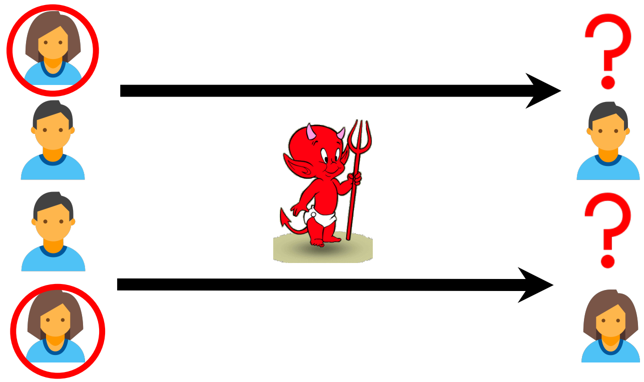
Fair Classification with Adversarial Perturbations

Adversary can pick any η fraction of samples



Adversary can replace selected samples arbitrarily

L. Elisa Celis

Anay Mehrotra

Nisheeth K. Vishnoi

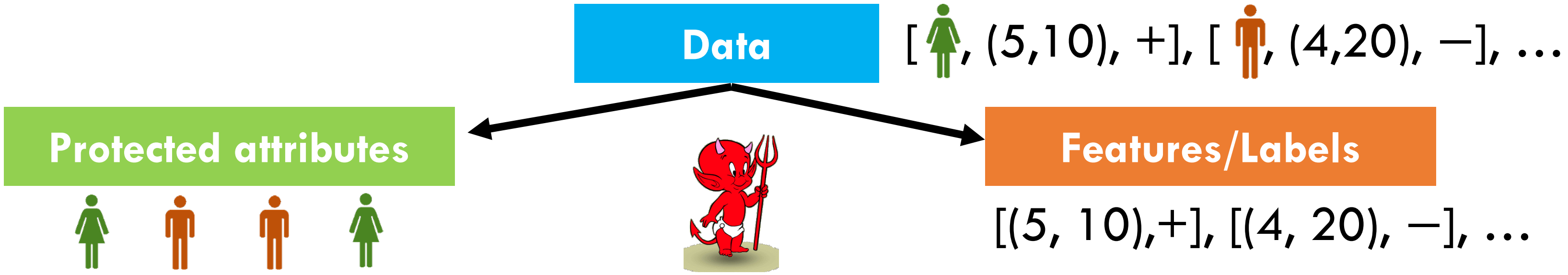
Yale

NEURAL INFORMATION PROCESSING SYSTEMS

Inaccuracies in data hamper existing fair classifiers

State-of-the-art approaches to mitigate the disparate impact of automated prediction find classifiers that are "fair" with respect to protected groups (e.g., defined by race and gender) [HPS16, ZVRG17, BDHH+18]

However, data may not be accurate…



- Data can be **strategically misreported** [Luh19] and have **missing protected attributes.** E.g., racial/ethnic information in health care [Eli04] and in data scraped from the internet [DDSL+09]
- Missing values can be **imputed.** But imputation is bound to **introduce errors,** which can be **correlated across samples** [MPRS+18] and **susceptible imperceptible changes** [GSS15]

Existing fair classification methods do not work when data has correlated/arbitrary perturbations

Is fair classification possible when a fraction of the data are arbitrarily perturbed?

