CERIAS

The Center for Education and Research in Information Assurance and Security

Unfair AI: It Isn't Just the Data

Chowdhury M R Haider, Chris Clifton

{chaider, clifton}@purdue.edu

Summary

Conventional wisdom: biased training data leads to biased models.

We show,

- Machine learning can be expected to introduce types of bias not found in the training data.
- Different group-wise optimal models with unequal accuracy leads to unfair optimal accuracy joint model w.r.t disparate impact.
- Likely occurrence due to systemic bias.
- *De-biasing* training data is insufficient to ensure machine learning fairness.

Example Scenario

- College admission prediction
- **Test score** best classifies **Caucasians,**
- **GPA** best classifies the **non-Caucasians,** but less accurate than test-score on Caucasian

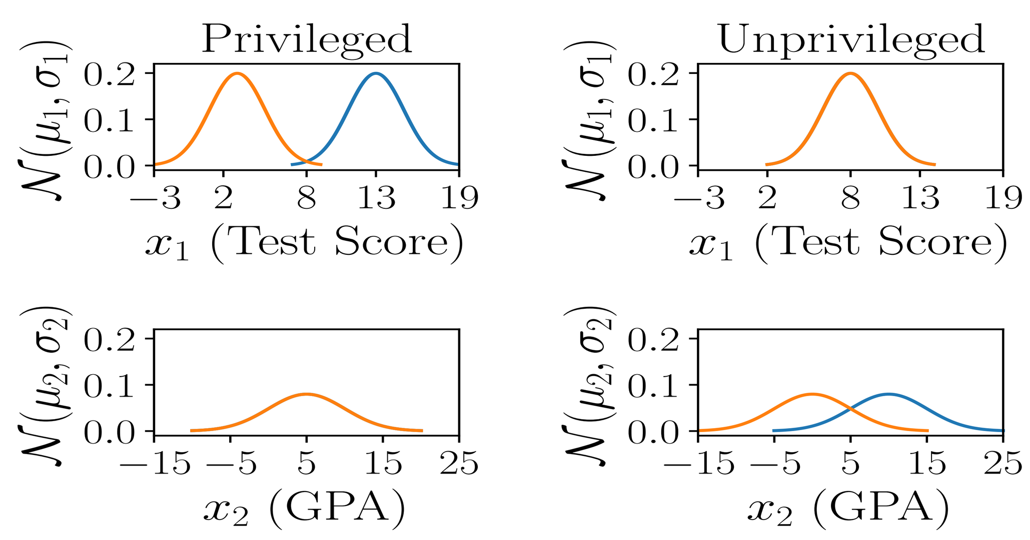

Figure 1: Feature Distributions were the unprivileged s less separable

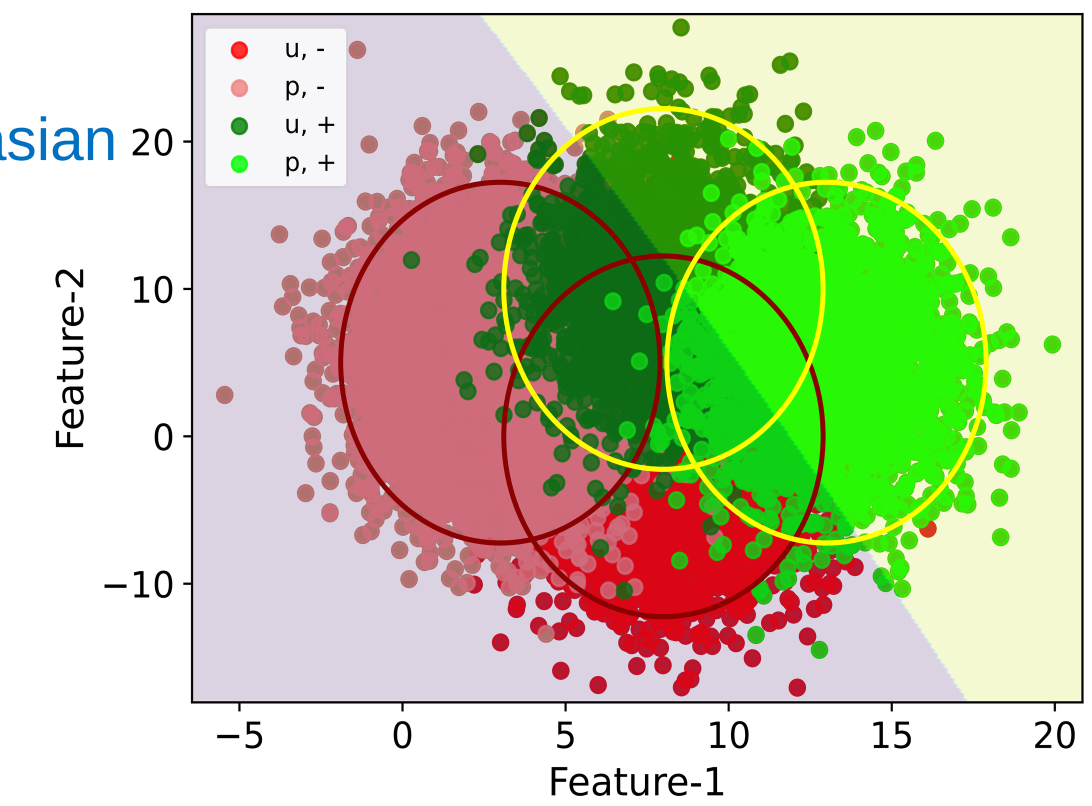

Figure 2: Decision Boundary and Classifications

Experiments

Datasets:

