

# Learning Fair Scoring Functions

## Bipartite Ranking under ROC-based Fairness Constraints

Louise Davy

Télécom Paris

June 4, 2024

# Outline

## ① Introduction

## ② Definitions

Ranking, bipartite ranking and pairwise bipartite ranking  
ROC curve and AUC  
ROC and bipartite ranking

## ③ Notations

## ④ Fairness in bipartite ranking

Motivation  
AUC-based constraints  
ROC based constraints

## ⑤ Applications

Compas dataset  
Adult dataset  
Results

## ⑥ Conclusion

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# Introduction

## Paper

- **Title** : Learning Fair Scoring Functions : Bipartite Ranking under ROC-based Fairness Constraints
- **Authors** : Robin Vogel, Aurélien Bellet, Stephan Cléménçon
- **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

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# Outline

## ① Introduction

## ② Definitions

Ranking, bipartite ranking and pairwise bipartite ranking  
ROC curve and AUC  
ROC and bipartite ranking

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# 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**. In the cases where a model needs to process a **large number of observations** and where a **human verification is needed**, a bipartite ranking model would be able to provide the **most likely positive instances** first, allowing the human to only investigate a limited number of instances.



# 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

## ① Introduction

## ② Definitions

Ranking, bipartite ranking and pairwise bipartite ranking  
ROC curve and AUC  
ROC and bipartite ranking

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# 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.

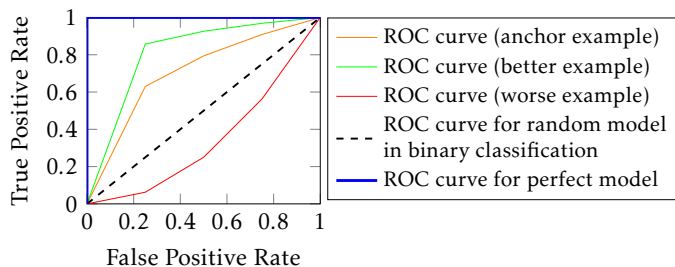


Figure: Different ROC curves

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 ?