Model of fair classification

- ***Data:*** N samples $S = \{(x_i, y_i, z_i)\}_{i=1,\dots,N} \in$ (features) × (labels) × (protected attributes)
- ***Loss function:*** $\mathrm{Err}(f, S) \in [0,1]$ measures fraction of incorrect predictions by f on S
- ***Fairness metric:*** E.g., statistical rate $\mathrm{SR}(f, S) = \frac{\min_\ell \Pr_S[f=1|Z=\ell]}{\max_\ell \Pr_S[f=1|Z=\ell]}$
- ***Desired fairness threshold:*** $\tau \in [0,1]$

Ideal fair classification problem:

$$f^\star := \mathrm{argmin}_{f\in\mathcal{F}} \ \ \mathrm{Err}(f, S), \text{ such that } \Omega(f, S) \geq \tau \quad (1)$$

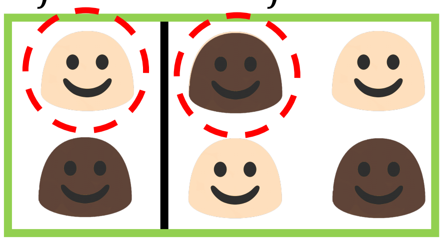
When S is known, (1) is a constrained optimization problem [HPS16, ZVRG17, BDHH+18]

Problem: We **observe $\widehat{S}$** that is a **perturbed version** of the **"true" data S**

Idea: Solve Program (1) by replacing S with the perturbed data $\widehat{\boldsymbol{S}}$

$\mathrm{Err}(f, S)$ & $\mathrm{SR}(f, S)$ can be different from $\mathrm{Err}(f, \hat{S})$ & $\mathrm{SR}(f, \hat{S})$ → Output can be inaccurate/unfair

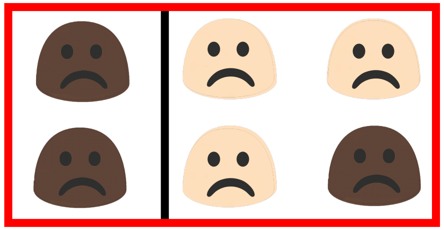
$f = 0$ $f = 1$



$\widehat{\boldsymbol{S}}$ (perturbed data)

$$\mathrm{SR}(f, \hat{S}) = \frac{2/3}{2/3} = 1$$

$f = 0$ $f = 1$



$\boldsymbol{S}$ (true data)

$$\mathrm{SR}(f, S) = \frac{1/3}{3/3} = \frac{1}{3}$$

Adversarial errors in data hinder prior approaches

Assumption: $\hat{S}$ has **IID perturbations** with **known distribution** $\mathcal{P}$ [LMZV19][AKM20][WLL21][CHKV21]

For all $i \in [N]$, $(\hat{x}_i, \hat{y}_i, \hat{z}_i) = (x_i, y_i, z_i) + \pi_i$, where $\pi_i \overset{\text{iid}}{\sim} \mathcal{P}$

Approach: Given $\mathcal{P}$ derive unbiased estimates $\mathrm{SR} \longrightarrow \widehat{\mathrm{SR}}$ and $\mathrm{Err} \longrightarrow \widehat{\mathrm{Err}}$ such that

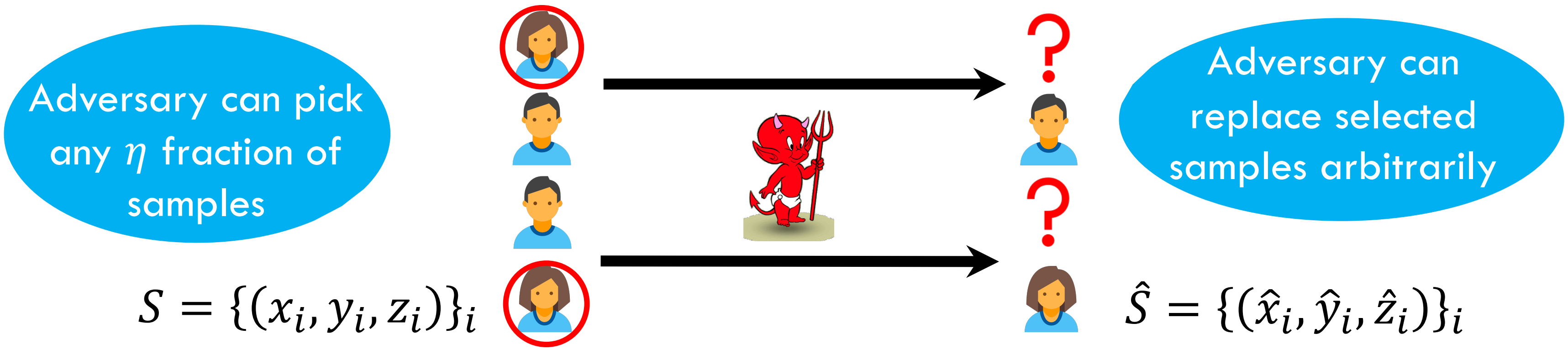
$$\mathbb{E}\left[\widehat{\mathrm{SR}}\left(f, \hat{S}\right)\right] = \Omega(f, S) \pm O(N^{-1}) \quad \text{and} \quad \mathbb{E}\left[\widehat{\mathrm{Err}}\left(f, \hat{S}\right)\right] = \mathrm{Err}(f, S) \pm O(N^{-1})$$

