unconstrained	**.80** (0)	.68 (.02)
DLR-SR τ=.9	.76 (.01)	.88 (.18)
Lamy et al. '19	.80 (0)	.70 (.01)
Awasthi et al.'20	.80 (0)	.72 (.02)
Wang et al. '20	.76 (.01)	.84 (.05)

Observations: (a) fairness close to τ, (b) better fairness-accuracy tradeoff than baselines

Paper contains additional experiments using other datasets and fairness metrics

Conclusion

- We propose a fair classification framework for the setting where protected attributes are perturbed according to a flipping noise model
- Output classifier guaranteed to be accurate and fair with high probability

Limitations and future work

- Extension to non-independent noise models
- Consider joint noise-models over both protected attributes and labels