Learning Fair Scoring Functions

Bipartite Ranking under ROC-based Fairness Constraints

Louise Davy

Télécom Paris

June 5, 2024

Outline

Outline

Introduction

Paper

- **Title** : Learning Fair Scoring Functions : Bipartite Ranking under ROC-based Fairness Constraints
- **Authors** : Robin Vogel, Aurélien Bellet, Stephan Clémençon
- **Year** : 2021
- **Arxiv link** : `https://arxiv.org/abs/2002.08159`
- **Github link** : `https://github.com/RobinVogel/Learning-Fair-Scoring-Functions`
- **Blogpost link** : `https://responsible-ai-datascience-ipparis.github.io/posts/lambert-davy/`

Summary

The paper addresses the problem of **fairness** in **bipartite ranking models**, which have different requirements than classification models.

The authors came up with **two contributions** to **improve fairness** of bipartite ranking models :

- AUC-based constraints
- ROC-based constraints

They show the **limitations** of the AUC-based constraints, and how the ROC-based constraints **address** them.

Outline

Outline

Ranking

What is ranking ?

Ranking is a class of machine learning algorithms aiming to **sort** a list of observations according to some **criterion**.

Examples

- Information retrieval : Sort documents according to their relevance to a query
- Recommendation systems : Recommend user's favourite songs first

Bipartite ranking

What is bipartite ranking ?

In bipartite ranking, we consider that all the observations that we want to sort can be partitioned into two classes : **positive** and **negative**. We want the positive instances to be consistently **ranked higher** than the negative ones.

Examples

- Fraud detection : Find the observations that are most likely to be fraudulent among fraudulent and non-fraudulent observations
- Recommendation systems : Recommend user's favourite songs first but this time we have songs that are liked by the user and songs that are disliked

Bipartite ranking

What is the difference between bipartite ranking and binary classification ?

Bipartite ranking is very close to binary classification since we are trying to distinguish positive instances from negative instances, but **serves a slightly different goal**.

Problem	Input space	Output	Risk
Classification	$\mathcal{X} \times \mathcal{Y}$	$c : \mathcal{X} \to \mathcal{Y}$	$\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim D} [\ell(\mathsf{Y}, s(\mathsf{X}))]$
Class-probability estimation	$\mathcal{X} \times \mathcal{Y}$	$\hat{\eta} : \mathcal{X} \to \Delta_{[\lvert\mathcal{Y}\rvert]}$	
Bipartite ranking	$\mathcal{X} \times \{\pm 1\}$	$s : \mathcal{X} \to \mathbb{R}$	$\mathbb{E}_{\mathsf{X}\sim P, \mathsf{X}'\sim Q} \ell_{\text{symm}}(s(\mathsf{X}) - s(\mathsf{X}'))$
Pairwise ranking	$\mathcal{X} \times \mathcal{X} \times \{\pm 1\}$	$s_{\text{Pair}} : \mathcal{X} \times \mathcal{X} \to \{\pm 1\}$	$\mathbb{E}_{(\mathsf{X},\mathsf{X}',\mathsf{Z})\sim R} \left[\ell(\mathsf{Z}, s_{\text{Pair}}(\mathsf{X}, \mathsf{X}'))\right]$

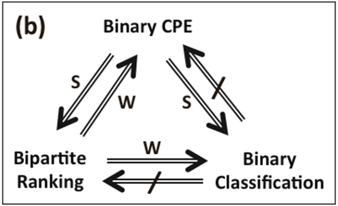
Figure: Differences summed up in[1]

[1] Menon and Williamson, "Bipartite Ranking: a Risk-Theoretic Perspective".

Bipartite ranking

What is the difference between bipartite ranking and binary classification ?

There are some works[2] working around the **link** between the two that were able to show that a **good ranking model**, once **transferred to binary classification**, will perform **well** (provided that the right threshold was found), while the **opposite is not always true**.



[2] Narasimhan and Agarwal, "On the relationship between binary classification, bipartite ranking, and binary class probability estimation".

