LEARNING FAIR SCORING FUNCTIONS FOR BIPARTITE RANKING

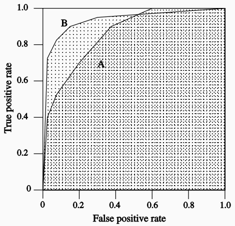
Aurélien Bellet (Inria Magnet)

Joint work with Robin Vogel and Stéphan Clémençon (Télécom Paris)

Workshop on Ethical AI at Inria Comète
September 30, 2022

LEARNING SCORING FUNCTIONS

- Algorithmic decisions often involve scoring individuals using a learned function of their attributes
- Decisions are usually taken based on whether the score exceeds a certain threshold, where the value of threshold depends on the context in which the decision is taken
- Examples: credit lending [Chen, 2018], medical diagnosis [Deo, 2015], recidivism prediction in criminal justice [Rudin et al., 2018]
- Fairness is a major concern in such applications!



- **Statistical framework:** same as in binary classification
  - Random variables (X, Y) with joint distribution P
  - $X \in \mathcal{X}$: observation (features)
  - $Y \in \{-1, +1\}$: binary label
- **Training dataset:** $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P$
- **Objective:** learn a scoring function $s : \mathcal{X} \to \mathbb{R}$ from $\mathcal{D}$ so that positive observations are ranked higher with high probability
  - Optimal scoring function orders elements by decreasing $\Pr[Y = +1 \mid X = x]$
- **Performance measures:** derived from the ROC curve
  - For any threshold $t \in \mathbb{R}$, we can define an induced classifier $g(X) = \mathbb{I}[s(X) > t]$
  - ROC: true positive rate (TPR) as a function of the false positive rate (FPR) when varying t
  - Common scalar summary: Area under the ROC curve (AUC)

- Sensitive group $Z \in \{0, 1\}$: we now have $\mathcal{D} = \{(X_i, Y_i, Z_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$

Motivating Example: credit-risk screening

- A bank grants a loan to a client with socio-economic features X if the score $s(X) > t$
- The risk aversion may vary so the precise value of t is unknown, but the bank is generally interested in regimes where the probability of default is small (low FPR).
- The bank would like to design a score function s that ranks higher the clients that are more likely to repay the loan (*ranking performance*), while ensuring that any t in the regime of interest leads to similar FNR across sensitive groups (*fairness constraint*)

- Learning a scoring function gives flexibility in thresholding the scores but we cannot rely on fairness notions that consider a single classifier!
- **How to define and guarantee fairness for a scoring function?**

- Previous work in different communities [Kallus and Zhou, 2019] [Beutel et al., 2019] [Borkan et al., 2019] introduced several fairness notions relevant to bipartite ranking
- For conciseness, denote the r.v. $s(X \mid y) := s(X)|Y = y$ and $s(X \mid y, z) := s(X)|Y = y, Z = z$

Intra-group pairwise	$\Pr[s(X \mid -1, 0) < s(X' \mid +1, 0)] = \Pr[s(X \mid -1, 1) < s(X' \mid +1, 1)]$
Inter-group pairwise	$\Pr[s(X \mid -1, 0) < s(X' \mid +1, 1)] = \Pr[s(X \mid -1, 1) < s(X' \mid +1, 0)]$
Background Neg. Subgroup Pos.	$\Pr[s(X, -1) < s(X', +1, 0)] = \Pr[s(X, -1) < s(X', +1, 1)]$

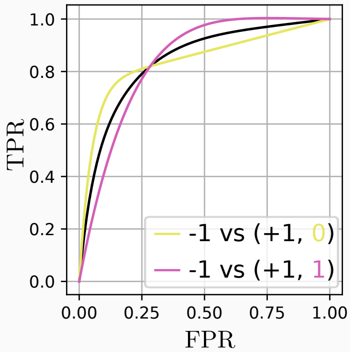
- We show that these are special cases of a general family AUC-based fairness notions, which we precisely characterize [Vogel et al., 2021]
- The choice of AUC-based fairness constraint depends on the use-case

- Recall our credit lending example and assume that the scoring function s satisfies Background Negative Subgroup Positive fairness:

$$\Pr[s(X,-1) < s(X',+1,0)] = \Pr[s(X,-1) < s(X',+1,1)]$$

- This means that creditworthy individuals from either group have the same probability of being ranked higher than a “bad borrower”
- Sounds good?

