THE HEBREW UNIVERSITY OF JERUSALEM

Learning fair classifiers: A regularization-inspired approach

FAT / ML

{ Yahav Bechavod and Katrina Ligett } School of Computer Science and Engineering, Hebrew University of Jerusalem, Israel 9190401
yahav.bechavod, katrina @ cs.huji.ac.il

Abstract

We present a regularization-inspired approach for reducing bias in learned classifiers. In particular, we focus on binary classification tasks over individuals from two populations, where, as our criterion for fairness, we wish to achieve similar false positive rates in both populations, and similar false negative rates in both populations. As a proof of concept, we implement our approach and empirically evaluate its ability to achieve both fairness and accuracy, using the COMPAS scores data for prediction of recidivism.

Setting & Approach

- **Dataset:** $S = (x_i, a_i, y_i)_{i=1}^n$
  - $x_i \in \mathbb{R}^{d-1}$ - Non-protected features
  - $a_i \in \{0,1\}$ - Protected feature (e.g. race)
  - $y_i \in \{0,1\}$ - Label
- **FPR & FNR:** Given a classifier $\hat{Y} : \mathbb{R}^d \to \{0,1\}$, writing $\hat{y}_i = \hat{Y}(x_i, a_i)$, we can express the FPR and FNR as follows:

$$FPR = \frac{|\{i : \hat{y}_i = 1, y_i = 0\}|}{|\{i : y_i = 0\}|}$$

$$FNR = \frac{|\{i : \hat{y}_i = 0, y_i = 1\}|}{|\{i : y_i = 1\}|}$$

- **Boundary-based classifier:** Trained in the form of a decision boundary in the feature space. We denote by $f = g \circ h_\theta : \mathbb{R}^d \to \{0,1\}$, where $h_\theta : \mathbb{R}^d \to [0,1]$, and $g : [0,1] \to \{0,1\}$.
- **Regularization:** Our approach to learning a fair classifier is inspired by the concept of *regularization*, a common technique in machine learning for preventing overfitting.
- **Penalizers:** We define two penalizers, to be added to the loss function, that serve as a proxy for the difference in the FPR and FNR (respectively) between the groups in the population:

$$R_{FPR}(\theta; S^{neg}) = \left| \frac{\sum\limits_{i \in N_A^{neg}} h_\theta(x_i)}{|N_A^{neg}|} - \frac{\sum\limits_{i \in N_B^{neg}} h_\theta(x_i)}{|N_B^{neg}|} \right|$$

$$R_{FNR}(\theta; S^{pos}) = \left| -\frac{\sum\limits_{i \in N_A^{pos}} h_\theta(x_i)}{|N_A^{pos}|} + \frac{\sum\limits_{i \in N_B^{pos}} h_\theta(x_i)}{|N_B^{pos}|} \right|$$

Where:

$$N_A^{pos} = \{i : a_i = A, y_i = 1\}, N_A^{neg} = \{i : a_i = A, y_i = 0\}$$

$$N_B^{pos} = \{i : a_i = B, y_i = 1\}, N_B^{neg} = \{i : a_i = B, y_i = 0\}$$

Introduction

Concerns of bias in classification were at the center of a recent media stir regarding the potential hazards of computer algorithms for risk assessment in the criminal justice system. The COMPAS system is a proprietary algorithm developed by Northpointe Inc., widely used in the United States for risk assessment and recidivism prediction. At the center of the controversy was an investigative report conducted by ProPublica, which observed that although the COMPAS algorithm, when used to label individuals as either high or low risk for recidivism, demonstrated similar accuracy on whites and blacks, the *direction* of errors made on whites vs. blacks was very different – **the rate of individuals who were classified using the COMPAS algorithm to be ”high risk” but did not re-offend was almost twice as high for black individuals as for whites; among those who were classified as ”low risk” and did re-offend, the rate was significantly higher for whites than it was for blacks.**