Solve: $\min_{f \in \mathcal{F}} \widehat{\mathrm{Err}}(f, \hat{S})$, such that $\widehat{\mathrm{SR}}\left(f, \hat{S}\right) \geq \tau$ (2)

Other prior work consider similar settings:

- $\mathcal{P}$ is not known but can be **"estimated" using auxiliary data** [WGNC+20]
- $\hat{S}$ has **arbitrary perturbations** on **samples selected uniformly** without replacement [KL21]

Problem: Rely on perturbations being independent and $\mathcal{P}$ being known or can be estimated



Perturbation model: Given $\eta \in [0,1]$, adversary chooses any ηN samples and corrupts them **arbitrarily**

Problem: Given $\eta > 0$, N samples S, $\widehat{\mathrm{SR}}$, and $\widehat{\mathrm{Err}}$, the adversary can perturb ηN samples to generate $\hat{S}$ such that $\mathrm{Err}(f, S)$ and $\mathrm{SR}(f, S)$ are "far" from $\mathbb{E}\left[\widehat{\mathrm{Err}}\left(f, \hat{S}\right)\right]$ and $\mathbb{E}\left[\widehat{\mathrm{SR}}\left(f, \hat{S}\right)\right]$

Theoretical results

Pathological case: If $\Pr[Z=\ell] \leq \eta$ (for some $\ell \in \{0,1\}$), the adversary can **perturb all samples** in the ℓ-th protected group – $\widehat{S}$ **gives "no information" about samples** ℓ-th group

- Information-theoretically impossible to find $f^\circ \in \mathcal{F}$, s.t., $\mathrm{Err}(f^\circ, S) < 1/2$ and $\mathrm{SR}\,(f^\circ, S) > 1/2$

λ-assumption: There is a known constant $\lambda > 0$ such that $\min_\ell \Pr_S[f^\star = 1, Z = \ell] \geq \lambda$

where $f^\star \coloneqq \mathrm{argmin}_{f\in\mathcal{F}}\ \mathrm{Err}(f,S)$, such that $\Omega(f,S) \geq \tau$

In particular, the **λ-assumption ensures** that for all $\ell \in \{0,1\}$, $\Pr[Z=\ell] \geq \lambda$.

Main result: There is an optimization program parameterized by perturbation rate $\eta \in (0,1)$, desired fairness threshold $\tau \in [0,1]$, hypothesis class $\mathcal{F}$, and perturbed data $\hat{S}$ with N samples, such that the optimal solution $f^\circ \in \mathcal{F}$ satisfies:

1. *Accuracy guarantee:* $\mathrm{Err}(f^\circ, S) \leq \mathrm{Err}(f^\star, S) + 2\eta$,
2. *Fairness guarantee:* $\mathrm{SR}(f^\circ, S) \geq \tau - O(\eta)$.