Pairwise bipartite ranking

What is pairwise bipartite ranking ?

Pairwise bipartite ranking is specific case of bipartite ranking, in which we rank each instance **relatively to another instance**. Instead of simply distinguishing between positive and negative items, pairwise bipartite ranking considers the **relative preference between pairs of items**.

Example

- Facial recognition : Find pairs of faces that are the most similar in a database

(This is not the focus of this presentation, but this is what I'm currently working on.)

Outline

ROC curve

What is a ROC curve ?

ROC stands for **Receiver Operating Characteristic** curve and is a graph showing the performance of a classification model **at all classification thresholds**. It plots the false positive rate in the x-axis against the true positive rate in the y-axis.

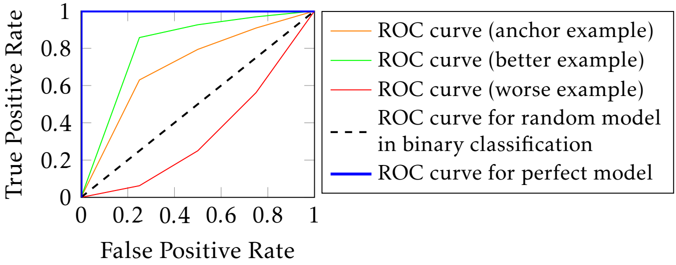


Figure: Different ROC curves

Warning : When a curve is below the diagonal (like for the red line), we can simply switch the labels of the classes and we will get a better model. This means that the worst possible model is actually the random model.

AUC

What is the AUC ?

AUC stands for **Area Under the Curve** and is a widely used metric for machine learning model evaluation that quantifies the overall performance of the model **across all possible classification thresholds**. AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).

Example

- A model who is 100% wrong has an AUC of 0.
- A model who is 100% correct has an AUC of 1.

AUC

What is the AUC ?

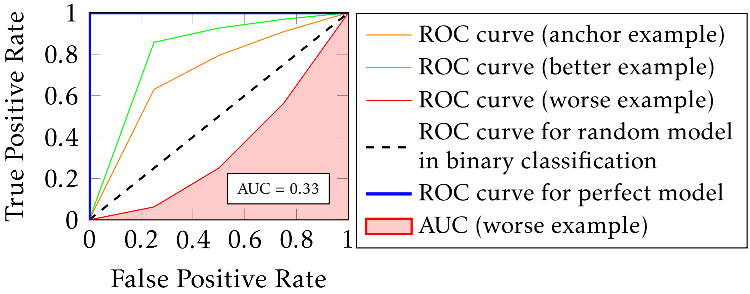


Figure: AUC for the worst ROC curve

AUC

What is the AUC ?

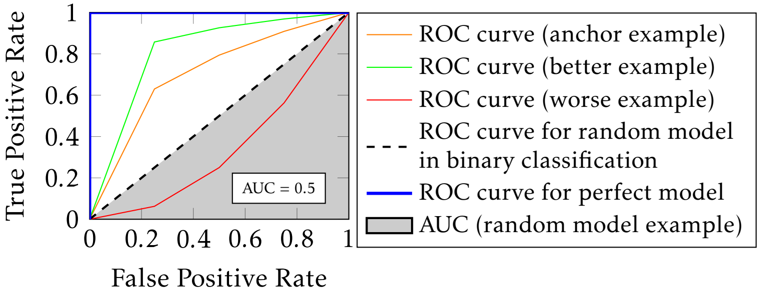


Figure: AUC for the random model ROC curve

AUC

What is the AUC ?