Ground truth **Prediction made**	Re-offended	Did not re-offend
Will re-offend	🟥 True Positive (TP)	🟩 False Positive (FP)
Will not re-offend	🟦 False Negative (FN)	🟨 True Negative (TN)

Table 1: Confusion matrix for recidivism prediction

Implementation & Experiments

- **Implementation:** We instantiate our method in the context of logistic regression.
- **Optimization problem:**

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & -ll(\theta; S) \\ & + C_1 R_{FPR}(\theta; S^{neg}) \\ & + C_2 R_{FNR}(\theta; S^{pos}) \\ & + \frac{1}{2} C_3 \,||\theta||_2^2 \end{aligned}$$

Where: ll stands for log-likelihood, $C_1 > 0$, $C_2 > 0$, $C_3 > 0$ are constants to be tuned according to the desired trade-off between the components.

- **Training data:**
  - **Dataset:** COMPAS records from Broward County, Florida 2013-2014. Contains information of 5278 individuals, and whether or not they recidivated within 2 years of the screening.
  - **Features:** Age (<25, 25-45, >45), gender (male or female), race (black or white), priors count (0-37), charge degree (misconduct or felony).
  - **Target variable:** 2-year recidivism.
- **Results:**

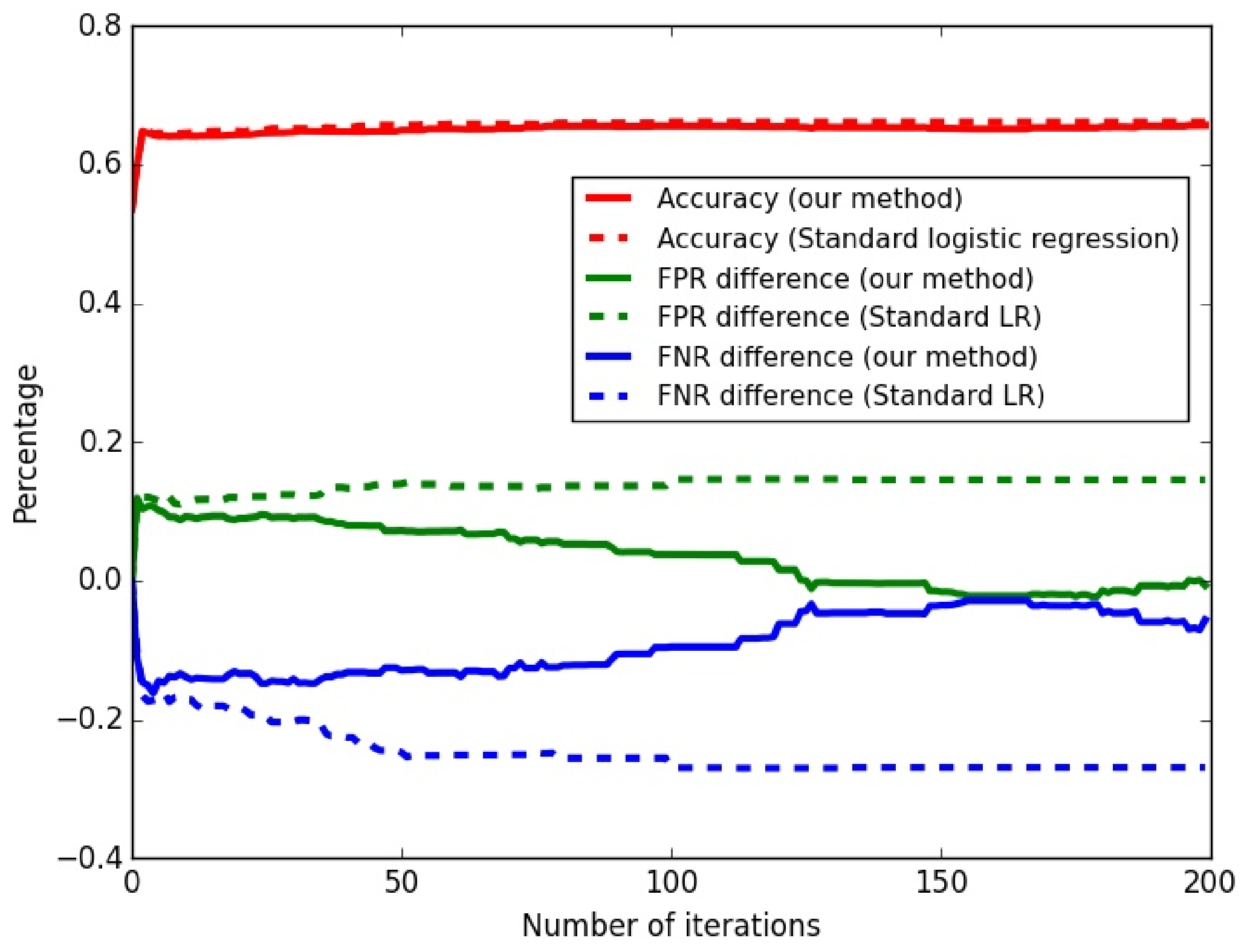

Figure 1: Accuracy, FPR difference, and FNR difference (evaluated on the hold-out data) of the learned classifier after each iteration of fairness-regularized gradient descent (”Our”), versus vanilla logistic regression (”LR”)

		FPR constraints			FNR constraints			Both constraints		
		Acc.	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$	Acc.	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$	Acc.	$\mathbf{D_{FPR}}$	$\mathbf{D_{FNR}}$
	Our Method	**0.646**	**0.0006**	**-0.0084**	**0.650**	**-0.010**	**-0.0008**	**0.646**	**0.0006**	**-0.0084**
	Bilal-Zafar et al.	0.660	0.06	−0.14	0.662	0.03	−0.10	0.661	0.03	−0.11