Lower bound: Given perturbation rate $\eta \in (0,1)$, hypothesis class $\mathcal{F}$, perturbed data $\hat{S}$, and fairness threshold $\tau \in [0,1]$, it is **information-theoretically impossible** to find $f^\circ \in \mathcal{F}$ such that:

1. *Accuracy guarantee:* $\mathrm{Err}(f^\circ, S) < \mathrm{Err}(f^\star, S) + \boldsymbol{\eta}$, and
2. *Fairness guarantee:* $\mathrm{SR}(f^\circ, S) \geq \tau - \boldsymbol{o(\eta)}$.

Related work: PAC learning + adversary [BEK02]. Output f s.t.: $\mathrm{Err}(f,S) \leq \min_f \mathrm{Err}(f,S) + 2\eta$,
But no fairness guarantee. We "match" their accuracy guarantee AND also give SR guarantee.

Adversary's effect on accuracy and stat. rate

Challenge: Cannot derive unbiased estimates of predictive error or statistical rate on S because the **adversary can adapt** perturbations to the estimates

Bound the "effect of the adversary:" Given a classifier $f \in \mathcal{F}$ and a perturbation rate $\eta > 0$: Bound $\left|\mathrm{Err}(f, S) - \mathrm{Err}(f, \hat{S})\right|$ and $\left|\mathrm{SR}(f, S) - \mathrm{SR}(f, \hat{S})\right|$

Adversary's effect on accuracy and stat. rate

1) Effect of adversary on accuracy: Let $\ell(x,z,y) := \mathbb{I}[f(x,z) \neq y]$

ηN samples perturbed

$$\mathrm{Err}(f,\hat{S}) = \frac{1}{N}\sum_i \ell(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{z}}_i, \hat{\boldsymbol{y}}_i) = \frac{1}{N}\sum_{i=1}^{N(1-\eta)} \ell(x_i,z_i,y_i) + \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N\eta}} \ell(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{z}}_i, \hat{\boldsymbol{y}}_i)$$
$$= \mathrm{Err}(f,S) \pm \boldsymbol{\eta}$$

Accuracy on S and $\hat{S}$ are close to each other if η is small

2) Effect of adversary on statistical rate:

$$\mathrm{SR}(f,\hat{S}) := \frac{\min_\ell \Pr_{\hat{S}}[f=1 | \hat{\boldsymbol{Z}}=\ell]}{\max_\ell \Pr_{\hat{S}}[f=1 | \hat{\boldsymbol{Z}}=\ell]} = \frac{\Pr[f=1 \wedge \hat{\boldsymbol{Z}}=\boldsymbol{\ell_1}]\cdot\Pr[\hat{\boldsymbol{Z}}=\boldsymbol{\ell_2}]}{\Pr[f=1 \wedge \hat{\boldsymbol{Z}}=\boldsymbol{\ell_2}]\cdot\Pr[\hat{\boldsymbol{Z}}=\boldsymbol{\ell_1}]}$$

(for some $\ell_1, \ell_2 \in \{0,1\}$)

$$= \frac{(\Pr[f=1 \wedge Z=\ell_1] \pm \boldsymbol{\eta})\cdot(\Pr[Z=\ell_2] \pm \boldsymbol{\eta})}{(\Pr[f=1 \wedge Z=\ell_2] \pm \boldsymbol{\eta})\cdot(\Pr[Z=\ell_1] \pm \boldsymbol{\eta})}$$

η-fraction of samples perturbed

$$= \frac{\Pr[f=1 \wedge Z=\ell_1]\cdot\Pr[Z=\ell_2] \pm \boldsymbol{2\eta}}{\Pr[f=1 \wedge Z=\ell_2]\cdot\Pr[Z=\ell_1] \pm \boldsymbol{2\eta}}$$
$$\in \mathrm{SR}(f,S) \pm \mathcal{E}(f,\eta,S)$$

Error $\mathcal{E}(f,\eta,S)$ can be large if denominator is small compared to η

Statistical rate on $\hat{S}$ and S can be very different!

