華東師範大學
EAST CHINA NORMAL UNIVERSITY

Penalizing Unfairness in Binary Classification

沈楚云

Multi-Agent Artificial Intelligence Laboratory
East China Normal University

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Contents

Multi-Agent Artificial Intelligence Laboratory
East China Normal University

Contribution

- propose a new, easy-to-use, general-purpose technique for mitigating unfairness in classification settings.
- We validate the ability of our approach to achieve both fairness and high accuracy, implementing and testing it on multiple datasets pertaining to recidivism, credit, loan defaults, and law school admissions.

Models

- Goal
  - to achieve similar false positive rates in both populations, and similar false negative rates in both populations. (the rate of individuals who were classified using the COMPAS algorithm to be "high risk" but who did not actually re-offend was almost twice as high for black individuals as for whites; among those who were classified as "low risk" and did actually re-offend, the rate was significantly higher for whites than it was for blacks)
- Preliminaries
  - Represent each data point as a pair (x, y) ∈ $R^d \times \{0, 1\}$
  - the first feature x_1 (which we assume to be binary) represents a protected attribute (e.g., subgroup membership, black vs. white) and we will also write it as A ∈ {0, 1}; and y ∈ {0, 1} represents the true label (e.g., "re-offended" or "did not re-offend").

Models

- A labeled data set S
  - $s_{ay} = \{x^{i} \in S: X_1^i = a, y^i = y\}$, a, y ∈ {0, 1}
- FPR & FNR

$$FPR(\hat{Y}) = \frac{\left|\{i : \hat{y}^i = 1, y^i = 0\}\right|}{\left|\{i : y^i = 0\}\right|}$$

$$FNR(\hat{Y}) = \frac{\left|\{i : \hat{y}^i = 0, y^i = 1\}\right|}{\left|\{i : y^i = 1\}\right|}$$

Models

- Penalizing Unfairness(based on relaxing the 0- 1 loss)
  - We will penalize the difference in the average distance from the decision boundary across different values of the protected attribute A.
  - Absolute Value Difference (AVD)   Squared Difference (SD) penalizer

$$R_{FP}^{AVD}(\theta; S) = \left| \frac{\sum_{x \in S_{00}} \theta^T x}{|S_{00}|} - \frac{\sum_{x \in S_{10}} \theta^T x}{|S_{10}|} \right| = \left| \theta^T \underbrace{\left(\frac{\sum_{x \in S_{00}} x}{|S_{00}|} - \frac{\sum_{x \in S_{10}} x}{|S_{10}|} \right)}_{\overline{x}} \right| = \left| \theta^T \overline{x} \right|$$

$$R_{FP}^{SD}(\theta; S) = \left(\theta^T \overline{x}\right)^2 .$$

The Importance of Incorporating Fairness in the Learning Phase

- An example
- X = (X1, X2) = {0, 1}^{2} —X_1= A is the protected attribute, and X_2 is a non-protected attribute—and a label in Y = {0, 1}. Given $\epsilon \in (0, \frac{1}{4})$
- The given distribution$D\varepsilon$
  - P[Y = 1] = 0.5
  - P[A = y|Y = y] = 1 − ϵ
  - P[X2 = y|Y = y] = 1 − 2 ϵ
  - that $D\varepsilon$ is defined s.t. A ⊥ X_2|Y

華東師範大學
EAST CHINA NORMAL UNIVERSITY

The Importance of Incorporating Fairness in the Learning Phase

- The Bayes optimal predictor with respect to the 0-1 loss is
  - $\hat{h}(X) = \underset{y \epsilon \{0, 1\}}{argmax}$ P[Y = y | X = x]
- which, in our case, gives $\hat{h}(X)$ = A, This classifier has 0-1 loss of only ϵ, However, in terms of fairness, it performs as badly as possible, as it induces the maximal possible differences in both the FPR and FNR rates across the two sub-populations in the distribution.

