Learning Fair Scoring Functions

Bipartite Ranking under ROC-based Fairness Constraints

Robin Vogel[1,2], Aurélien Bellet[3] and Stephan Clémençon[1]   [1] Télécom Paris, [2] IDEMIA, [3] Inria

Introduction

Fairness is crucial to machine learning systems operating in very sensitive contexts, such as:
· in the banking sector,
· for diagnosis in medicine,
· for recidivism prediction in criminal justice.

Bipartite ranking formalizes many problems naturally such as **credit scoring** or biometric authentification.

Example 1 (Credit-risk screening**).**
A bank assigns the score $s(X)$ to a client and grants a loan if $s(X) > t$. The threshold t is unknown when learning s, as it depends on their risk aversion (low).

Contributions. We propose:
· a general formulation for AUC constraints,
· a **new** ROC-**based fairness constraint**,
· generalization guarantees for fair scoring,
· to learn fair scoring functions by gradient descent.

Preliminaries

Definitions. (X, Y, Z) r.v.'s in $\mathbb{R}^d \times \{-1, 1\} \times \{0, 1\}$. We predict Y using X, while Z is the sensitive group.

For any $z \in \{0, 1\}$, we set:
· $H^{(z)}$ is the distribution of $X \mid Y = -1, Z = z$,
· $G^{(z)}$ is the distribution of $X \mid Y = +1, Z = z$.

For any $s : \mathbb{R}^d \to \mathbb{R}$ and $F \in \{H, G\}$, we set $F_s^{(z)}$ as the distribution on $\mathbb{R}$ induced by s using $F^{(z)}$.
Notably $H_s^{(0)}(t) = \mathbb{P}\{s(X) \le t \mid Y = -1, Z = 0\}$.

The ROC **curve** is used to visualize the dissimilarity between two distributions h, g on $\mathbb{R}$,

$$\mathrm{ROC}_{h,g} : \alpha \in [0, 1] \to 1 - g \circ h^{-1}(1 - \alpha).$$

The $\mathrm{AUC}_{h,g}$ is the area under the $\mathrm{ROC}_{h,g}$ curve .

References

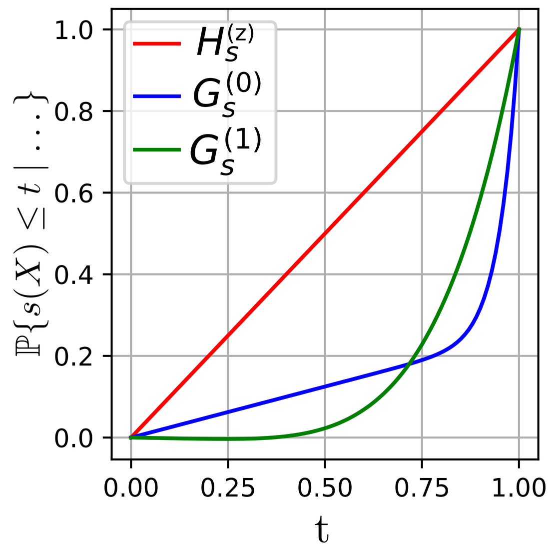
[1] Alex Beutel et al. Fairness in recommendation ranking through pairwise comparisons. In *SigKDD*, 2019.

[2] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, et al. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, 2019.

[3] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the XAUC metric. In *NeurIPS*. 2019.

Illustrating AUC fairness

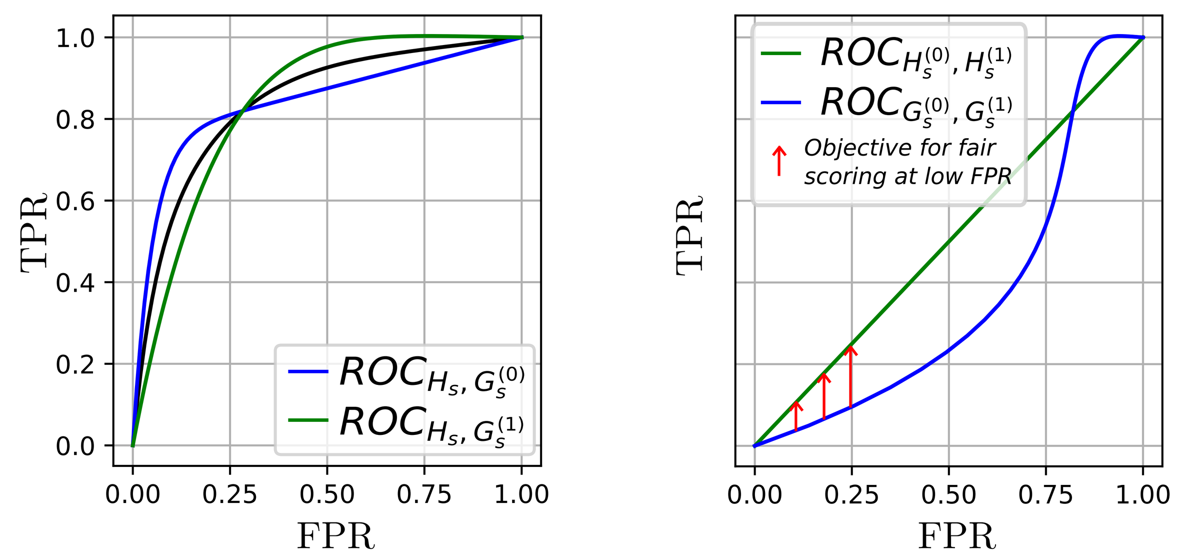
Consider s with the following distributions:



Notations for conditional score distributions

Group×Class	Y = −1	Y = +1
Z = 0	$H_s^{(0)}$	$G_s^{(0)}$
Z = 1	$H_s^{(1)}$	$G_s^{(1)}$
Z ∈ {0, 1}	H_s	G_s

Then $\mathrm{AUC}_{H_s, G_s^{(0)}} = \mathrm{AUC}_{H_s, G_s^{(1)}}$ (BNSP AUC [1]), but we have very different TPR's for low FPR's.