Adversary's effect on accuracy and stat. rate

1) Effect of adversary on accuracy: Let $\ell(x,z,y) := \mathbb{I}[f(x,z) \neq y]$

ηN samples perturbed

$$\mathrm{Err}(f,\hat{S}) = \frac{1}{N}\sum_i \ell(\hat{x}_i,\hat{z}_i,\hat{y}_i) = \frac{1}{N}\sum_{i=1}^{N(1-\eta)} \ell(x_i,z_i,y_i) + \frac{1}{N}\sum_{i=1}^{N\eta} \ell(\hat{x}_i,\hat{z}_i,\hat{y}_i)$$

$$= \mathrm{Err}(f,S) \pm \eta$$

Accuracy on S and $\hat{S}$ are close to each other if η is small

2) Effect of adversary on statistical rate:

$$\mathrm{SR}(f,\hat{S}) := \frac{\min_\ell \Pr_{\hat{S}}[f=1|\hat{Z}=\ell]}{\max_\ell \Pr_{\hat{S}}[f=1|\hat{Z}=\ell]} = \frac{\Pr[f=1 \wedge \hat{Z}=\ell_1]\cdot\Pr[\hat{Z}=\ell_2]}{\Pr[f=1 \wedge \hat{Z}=\ell_2]\cdot\Pr[\hat{Z}=\ell_1]}$$

(for some $\ell_1, \ell_2 \in [p]$)

$$= \frac{(\Pr[f=1 \wedge Z=\ell_1]\pm\eta)\cdot(\Pr[Z=\ell_2]\pm\eta)}{(\Pr[f=1 \wedge Z=\ell_2]\pm\eta)\cdot(\Pr[Z=\ell_1]\pm\eta)}$$

η-fraction of samples perturbed

$$= \frac{\Pr[f=1 \wedge Z=\ell_1]\cdot\Pr[Z=\ell_2]\pm 2\eta}{\Pr[f=1 \wedge Z=\ell_2]\cdot\Pr[Z=\ell_1]\pm 2\eta}$$

$$\in \mathrm{SR}(f,S) \pm \mathcal{E}(f,\eta,S)$$

Error $\mathcal{E}(f,\eta,S)$ can be large if denominator is small compared to η

Statistical rate on $\hat{S}$ and S can be very different!

Definition (r-stability). A classifier is r-stable for S and $\hat{S}$ if $r \leq \mathrm{SR}(f,S)/\mathrm{SR}(f,\hat{S}) \leq 1/r$

Consequence of r-stability: If f is r-stable, then $\mathrm{SR}(f,\hat{S}) \geq \tau \Rightarrow \mathrm{SR}(f,S) \geq \tau \cdot r$

Direct approach: Compute $\mathrm{SR}(f,S)$ and $\mathrm{SR}(f,\hat{S})$ to check if f is r-stable

The direct approach is **not possible** because **S is not observed!**

Lemma. Given $\eta \in (0,1)$, $r \in (0,1)$, $f \in \mathcal{F}$, and S and $\hat{S}$ (which has $\eta \cdot N$ perturbed samples), if for all $\ell \in [p]$, $\Pr_{\hat{S}}[f = 1 \wedge \hat{Z} = \ell] \geq 2\eta(1-\sqrt{r})^{-1} - \eta$, then f is r-stable.