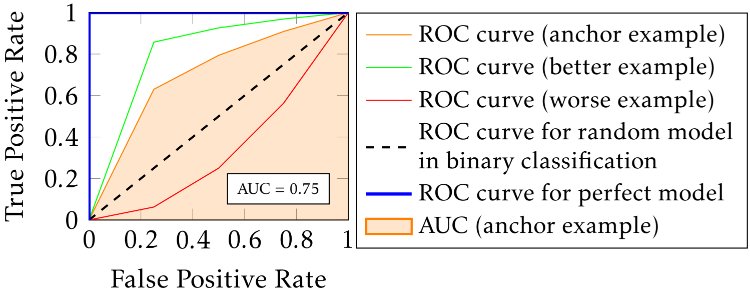


Figure: AUC for the anchor ROC curve

AUC

What is the AUC ?

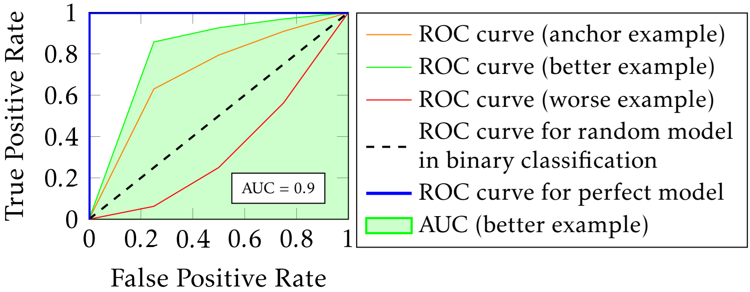


Figure: AUC for the better ROC curve

AUC

What is the AUC ?

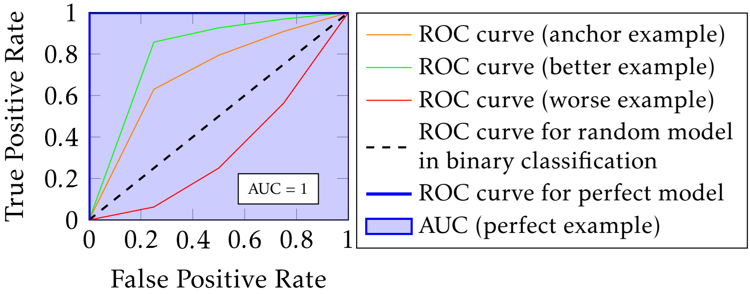


Figure: AUC for the perfect ROC curve

Outline

ROC and bipartite ranking

What is the link between ROC curves and bipartite ranking ?

- Different tasks require different metrics.
- **Classification** : accuracy, precision, recall, f1 score, etc.
- **Regression** : mean squared error, mean absolute error, etc.
- **None of these metrics take the rank into account.** They freeze the number of true/false positives/negatives for a **particular threshold** (usually 0.5).

ROC and bipartite ranking

- The ROC curve **intrinsically embeds the information of the rank** by giving information on the confusion matrix for all possible thresholds.

Rank	Real	Score
1	1	0.9
2	1	0.8
3	0	0.7
4	1	0.6
5	1	0.55
6	1	0.54
7	0	0.53
8	0	0.52
9	1	0.51
10	0	0.5
11	1	0.4
12	0	0.39
13	1	0.38
14	0	0.37
15	0	0.36
16	0	0.35
17	1	0.34
18	0	0.33
19	1	0.3
20	0	0.1

Table: Predictions

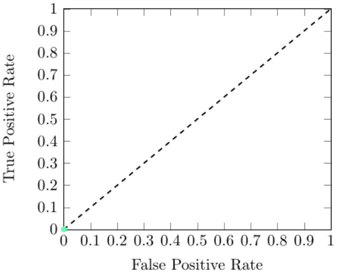


Figure: Construction of the ROC curve

ROC and bipartite ranking

- Therefore, the analysis of the ROC curve and its AUC are **the gold standard** to assess the performance of a **bipartite ranking model**.
- The ROC curve and AUC can also be directly used to **learn the ranking of the instances** in bipartite ranking models.
- Examples for AUC optimisation :[3][4]
- Examples for ROC curve pointwise optimisation :[5][6]

---

[3] Clémençon, Lugosi, and Vayatis, "Ranking and empirical minimization of U-statistics".