Therefore, any classifier $g_{s,t} : x \mapsto 2 \cdot \mathbb{I}\{s(x) > t\} - 1$ derived from s can be very **unfair in TPR**.

AUC-based fairness

Denote by (e_1, e_2, e_3, e_4) the canonical basis of $\mathbb{R}^4$, AUC constraints are equalities of AUC's between mixtures of $D(s) := (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})^\top$.
Given probability vectors $\alpha, \beta, \alpha', \beta'$, they write as:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)}. \qquad (1)$$

For example, [2] proposed the BNSP AUC, [1] (r. [3]) the intra-group (r. inter) pairwise AUC fairness.

We show that fairness constraints of the form eq. (1) are combinations of elementary constraints $C_l(s) = 0$:

$$\mathcal{C}_\Gamma(s) : \quad \Gamma^\top C(s) = \textstyle\sum_{l=1}^5 \Gamma_l C_l(s) = 0, \qquad (2)$$

where $\Gamma = (\Gamma_1, \dots, \Gamma_5)^\top \in \mathbb{R}^5$.

Theorem 1. *The following statements are equivalent:*
1. Eq. (1) is satisfied for any s when $H^{(0)} = H^{(1)}$, $G^{(0)} = G^{(1)}$ and $\eta(X)$ not a.s. constant.
2. Eq. (1) is equivalent to $\mathcal{C}_\Gamma(s)$ for some $\Gamma \in \mathbb{R}^5$,
3. $(e_1 + e_2)^\top [(\alpha - \alpha') - (\beta - \beta')] = 0$.

ROC-based fairness

AUC-based fairness implies that the ROC's intersect at some **unknown point** in the ROC plane.

We propose **pointwise** ROC **fairness constraints** as an alternative to AUC-based constraints.
For $\alpha \in [0, 1]$, consider:

$$\Delta_{G,\alpha}(s) := \mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) - \alpha,$$
$$(\text{resp. } \Delta_{H,\alpha}(s) := \mathrm{ROC}_{H_s^{(0)}, H_s^{(1)}}(\alpha) - \alpha).$$

Enforcing $G_s^{(0)} = G_s^{(1)}$ (resp. $H_s^{(0)} = H_s^{(1)}$) is equivalent to $\forall \alpha \in [0, 1]$, $\Delta_{G,\alpha}(s) = 0$ (resp. $\Delta_{H,\alpha}(s) = 0$).

We propose to satisfy a **finite number of constraints** on $\Delta_{H,\alpha}(s)$ and $\Delta_{G,\alpha}(s)$ for relevant values of α.
We denote them as $\alpha_F = [\alpha_F^{(1)}, \dots, \alpha_F^{(m_F)}]$ where $F = G$ for $\Delta_{G,\alpha}$ (resp. $F = H$ for $\Delta_{H,\alpha}$).

Constraints in **sup norm on an entire interval** can be derived from a small number of pointwise constraints.

Learning Scoring Functions

AUC-based fairness. Minimize $L_\lambda(s)$, *e.g.* equal to:

$$\mathrm{AUC}_{H_s, G_s} - \lambda \left| \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}} \right|,$$

where λ is a fairness regularization hyperparameter.

Generalization guarantees for the ERM of L_λ:
→ Rely on the theory of U-processes.

ROC-based fairness. Introducing $\Lambda := (\alpha, \lambda_H, \lambda_G)$, we minimize $L_\Lambda(s)$ defined as:

$$\mathrm{AUC}_{H_s, G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} \left| \Delta_{H, \alpha_H^{(k)}} \right| - \sum_{k=1}^{m_G} \lambda_G^{(k)} \left| \Delta_{G, \alpha_G^{(k)}} \right|,$$

where $\lambda_F = [\lambda_F^{(1)}, \dots, \lambda_F^{(m_F)}]$ are fairness regularization hyperparameters for any $F \in \{H, G\}$.

Generalization guarantees for the ERM of L_Λ:
→ the empirical ROC curve is almost a composition of empirical processes, we study its uniform deviation.

Experiments

We smooth empirical losses $\widehat{L}_\lambda$ and $\widehat{L}_\Lambda$ with the logistic function $x \mapsto 1/(1 + e^{-x})$ and maximize them with SGD.
Following the **low FPR objective**, ROC constraints penalize high $|\Delta_{G,1/8}|$, $|\Delta_{G,1/4}|$, $|\Delta_{H,1/8}|$ and $|\Delta_{H,1/4}|$.

Compas is a recidivism prediction dataset. Then $Z = 1$ if a sample is African-American, $Z = 0$ otherwise.
Being labeled **positive is a disadvantage**, thus we chose the BPSN AUC constraint $\mathrm{AUC}_{H_s^{(0)}, G_s} = \mathrm{AUC}_{H_s^{(1)}, G_s}$.

Adult is a salary prediction ($Y = 1$ if above 50K\$) dataset. Then $Z = 1$ if a sample is male, $Z = 0$ if female.
No obvious disadvantage from $Y = 1$ or $Y = -1$, thus we chose $\mathrm{AUC}_{H_s^{(0)}, G_s^{(1)}} = \mathrm{AUC}_{H_s^{(1)}, G_s^{(0)}}$