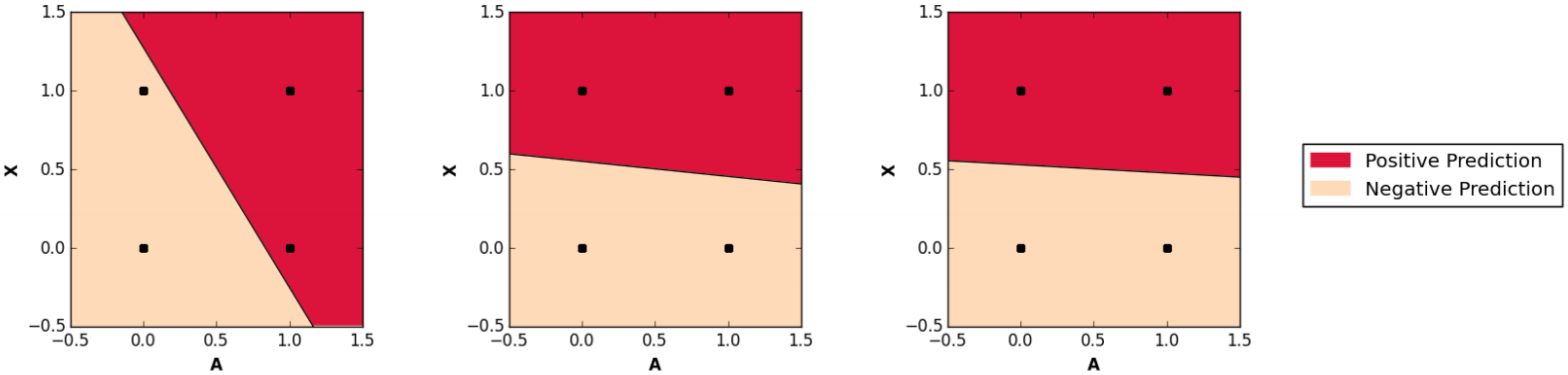
The Importance of Incorporating Fairness in the Learning Phase

- Any approach to post-processing this classifier for fairness (including, for example, the technique proposed by Hardt et al. (2016)) yields a classifier $\hat{Y}$ that predicts 0 or 1 at random, each with probability 0.5. While $\hat{Y}$ is a completely fair classifier, it only achieves trivial 0-1 loss of 0.5.

The Importance of Incorporating Fairness in the Learning Phase

$$
\begin{aligned}
\underset{\theta}{\text{minimize}} \quad & -ll(\theta; S) \\
& + c_1 R_{FP}(\theta; S) \\
& + c_2 R_{FN}(\theta; S) \\
& + q\,||\theta||_2^2
\end{aligned}
$$

placing weights of c1 = c2 = c for c ∈ {0, 300, 600}.



Experiments

$$\mathbf{D_{FPR}} = \left| FPR_{A=0}(\hat{Y}) - FPR_{A=1}(\hat{Y}) \right|$$

$$\mathbf{D_{FNR}} = \left| FNR_{A=0}(\hat{Y}) - FNR_{A=1}(\hat{Y}) \right|$$

	COMPAS Dataset								
	FPR Considerations			**FNR Considerations**			**Both Considerations**		
	Acc.	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**
Our Method (AVD Penalizers)	**0.660**	**0.01**	0.04	**0.653**	0.02	**0.04**	**0.654**	**0.02**	**0.04**
Our Method (SD Penalizers)	**0.664**	**0.02**	0.09	**0.661**	0.05	**0.03**	**0.661**	**0.02**	**0.03**
Zafar et al. (2017)	0.660	0.06	0.14	0.662	0.03	0.10	0.661	0.03	0.11
Zafar et al. Baseline (2017)	0.643	0.03	0.11	0.660	0.00	0.07	0.660	0.01	0.09
Hardt et al. (2016)	0.659	0.02	0.08	0.653	0.06	0.01	0.645	0.01	0.01
Vanilla Regularized Logistic Regression	**0.672**	**0.20**	**0.30**	**0.672**	**0.20**	**0.30**	**0.672**	**0.20**	**0.30**

Experiments