Our framework

Parameter: $r := 1 - O(\eta)$

$$\begin{aligned} &\min_{f\in\mathcal{F}} && \mathrm{Err}(f,\hat{S}) && (1)\\ &\text{s.t.,} && \mathrm{SR}(f,\hat{S}) \geq \tau\cdot r && (2)\\ &\forall \ell\in[p] && \Pr_{\hat{S}}\left[f=1 \wedge \hat{Z}=\ell\right] \geq \frac{2\eta}{1-\sqrt{r}} - \eta && (3)\end{aligned}$$

Intuition: Find the classifier with **min. predictive error on $\hat{S}$** that has **SR $\geq \tau r$ on $\hat{S}$** and is **r-stable**

1) **Fairness guarantee:** Any feasible solution has statistical rate $\geq \tau \cdot r^2$ due to Constraints (2)&(3)

- From Constraint (2), $\mathrm{SR}(f,\hat{S}) \geq \tau\cdot r$
- From Constraint (3), any solution is r-stable for $r = 1 - O(\eta)$
- Combining these, $\mathrm{SR}(f,S) \geq r\cdot \mathrm{SR}(f,\hat{S})$ (definition of r-stability)

$$\geq r\cdot\tau\cdot r$$
$$\geq \tau\cdot r^2$$

2) **Accuracy guarantee:** Follows because, under the λ-assumption, $f^\star$ is feasible for Program (1)

Let f° be the optimal solution of Program (1)

$$\begin{aligned} \mathrm{Err}(f^\circ,S) &\leq \mathrm{Err}(f^\circ,\hat{S}) + \eta && (\forall f,\ \mathrm{Err}(f,S) = \mathrm{Err}(f,\hat{S}) \pm \eta)\\ &\leq \mathrm{Err}(f^\star,\hat{S}) + \eta && (f^\circ \text{ is optimal for Program (1)})\\ &\leq \mathrm{Err}(f^\star,S) + 2\eta && (\forall f,\ \mathrm{Err}(f,S) = \mathrm{Err}(f,\hat{S}) \pm \eta)\end{aligned}$$

The paper extends the framework to other fairness metrics Ω and multiple protected attributes

Empirical results on real-world data

COMPAS data: Size ≈6000, protected attribute: gender (encoded as binary)

Two adversaries (η=3.5%): A$_{TN}$ and **A$_{FN}$** construct $\hat{S}$ to heuristically increase $\mathrm{SR}(f^\star, \hat{S})$

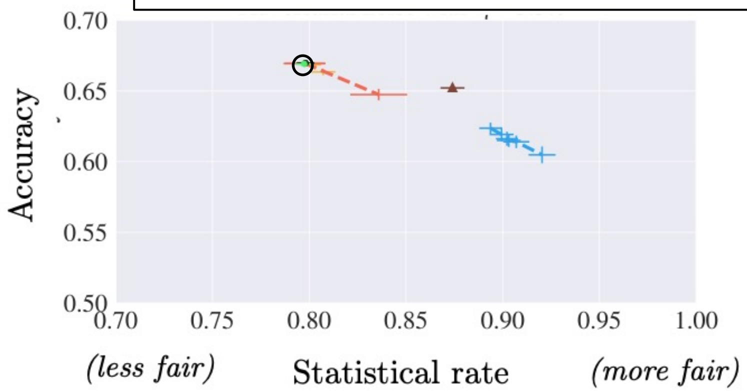
"Select ηN samples" furthest from $f^\star$'s decision boundary with $Z = 1$. **for each** sample, **set** $\hat{Z} = 2$

Idea: *Samples far from decision boundary of* $f^\star$ *are "confident." Perturbing their protected attributes also increases the statistical rate of other classifiers along with* $f^\star$

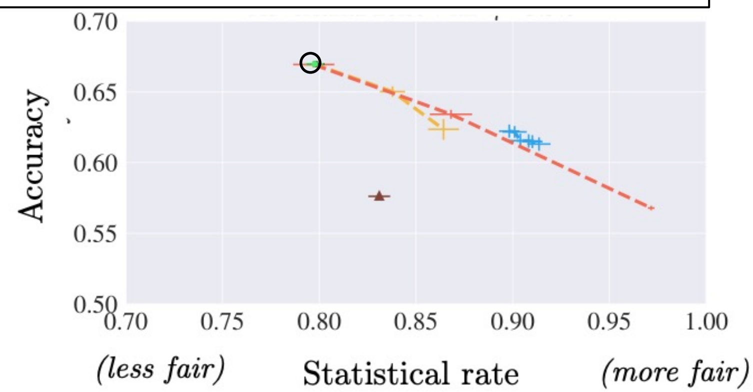
These are not intended to be worst-case. But our guarantees hold for worst-case adversaries