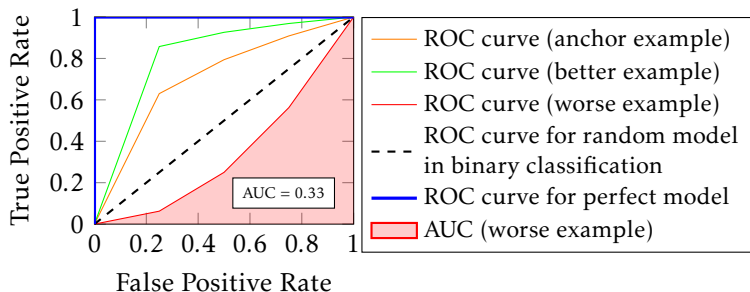


Figure: AUC for the worst ROC curve



# AUC

## What is the AUC ?

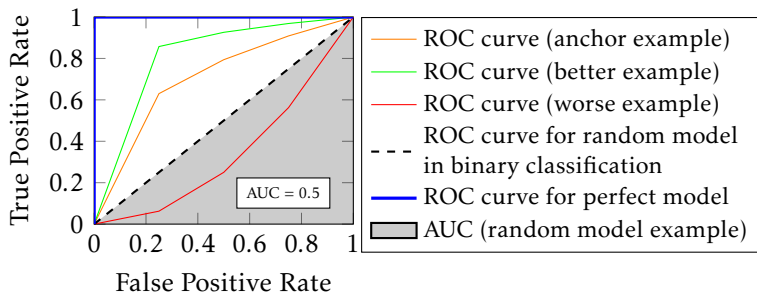


Figure: AUC for the random model ROC curve

# AUC

## What is the AUC ?

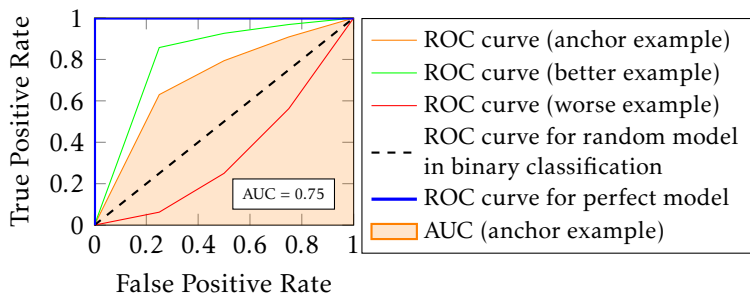


Figure: AUC for the anchor ROC curve

# AUC

## What is the AUC ?

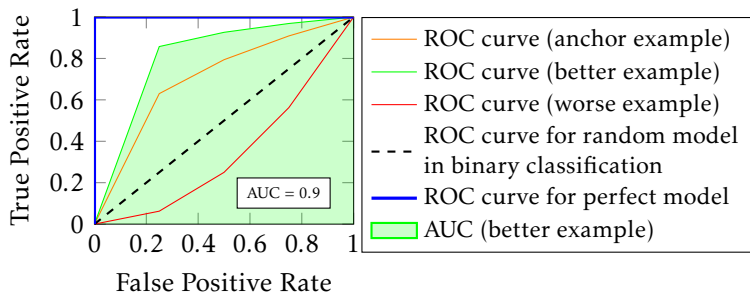


Figure: AUC for the better ROC curve

# AUC

## What is the AUC ?

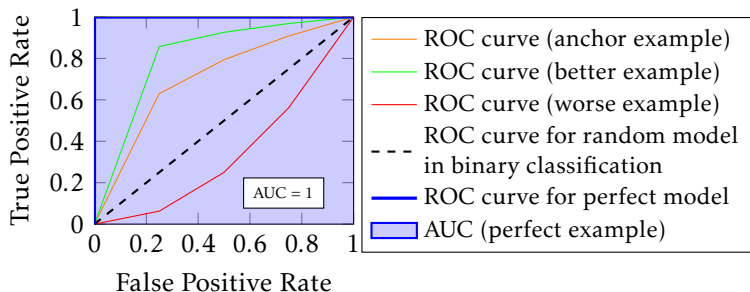


Figure: AUC for the perfect ROC curve

# Outline

## ① Introduction

## ② Definitions

Ranking, bipartite ranking and pairwise bipartite ranking  
ROC curve and AUC  
ROC and bipartite ranking

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# 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.

# 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 :<sup>23</sup>
- Examples for ROC curve pointwise optimisation :<sup>45</sup>

---

<sup>2</sup>Cléménçon, Lugosi, and Vayatis, “Ranking and empirical minimization of U-statistics”.

<sup>3</sup>Zhao et al., “Online AUC maximization”.

<sup>4</sup>Vogel, Bellet, and Cléménçon, “A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization”.

<sup>5</sup>Lieberman et al., “Optimizing for ROC Curves on Class-Imbalanced Data by Training over a Family of Loss Functions”.



# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# 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** :  $s : \begin{smallmatrix} X \rightarrow Y \\ x \mapsto s(x) \end{smallmatrix}$
- **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) :**

$$G_s(t) := \mathbb{P}\{s(X) \leq t|Y = +1\}$$

$$= G(s(X) \leq t)$$

$$= \mathbf{FNR}(t)$$

$$H_s(t) := \mathbb{P}\{s(X) \leq t|Y = -1\}$$

$$= H(s(X) \leq t)$$

$$= \mathbf{TNR}(t)$$

# Notations

- **ROC curve** : For a fixed **FPR** that we write  $\alpha \in [0, 1]$  :  
 $ROC(\alpha) := \mathbf{TPR}(\alpha)$

$$= 1 - \mathbf{FNR}(\mathbf{TNR}^{-1}(1 - \alpha))$$

$$= 1 - G_s(H_s^{-1}(1 - \alpha))$$

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  $H_s$**  :

$$\widehat{G}_s(t) := \left(\frac{1}{n_+}\right) \sum_{i=1}^n \mathbb{1}\{Y_i = +1, s(X_i) \leq t\}$$

$$\widehat{H}_s(t) := \left(\frac{1}{n_-}\right) \sum_{i=1}^n \mathbb{1}\{Y_i = -1, s(X_i) \leq t\}$$

- **Empirical ROC curve** :

$$\widehat{ROC}_{H_s, G_s} := ROC_{\widehat{H}_s, \widehat{G}_s}$$

- **Empirical AUC** :

$$\begin{aligned} \widehat{AUC}_{H_s, G_s} &:= AUC_{\widehat{H}_s, \widehat{G}_s} \\ &= \frac{1}{n_+ n_-} \sum_{i < j} K((s(X_i), Y_i), (s(X_j), Y_j)) \end{aligned}$$

where, for any  $t, t' \in \mathbb{R}^2, y, y' \in \{-1, +1\}^2$  :

$$K((t, y), (t', y')) = \mathbb{1}\{(y - y')(t - t') > 0\} + \frac{1}{2} \mathbb{1}\{y \neq y', t = t'\}.$$