[4] Zhao et al., "Online AUC maximization".

[5] Vogel, Bellet, and Clémençon, "A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization".

[6] Lieberman et al., "Optimizing for ROC Curves on Class-Imbalanced Data by Training over a Family of Loss Functions".

Outline

Notations

- **Input space** : X, taking values in $\mathcal{X} \subset \mathbb{R}^d$, with $d \geq 1$
- **Output space** : Y, taking values in $\{-1, +1\}$
- **Sensitive attribute** : Z, taking values in $\{0, 1\}$ (when $Z = 1$, the individual is part of the sensitive group)
- **Scoring function** : $s : \substack{X \longrightarrow Y \\ x \longmapsto s(x)}$
- **TPR** = True Positive Rate = $\mathbb{P}\{s(X) > t | Y = +1\}$
- **TNR** = True Negative Rate = $\mathbb{P}\{s(X) \leq t | Y = -1\}$
- **FPR** = False Positive Rate = $\mathbb{P}\{s(X) > t | Y = -1\}$
- **FNR** = False Negative Rate = $\mathbb{P}\{s(X) \leq t | Y = +1\}$

Notations

- **Conditional distributions of X given Y :**
  $G = \mathbb{P}\{X|Y = +1\}$
  $H = \mathbb{P}\{X|Y = -1\}$
- **Cumulative distribution functions (CDFs) :**

$$\begin{aligned} G_s(t) &:= \mathbb{P}\{s(X) \leq t|Y = +1\} \\ &= G(s(X) \leq t) \\ &= \mathbf{FNR}(t) \end{aligned}$$

$$\begin{aligned} H_s(t) &:= \mathbb{P}\{s(X) \leq t|Y = -1\} \\ &= H(s(X) \leq t) \\ &= \mathbf{TNR}(t) \end{aligned}$$

Notations

- **ROC curve** : For a fixed **FPR** that we write $\alpha \in [0,1]$:
$$\begin{aligned} ROC(\alpha) &:= \mathbf{TPR}(\alpha) \\ &= 1 - \mathbf{FNR}(\mathbf{TNR}^{-1}(1-\alpha)) \\ &= 1 - G_s(H_s^{-1}(1-\alpha)) \end{aligned}$$
where $\mathbf{TNR}^{-1}(1-\alpha) = \mathbf{FPR}^{-1}(\alpha) = t_\alpha$.
- From now on, we will write $ROC_{H_s,G_s}(\alpha)$.
- Why do we use **FNR** and **TNR** instead of **TPR** and **FPR** ?
- Because they are cumulative distribution functions.
- We can finally define the **AUC** :
$$AUC_{H_s,G_s} = \int_0^1 ROC_{H_s,G_s}(\alpha) d\alpha = \mathbf{P}\{G_s > H_s\} + \frac{1}{2}\mathbf{P}\{G_s = H_s\}$$

Empirical counterparts

- **Training set** : $(X_i, Y_i)_{i=1}^n$ with n_+ positive examples and n_- negative examples.
- **Empirical** G_s **and** H_s :
  $\widehat{G}_s(t) := (\frac{1}{n_+}) \sum_{i=1}^n \mathbb{1}\{Y_i = +1, s(X_i) \leq t\}$
  $\widehat{H}_s(t) := (\frac{1}{n_-}) \sum_{i=1}^n \mathbb{1}\{Y_i = -1, s(X_i) \leq t\}$
- **Empirical ROC curve** :
  $\widehat{ROC}_{H_s,G_s} := ROC_{\widehat{H}_s,\widehat{G}_s}$
- **Empirical AUC** :
  $$\widehat{AUC}_{H_s,G_s} := AUC_{\widehat{H}_s,\widehat{G}_s} = \frac{1}{n_+ n_-} \sum_{i<j} K((s(X_i), Y_i), (s(X_j), Y_j))$$

  where, for any $t, t' \in \mathbf{R}^2, y, y' \in \{-1, +1\}^2$:
  $K((t, y), (t', y')) = \mathbb{1}\{(y - y')(t - t') > 0\} + \frac{1}{2}\mathbb{1}\{y \neq y', t = t'\}$.