	Bilal-Zafar et al. Baseline	0.643	0.03	−0.11	0.660	0.00	−0.07	0.660	0.01	−0.09
	Hardt et al.	0.659	0.02	−0.08	0.653	−0.06	−0.01	0.645	−0.01	−0.01
	Logistic Regression (No fairness Constraints)	0.665	0.187	−0.326	0.665	0.187	−0.326	0.665	0.187	−0.326

Table 2: Performance comparison. Accuracy, FPR difference and FNR difference of the different methods as evaluated on hold-out set ($\mathbf{D_{FPR}} = FPR(Blacks) - FPR(Whites)$, $\mathbf{D_{FNR}} = FNR(Blacks) - FNR(Whites)$)

Discussion & Conclusions

- **Fairness and accuracy:** Our method succeeds in eliminating unfairness on the given dataset almost completely, while retaining high accuracy.
- **Compatibility:** Penalizers can be incorporated regardless of the specific boundary-based classifier in use.
- **Flexibility:** Parameters dictating the significance given to each of the penalizers can be tuned based on the requested trade-off of fairness and accuracy.
- **Computational Efficiency:** For gradient-based methods of training, additional cost of incorporating suggested penalizers is generally small.
- **Hardness of the general case:** As the general case of learning a fair classifier is hard, we resort to relaxations. Each of the methods compared in the experimental section bears its pros and cons.
- **Toolkit:** As no single method has proven to perform well in every setting, it is reasonable to expect that state-of-the-art in practical fair learning will best be served by a diverse toolkit of approaches.

Acknowledgements

This work was supported in part by NSF grants CNS-1254169 and CNS-1518941, US-Israel Binational Science Foundation grant 2012348, Israeli Science Foundation (ISF) grant #1044/16, a subcontract on the DARPA Brandeis Project, and the HUJI Cyber Security Research Center in conjunction with the Israel National Cyber Bureau in the Prime Minister's Office.