Metrics: Accuracy and statistical rate (w.r.t. the unperturbed dataset S); τ varies from 0 to 1

Unconstrained **(This work)** **[LMZV19]** **[AKM20]** **[CHKV21]** **[KL21]**



A$_{TN}$ (η=3.5%)



A$_{FN}$ (η=3.5%)

Observations: • Better stat. rate than uncons. classifier (12%), with minimal loss in accuracy (7%)

- Similar (or better) fairness-accuracy trade-off than baselines

The paper also contains empirical results on UCI Adult data, other fairness metrics, and adversaries

Key takeaways

- Most existing frameworks for fair decision making **assume data is accurate,** or make **independence assumptions** on the errors
- In many applications, **data has perturbations** that are across samples, and may even be **correlated strategically chosen**
- Such errors **hurt both fairness and accuracy guarantees** of existing frameworks

Conclusion

- We study fair classification with **adversarial perturbations** in the data
- Give a framework for fair classification whose optimal solution classifier has **provable guarantees on fairness and accuracy**
- Both the fairness and accuracy guarantees are tight up to constants

Limitations and future work

- Efficacy depends on appropriate choices of parameters: τ and η; e.g., either overly conservative or optimistic η can decrease accuracy and fairness

Must be considered as a part of a broader system for mitigating bias

- Is there a different model of perturbations that is also realistic and but allows for fairness and accuracy guarantees without additional assumptions?

https://controlling-bias.github.io/

Bibliography

[Hamming 1950] Richard W Hamming.
Error detecting and error correcting codes. The Bell system technical journal, 1950.

[Eli04] N.R. Council, D.B.S.S. Education, C.N. Statistics, P.D.C.R.E. Data, E. Perrin, & M.V. Ploeg.
Eliminating Health Disparities: Measurement and Data Needs.

[DDSL+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei.
Imagenet: A large-scale hierarchical image database. CVPR 2009

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy.
Explaining and harnessing adversarial examples. In ICLR, 2015.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro.
Equality of opportunity in supervised learning. NeurIPS 2016.

[ZVRG17] Muhammad B. Zafar, Isabel Valera, Manuel G. Rodriguez, and Krishna P. Gummadi.
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. WWW 2017.

[BDHH+18] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang.
AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[MPRS+18] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kings- bury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush Varshney
Understanding unequal gender classification accuracy from face images. 2018.

Bibliography

[MPRS+18] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kings- bury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney.
Understanding unequal gender classification accuracy from face images. arXiv preprint arXiv:1812.00099, 2018.

[Luh19] Elizabeth Luh.
Not so black and white: Uncovering racial bias from systematically misreported trooper reports. Available at SSRN 3357063, 2019.

[LMZV19] Alexandre Louis Lamy, Aditya Krishna Menon, Ziyuan Zhong, Nakul Verma.
Noise-tolerant fair classification. NeurIPS, 2019.

[AKM20] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern.
Equalized odds postprocessing under imperfect group information. AISTATS 2020.

[WGNC+20] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Michael I. Jordan.
Robust optimization for fairness with noisy protected groups. NeurIPS, 2020.

[CHKV21] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi.
Fair classification with noisy protected attributes. ICML 2021.

[WLL21] Jialu Wang, Yang Liu, and Caleb Levy.
Fair classification with group-dependent label noise. FAccT 2021.

[KL21] Nikola Konstantinov and Christoph H. Lampert.
Fairness-aware learning from corrupted data. CoRR, abs/2102.06004, 2021.