Outline

Outline

Motivation

- Most of **fairness-aware** machine learning research focuses on **classification** models.
- Because bipartite ranking models are evaluated differently (*i.e*, with ROC curves), evaluating fairness for bipartite ranking models might be **more challenging**.
- However, learning a scoring function over a classifier adds **more flexibility** to the thresholds, which means that a fair scoring function will lead to **fair decisions for all thresholds of interest**.

Outline

AUC-based constraints

Previous works proposed to use **AUC-based constraints** to ensure fairness in bipartite ranking models. Let's denote $G_s^{(i)}$ (resp. $H_s^{(i)}$) the CDF of the score on the positive (resp. negatives) of group $z \in \{0, 1\}$.
Beutel et al.[7] proposed to use **intra-group pairwise** AUC fairness :

$$AUC_{H_s^{(0)}, G_s^{(0)}} = AUC_{H_s^{(1)}, G_s^{(1)}} \quad (1)$$

Reminder : $G_s(t) = \mathbf{P}\{s(X) \leq t | Y = +1\}$ and $H_s(t) = \mathbf{P}\{s(X) > t | Y = -1\}$

[7] Beutel et al., "Fairness in Recommendation Ranking through Pairwise Comparisons".

AUC-based constraints

Borkan et al.[8] proposed to use **Background Negative Subgroup Positive (BNSP)** AUC fairness :

$$AUC_{H_s, G_s^{(0)}} = AUC_{H_s, G_s^{(1)}} \tag{2}$$

Finally, Kallus and Zhou[9] proposed to use **inter-group pairwise** AUC fairness :

$$AUC_{H_s^{(0)}, G_s^{(1)}} = AUC_{H_s^{(1)}, G_s^{(0)}} \tag{3}$$

Reminder : $G_s(t) = \mathbf{P}\{s(X) \leq t | Y = +1\}$ and $H_s(t) = \mathbf{P}\{s(X) > t | Y = -1\}$

[8] Borkan et al., "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification".

[9] Kallus and Zhou, "The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric".

AUC-based constraints

What is the difference ?

- **Intra-group fairness** : equal performance **within** groups
- **Background Negative Subgroup Positive fairness** : positives **from either group** have the **same probability of being ranked higher** than a negative example
- **Inter-group fairness** (in this specific case) : positives of a group can be distinguished from the **negatives of the other group** as effectively for both groups

AUC-based constraints

There are **many more constraints** available ![10]. This is one of the **strengths** of the method proposed by the authors : **it can be adapted to any fairness constraint** we might be interested in thanks to the following framework :

$$AUC_{\alpha^\top D(s),\beta^\top D(s)} = AUC_{\alpha'^\top D(s),\beta'^\top D(s)}. \tag{4}$$

with $D(s) := (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})^\top$ and the probability vectors $\alpha, \beta, \alpha', \beta' \in \mathcal{P}$ where $\mathcal{P} = \{v \mid v \in \mathbb{R}_+^4, \mathbf{1}^\top v = 1\}$.

[10] If you're interested, check out the supplementary material of the original paper

AUC-based constraints

Equation 4 is **under-specified**, so the authors finally formulate all **relevant** constraints as a linear combination of **5 elementary constraints**.

$$C_1(s) = AUC_{H_s^{(0)},H_s^{(1)}} - 1/2,$$
$$C_2(s) = 1/2 - AUC_{G_s^{(0)},G_s^{(1)}},$$
$$C_3(s) = AUC_{H_s^{(0)},G_s^{(0)}} - AUC_{H_s^{(0)},G_s^{(1)}},$$
$$C_4(s) = AUC_{H_s^{(0)},G_s^{(1)}} - AUC_{H_s^{(1)},G_s^{(0)}},$$
$$C_5(s) = AUC_{H_s^{(1)},G_s^{(0)}} - AUC_{H_s^{(1)},G_s^{(1)}}.$$