- The ROC curves associated with such s might look like this:
- High thresholds (low prob. of default) lead to unfair decisions
  - @FPR=10%, the FNR is 30% for group 0 and 60% for group 1
- There is a single threshold t at which the scoring function induces a classifier satisfying equal opportunity
- This threshold is **not** relevant for the use-case of interest (probability of default is too high!)



More generally, AUC-based fairness constraints only guarantee that there exists *some* $t \in \mathbb{R}$ for which s induces a fair classifier

ROC-BASED FAIRNESS CONSTRAINTS

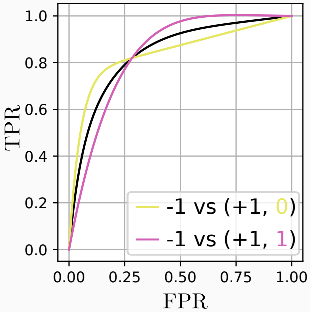
- We propose richer and more targeted fairness constraints
- Given a scoring function s, consider the conditional c.d.f.'s of s:

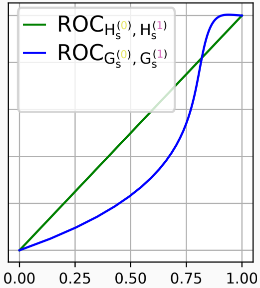
$$G_s^{(z)}(t) = \Pr[s(X) \leq t \mid Y = +1, Z = z]$$
$$H_s^{(z)}(t) = \Pr[s(X) \leq t \mid Y = -1, Z = z]$$

- Let's start from the "ideal fairness goal": enforcing $G_s^{(0)} = G_s^{(1)}$ and $H_s^{(0)} = H_s^{(1)}$
- This can be expressed in terms of ROC curves: for any $\alpha \in [0, 1]$

$$\mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) = \alpha$$
$$\mathrm{ROC}_{H_s^{(0)}, H_s^{(1)}}(\alpha) = \alpha$$

- When these conditions are satisfied, all AUC-based fairness constraints are satisfied and all induced classifiers are fair, but ranking performance is typically destroyed

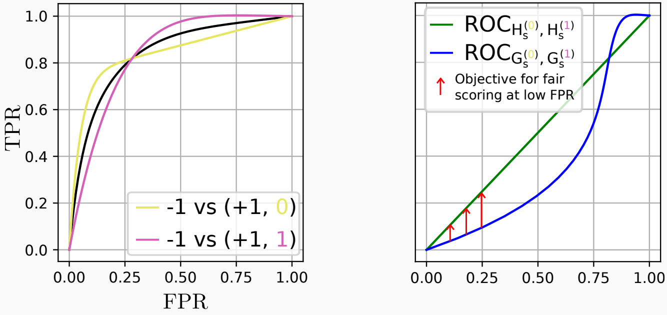




- Instead, we propose to enforce a finite number of pointwise constraints, providing fair classifiers when thresholding at the desired trade-offs (e.g., FPR vs FNR)
  - Discretization of interval $[\alpha_1, \alpha_2] \rightarrow$ classifiers are approximately fair in the whole interval
- For credit lending, we want fair classifiers in FNR for low FPR regimes: one could use

$$\mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) = \alpha, \quad \text{for } \alpha \in [0, \alpha_{\max}]$$

ROC-BASED FAIRNESS CONSTRAINTS



- Instead, we propose to enforce a finite number of pointwise constraints, providing fair classifiers when thresholding at the desired trade-offs (e.g., FPR vs FNR)
  - Discretization of interval $[\alpha_1, \alpha_2] \to$ classifiers are approximately fair in the whole interval
- For credit lending, we want fair classifiers in FNR for low FPR regimes: one could use

