Learning Fair Scoring Functions

Bipartite Ranking under ROC-based Fairness Constraints

Louise Davy

Télécom Paris

June 3, 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/`

Outline

Outline

Ranking

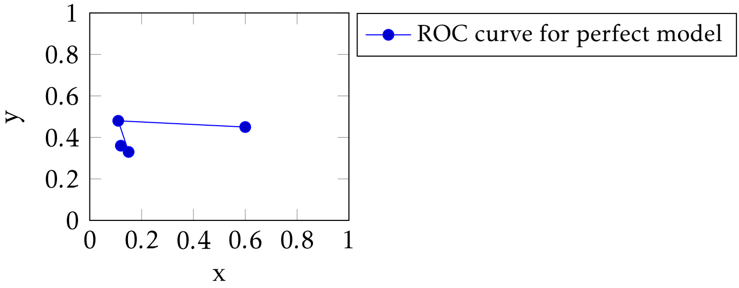
What is ranking ?

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

Ranking



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

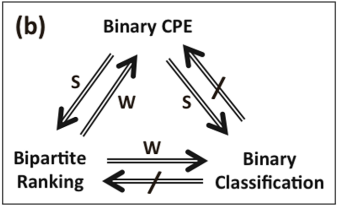
What is the difference between bipartite ranking and binary classification ?

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.

Bipartite ranking

What is the difference between bipartite ranking and binary classification ?

There are some works[1] 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.



[1] 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.

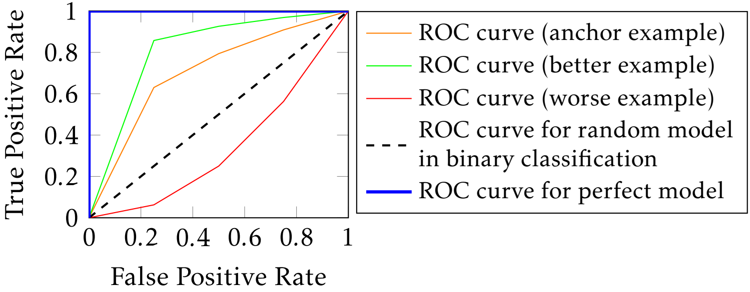


Figure: Different ROC curves

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 ?

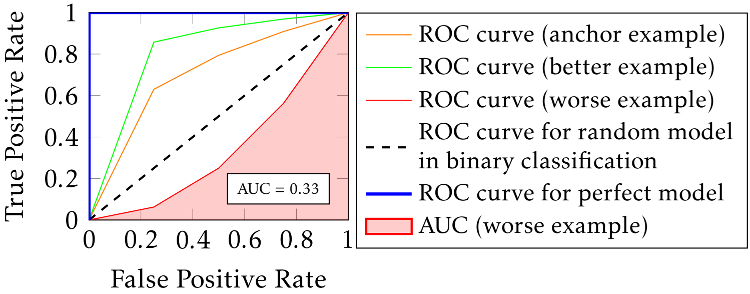


Figure: AUC for the worst ROC curve

AUC

What is the AUC ?

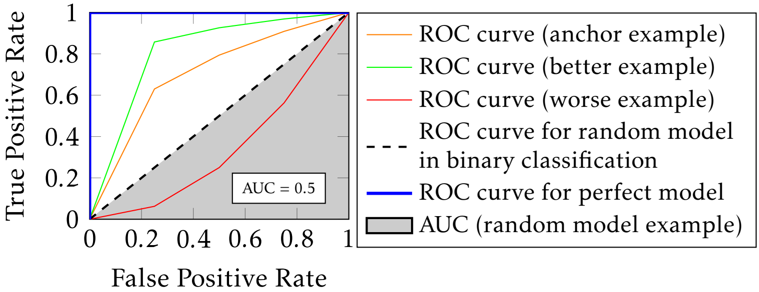


Figure: AUC for the random model ROC curve

AUC

What is the AUC ?

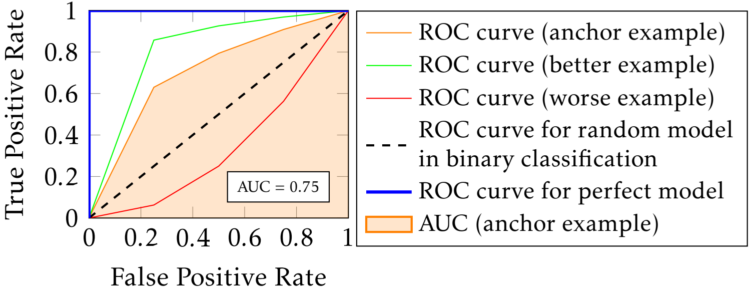


Figure: AUC for the anchor ROC curve

AUC

What is the AUC ?

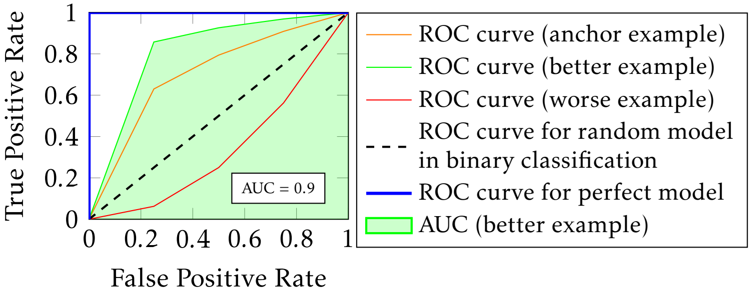


Figure: AUC for the better ROC curve

AUC

What is the AUC ?