The family of fairness constraints we consider is then the set of linear combinations of the $C_l(s) = 0$:

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

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

AUC-based constraints

The **learning problem** is defined as :

$$\max_{s \in \mathcal{S}} \quad AUC_{H_s,G_s} - \lambda |\Gamma^\top C(s)|,$$

where $\lambda \geq 0$ is a hyperparameter balancing ranking performance and fairness.
For example, for the **intra-group pairwise** AUC fairness constraint, we have :

$$L_\lambda(s) := AUC_{H_s,G_s} - \lambda \left| AUC_{H_s^{(0)},G_s^{(0)}} - AUC_{H_s^{(1)},G_s^{(1)}} \right|.$$

Limits of AUC-based constraints

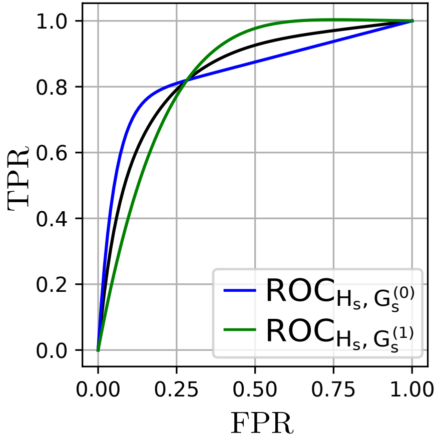


Figure: Illustrating the limitations of AUC-based fairness.

Outline

ROC-based constraints

For $\alpha \in [0,1]$, consider the deviations between the **positive** (resp. **negative**) **inter-group ROCs** and the **identity function**:

$$\Delta_{G,\alpha}(s) := ROC_{G_s^{(0)},G_s^{(1)}}(\alpha) - \alpha,$$
$$\left(\text{resp. } \Delta_{H,\alpha}(s) := ROC_{H_s^{(0)},H_s^{(1)}}(\alpha) - \alpha\right).$$

Ideally, we would want $\Delta_{G,\alpha}(s)$ and $\Delta_{H,\alpha}(s)$ to be equal to 0 for all $\alpha \in [0,1]$.
But this will most likely **jeopardize** the ranking performance as it is **too restrictive** !

ROC-base constraints

Instead, the authors propose a general approach to implement the satistifaction of a **finite number of fairness constraints** denoted by $m_H, m_G \in \mathbf{N}$ for the negatives and the positives respectively. We define $\alpha_H = [\alpha_H^{(1)}, \ldots, \alpha_H^{(m_H)}] \in [0,1]^{m_H}$ and $\alpha_G = [\alpha_G^{(1)}, \ldots, \alpha_G^{(m_G)}] \in [0,1]^{m_G}$ the points at which they apply (sorted in strictly increasing order).
The **learning objective** becomes:

$$L_\Lambda(s) = AUC_{H_s,G_s} - \sum_{k=1}^{m_H} \lambda_H^{(k)} \left|\Delta_{H,\alpha_H^{(k)}}(s)\right| - \sum_{k=1}^{m_G} \lambda_G^{(k)} \left|\Delta_{G,\alpha_G^{(k)}}(s)\right|,$$

$\lambda_H = [\lambda_H^{(1)}, \ldots, \lambda_H^{(m_H)}] \in \mathbf{R}_+^{m_H}$ and $\lambda_G = [\lambda_G^{(1)}, \ldots, \lambda_G^{(m_G)}] \in \mathbf{R}_+^{m_G}$ are hyperparameters.

Outline

Outline

Compas dataset

- The **COMPAS dataset** contains information about **defendants** and their **criminal history**.
- The goal is to predict whether a defendant will **re-offend**.
- It has been shown that the COMPAS algorithm is **biased against certain demographic populations** like African-Americans even though ethnicity was not included in the features.