- Synthetic Fair Balanced Dataset (SFBD)
- COMPAS dataset (De-biased, balanced)

Algorithms:

- Naïve Bayes,
- Prejudice Remover [1],
- Reduction Based Model [2]

Table 1: Classifier Performance on SFBD

	α	ACC_p	ACC_u	SR_p	SR_u	FPR_p	FPR_u
NBC	0.25	99.6	87.4	21.1	15.4	00.3	07.2
	0.50	99.4	84.4	49.9	49.4	03.0	25.9
	0.75	99.5	87.1	78.3	85.6	13.6	61.3
PR	0.25	99.5	87.5	25.0	21.4	00.3	06.1
	0.50	99.4	84.4	50.1	49.1	00.6	14.7
	0.75	99.5	87.1	75.6	75.8	02.8	30.2
RBC	0.25	99.5	87.5	22.1	21.8	01.1	10.1
	0.50	99.4	84.4	50.1	49.0	04.7	22.1
	0.75	99.5	87.1	76.8	78.4	11.9	43.2

Table 2: NBC Performance on COMPAS FBD

α	ACC_p	ACC_u	SR_p	SR_u	FPR_p	FPR_u
0.25	73.3	74.0	21.1	10.9	15.5	06.5
0.50	61.8	61.6	62.4	41.6	52.1	30.3
0.75	75.9	72.2	89.0	81.9	76.5	68.7

Result on synthetic and COMAPAS data support our claim. COMPAS dataset contains unfair labels which leads to deviation for $\alpha = 0.75$.

Problem Formulation

- Dataset, $\mathcal{D} = \{\boldsymbol{x}^{(k)},\ y^{(k)}, s^{(k)}\}_{k=1}^{N}$
- Feature vector, $\boldsymbol{x}^{(k)} = \{x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\}$
- Class labels, $y_i \in \{+,-\}$
- Sensitive attribute, $s_i \in \{p,u\}$
- Base rate, $\alpha = \mathbb{P}(y^+ \mid s) = \mathbb{P}(y^+)$
- Ratio of groups, $\beta = \mathbb{P}(p)/\mathbb{P}(u)$
- $x_i^{sy} \sim \mathcal{N}(\mu_i^{sy}, \sigma_i^{sy})$ with $\sigma_i^{sy} = \sigma_i^{s} \perp\!\!\!\perp s, y$
- Number of redlining features is $2r$
- The rest $n - 2r$ are independent

Disparate Impact of dataset $\mathcal{D}$,

$$DI(\mathcal{D}) = \frac{\mathbb{P}(y^+ \mid u)}{\mathbb{P}(y^+ \mid p)}$$

Disparate Impact of model θ ($\hat{y}$ prediction),

$$DI(\theta) = \frac{\mathbb{P}(\hat{y}^+ \mid u)}{\mathbb{P}(\hat{y}^+ \mid p)}$$

Assumptions

- $r = 1, n = 2, \beta = 1$
- $\mu_1^{p+} - \mu_1^{p-} = \mu_1^{u+} - \mu_1^{u-} = \delta$
- $\mu_1^{u+} - \mu_1^{u-} = \mu_2^{p+} - \mu_2^{p-} = 0$
- p is more separable than u, $\Rightarrow \sigma_1 < \sigma_2$

Result Summary

In this scenario,

We show that, if $DI(\mathcal{D}) = 1$, the Bayesian joint optimal model θ satisfies,

$$\mathbb{P}(\hat{y}^+|y^+,p) > \mathbb{P}(\hat{y}^+|y^+,u)$$
$$\mathbb{P}(\hat{y}^+|y^-,p) < \mathbb{P}(\hat{y}^+|y^-,u)$$

- ✓ **True Positive Rate is higher for p**
- ✓ **False Positive Rate is higher for u**

In other words,

- ✓ **Favorable for the privileged**
- ✓ **Unfavorable for the unprivileged**

Furthermore, with $\alpha < 0.5$,

$$\mathbb{P}(\hat{y}^+|p) > \mathbb{P}(\hat{y}^+|u)$$

Similarly, $\alpha > 0.5$ shows,

$$\mathbb{P}(\hat{y}^+|p) < \mathbb{P}(\hat{y}^+|u)$$

- The joint optimal model is expected to induce disparate impact even when the training data with imbalanced base rates is free from such bias.

Discussion

- Systemic bias resulting from lack of diversity in feature design
- Dataset repair doesn't always work
- Joint optimization of fairness and accuracy is a step towards right direction
- Effect of other systemic bias, i.e., disparity in noise rate, missing value rate, representation, etc., needs further study

Reference

1. Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012, September). Fairness-aware classifier with prejudice remover regularizer. In ECML PKDD (pp. 35-50). Springer, Berlin, Heidelberg.
2. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In *ICML* (pp. 60-69). PMLR.R

PURDUE UNIVERSITY Discovery Park

CERIAS