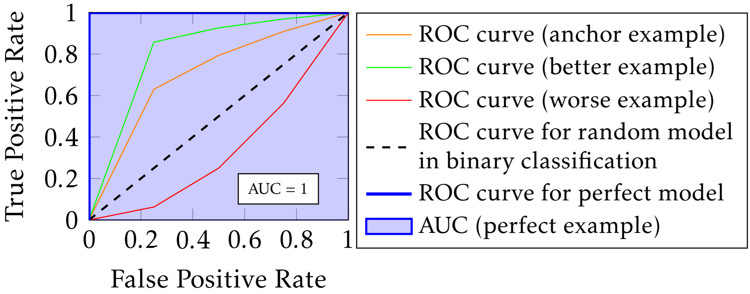


Figure: AUC for the perfect ROC curve

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

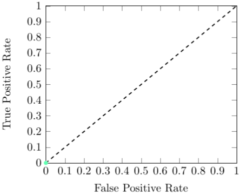


Figure: Construction of the ROC curve

Therefore, the analysis of the ROC curve is a **common solution** to assess the performance of a **ranking model**.

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 $\{0, 1\}$
- **Scoring function** : $s : \substack{X \to Y \\ x \mapsto s(x)}$
- **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 :**

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

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

Notations

- **ROC curve** : For a fixed **FPR** that we note $\alpha \in [0,1]$:
$$\begin{aligned} ROC(\alpha) &= \mathbf{TPR}(\alpha) \\ &= 1 - \mathbf{FNR}(\mathbf{TNR}^{-1}(1-\alpha)) \\ &= 1 - G_s(H_s^{-1}(1-\alpha)) \end{aligned}$$
- Where $\mathbf{TNR}^{-1}(1-\alpha) = \mathbf{FPR}^{-1}(\alpha) = t_\alpha$
- From now on, we will note $ROC_{H_s,G_s}(\alpha) = ROC(\alpha)$

Notations

- Why do we use **FNR** and **TNR** instead of **TPR** and **FPR** ?
- Because they are cumulative distribution functions and **FPR** and **TPR** are not.
- We can finally define **AUC** : $AUC_{H_s,G_s} = \int_0^1 ROC_{H_s,G_s}(\alpha)d\alpha$

Empirical counterparts

- **Training set** : $(X_i, Y_i)_{i=1}^n$ with n_+ positive examples and n_- negative examples.
- **Empirical of G_s and H_s** :

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

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

Outline

Contributions

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.

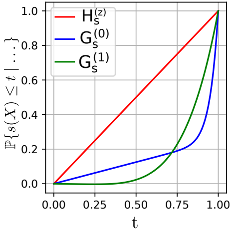
Motivation

- The vast majority of fairness-aware machine learning research focuses on classification models.
- However, many real-worl applications require bipartite ranking models.
- Because they are evaluated differently (*i.e*, with ROC curves), evaluating fairness for bipartite ranking models might also 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

Limits of AUC-based constraints



Notations for conditional score distributions

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

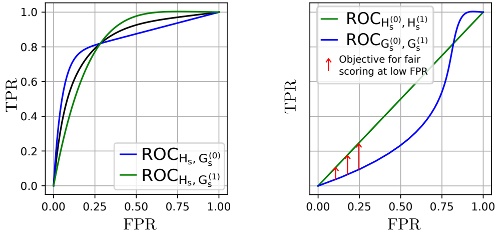


Figure: Illustrating the limitations of AUC-based fairness.

Outline

ROC based constraints

small change another change final change hopefully ok now it works

Outline

1. Introduction
2. Definitions
3. Notations
4. Contributions
5. Applications
6. Conclusion

Outline

Compas dataset

Outline

Adult dataset

Outline

Results

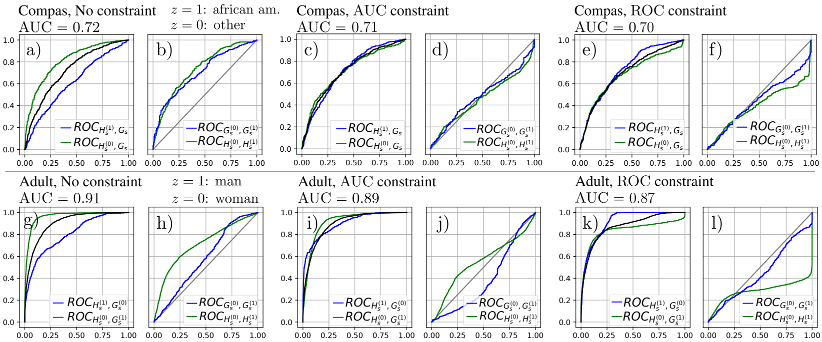


Figure: ROC curves on the test set of Adult and Compas for a score learned without and with fairness constraints. Black curves represent ROC_{H_s,G_s}. We also report the corresponding ranking performance AUC_{H_s,G_s}.

Outline

Conclusion

ROC based constraints

Bibliography

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).

.