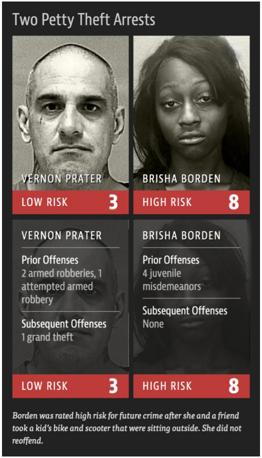


Figure: Bias from the COMPAS Algorithm

Outline

Adult dataset

- The **Adult dataset** contains information about **individuals** and their **income**.
- The goal is to predict whether an individual earns more than $50,000 a year.
- It has been shown that the Adult dataset is **biased against women**.

Outline

Results

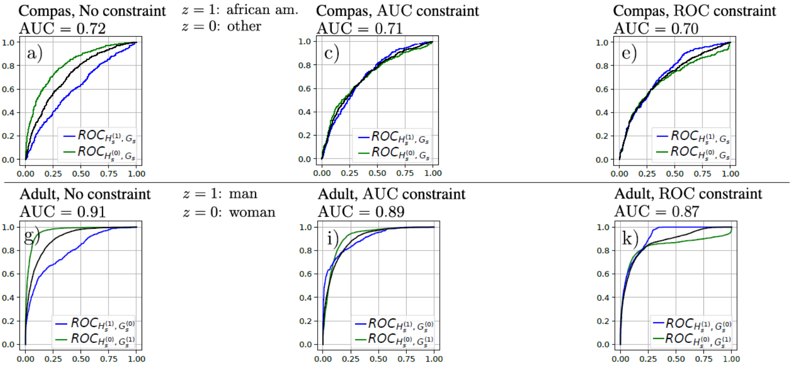


Figure: *ROC* curves on the test set of Adult and Compas for a score learned without and with fairness constraints. Black curves represent ROC_{H_s,G_s}.

Outline

1. Introduction
2. Definitions
3. Notations
4. Fairness in bipartite ranking
5. Applications
6. Conclusion

Conclusion

AUC-based fairness constraints

ROC-based fairness constraints

Bibliography I

- Beutel, Alex, Jilin Chen, Tulsee Doshi, et al. "Fairness in Recommendation Ranking through Pairwise Comparisons". In: *CoRR* abs/1903.00780 (2019). arXiv: 1903.00780. URL: http://arxiv.org/abs/1903.00780.
- Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, et al. "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification". In: *CoRR* abs/1903.04561 (2019). arXiv: 1903.04561. URL: http://arxiv.org/abs/1903.04561.
- Clémençon, Stéphan, Gábor Lugosi, and Nicolas Vayatis. "Ranking and empirical minimization of U-statistics". In: (2008).

Bibliography II

- Kallus, Nathan and Angela Zhou. "The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric". In: *CoRR* abs/1902.05826 (2019). arXiv: 1902.05826. URL: http://arxiv.org/abs/1902.05826.
- Lieberman, Kelsey, Shuai Yuan, Swarna Kamlam Ravindran, et al. "Optimizing for ROC Curves on Class-Imbalanced Data by Training over a Family of Loss Functions". In: *arXiv preprint arXiv:2402.05400* (2024).
- Menon, Aditya Krishna and Robert C. Williamson. "Bipartite Ranking: a Risk-Theoretic Perspective". In: *Journal of Machine Learning Research* 17.195 (2016), pp. 1–102. URL: http://jmlr.org/papers/v17/14-265.html.

Bibliography III

Narasimhan, Harikrishna and Shivani Agarwal. "On the relationship between binary classification, bipartite ranking, and binary class probability estimation". In: *Advances in neural information processing systems* 26 (2013).

Vogel, Robin, Aurélien Bellet, and Stéphan Clémençon. "A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 5065–5074. URL: https://proceedings.mlr.press/v80/vogel18a.html.

Zhao, Peilin, Steven CH Hoi, Rong Jin, et al. "Online AUC maximization". In: (2011).

.