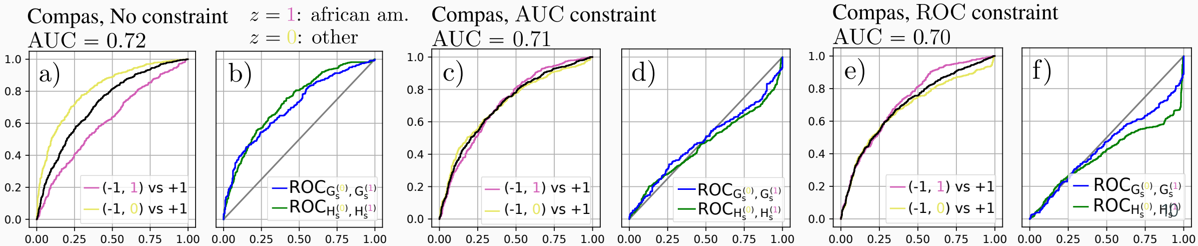
$$\mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) = \alpha, \quad \text{for } \alpha \in [0, \alpha_{\max}]$$

- We introduce empirical risk minimization formulations for learning fair scoring functions under AUC and ROC-based constraints
- We establish generalization bounds for fair bipartite ranking
- We propose efficient gradient-based training algorithms (*in-processing* approach)
- See the paper [Vogel et al., 2021] for details

ILLUSTRATION ON COMPAS

- Compas is a recidivism prediction dataset provided by ProPublica in their investigation of the COMPAS algorithm used in US courts
- No fairness constraint → more ranking errors for non-recidivist African-Americans
- As being labeled +1 (recidivist) is a disadvantage, we use BPSN AUC → still more of such errors in top 25% (the potential region of interest for decisions like denying bail)
- To address limitations of AUC-based fairness, we enforce:

$$\mathrm{ROC}_{G_s^{(0)}, G_s^{(1)}}(\alpha) = \alpha, \quad \mathrm{ROC}_{H_s^{(0)}, H_s^{(1)}}(\alpha) = \alpha, \quad \text{for } \alpha \in \{1/8, 1/4\}$$



- Predictive risk scores are used in many real-world applications of AI/ML
- The fairness of a scoring function can be defined based on ROC curves
- AUC-based fairness sets a global constraint on the full ordering $\rightarrow$ not so relevant when decisions are taken by thresholding the scores
- Pointwise ROC-based fairness allows more focused constraints and can ensure fairness for classifiers obtained by thresholding in a certain range
- Both types of constraints can used for training of the scoring function, with efficient algorithms and generalization guarantees

THANK YOU FOR YOUR ATTENTION!

QUESTIONS?

[Beutel et al., 2019] Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. (2019).
Fairness in recommendation ranking through pairwise comparisons.
In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*, pages 2212–2220. ACM.

[Borkan et al., 2019] Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019).
Nuanced metrics for measuring unintended bias with real data for text classification.
In *Companion of The 2019 World Wide Web Conference (WWW)*.

[Chen, 2018] Chen, J. (2018).
Fair lending needs explainable models for responsible recommendation.
CoRR, abs/1809.04684.

[Deo, 2015] Deo, R. (2015).
Machine learning in medicine.
Circulation, 132(20):1920–1930.

[Kallus and Zhou, 2019] Kallus, N. and Zhou, A. (2019).
The fairness of risk scores beyond classification: Bipartite ranking and the XAUC metric.
In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 3433–3443.

[Rudin et al., 2018] Rudin, C., Wang, C., and Coker, B. (2018).
The age of secrecy and unfairness in recidivism prediction.
CoRR, abs/1811.00731.

[Vogel et al., 2021] Vogel, R., Bellet, A., and Clémençon, S. (2021).
Learning Fair Scoring Functions: Bipartite Ranking under ROC-based Fairness Constraints.
In *AISTATS*.