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

### Motivation

AUC-based constraints

ROC based constraints

## ⑤ Applications

## ⑥ Conclusion

# 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

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

Motivation

AUC-based constraints

ROC based constraints

## ⑤ Applications

## ⑥ Conclusion

# 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.<sup>6</sup> proposed to use **intra-group pairwise** AUC fairness :

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

---

<sup>6</sup>Beutel et al., "Fairness in Recommendation Ranking through Pairwise Comparisons".

# AUC-based constraints

Borkan et al.<sup>7</sup> proposed to use **Background Negative Subgroup Positive (BNSP)** AUC fairness :

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

Finally, Kallus and Zhou<sup>8</sup> proposed to use **inter-group pairwise** AUC fairness :

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

<sup>7</sup>Borkan et al., “Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification”.

<sup>8</sup>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!<sup>9</sup>. 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 :

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

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  $\mathcal{P} = \{v \mid v \in \mathbf{R}_+^4, \mathbf{1}^\top v = 1\}$ .

---

<sup>9</sup>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  $C_l(s) = 0$ :

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

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



# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

Motivation

AUC-based constraints

ROC based constraints

## ⑤ Applications

## ⑥ Conclusion



Introduction  
○○○

Definitions  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Notations  
○○○○○

**Fairness in bipartite ranking**  
○○○○○○○○○○○○●

Applications  
○○○○○○○

Conclusion  
○○○○○

# ROC based constraints

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

Compas dataset

Adult dataset

Results

## ⑥ Conclusion

Introduction  
○○○

Definitions  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Notations  
○○○○○

Fairness in bipartite ranking  
○○○○○○○○○○○○○○○

**Applications**  
○○●○○○○

Conclusion  
○○○○○

# Compas dataset

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

Compas dataset

**Adult dataset**

Results

## ⑥ Conclusion

Introduction  
○○○

Definitions  
○○○○○○○○○○○○○○○○○○○○○○○○○○○○○○

Notations  
○○○○○

Fairness in bipartite ranking  
○○○○○○○○○○○○○○○○

**Applications**  
○○○○●○○

Conclusion  
○○○○○

# Adult dataset

# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

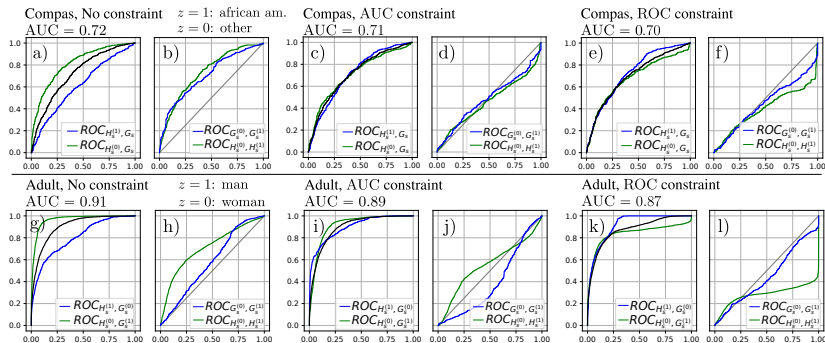
Compas dataset

Adult dataset

Results

## ⑥ Conclusion

# Results



**Figure:** ROC curves on the test set of Adult and Compas for a score learned without and with fairness constraints. Black curves represent  $ROC_{H_s, G_s}$ . We also report the corresponding ranking performance  $AUC_{H_s, G_s}$ .



# Outline

## ① Introduction

## ② Definitions

## ③ Notations

## ④ Fairness in bipartite ranking

## ⑤ Applications

## ⑥ Conclusion

# Conclusion

ROC based constraints are great

# 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>.



Cléménçon, Stéphan, Gábor Lugosi, and Nicolas Vayatis. “Ranking and empirical minimization of U-statistics”. In: (2008).



# Bibliography III



Vogel, Robin, Aurélien Bellet, and Stéphan Cléménçon. “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).

.