Learning Fair Scoring Functions

Robin Vogel[1,2] Aurélien Bellet[3] Stephan Clémençon[2]

[1] IDEMIA, [2] Télécom Paris, [3] Inria

AISTATS 2021

Fairness for ranking/scoring

Lots of recent papers focus on **fairness in classification**.

Binary classification: $(X, Y) \sim P$ and $(X, Y) \in \mathcal{X} \times \{-1, 1\}$,
learn a classifier $g : \mathcal{X} \to \{-1, 1\}$ from data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$.

Fairness: Sensitive information $Z \in \{0, 1\}$, a Z_i for each (X_i, Y_i).
e.g. gender, ethnicity, …

Example of constraints: Parity in …

· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$,

· FPR: $\mathbb{P}\{g(X) = 1 \mid Z = 0, Y = -1\} = \mathbb{P}\{g(X) = 1 \mid Z = 1, Y = -1\}$,

· TPR, …

Fairness for ranking/scoring

Lots of recent papers focus on **fairness in classification**.

Binary classification: $(X, Y) \sim P$ and $(X, Y) \in \mathcal{X} \times \{-1, 1\}$,
learn a classifier $g : \mathcal{X} \to \{-1, 1\}$ from data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$.
Fairness: Sensitive information $Z \in \{0, 1\}$, a Z_i for each (X_i, Y_i).
e.g. gender, ethnicity, ...

Example of constraints: Parity in ...
· Error: $\mathbb{P}\{g(X) \neq Y \mid Z = 0\} = \mathbb{P}\{g(X) \neq Y \mid Z = 1\}$,
· FPR: $\mathbb{P}\{g(X) = 1 \mid Z = 0, Y = -1\} = \mathbb{P}\{g(X) = 1 \mid Z = 1, Y = -1\}$,
· TPR, ...

Fairness in scoring/ranking is a recent a research topic.

Scoring: $(X, Y) \sim P$ and $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$,
learn a score $s : \mathcal{X} \to \mathbb{R}$ from data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P$.

Contributions in fairness for ranking

See *r.h.s.* table for the problem distributions.
e.g. $H_s^{(0)} = \mathbb{P}\{s(X) \leq t \mid Y = -1, Z = 0\}$.

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

Contributions in fairness for ranking

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

See *r.h.s.* table for the problem distributions.
e.g. $H_s^{(0)} = \mathbb{P}\{s(X) \leq t \mid Y=-1, Z=0\}$.

The ROC **curve** represents dissimilarity between dist. 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 .

Perf. measure: ROC_{H_s,G_s}: the true positive rate (TPR) for a false positive rate (FPR) for the test $Y=+1$ with $s(X) > t$.

Contributions in fairness for ranking

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

See *r.h.s.* table for the problem distributions.
e.g. $H_s^{(0)} = \mathbb{P}\{s(X) \leq t \mid Y = -1, Z = 0\}$.

The ROC **curve** represents dissimilarity between dist. 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 .

Perf. measure: ROC_{H_s,G_s}: the true positive rate (TPR) for a false positive rate (FPR) for the test $Y = +1$ with $s(X) > t$.

Fairness measure: BNSP constraint $\mathrm{AUC}_{H_s,G_s^{(0)}} = \mathrm{AUC}_{H_s,G_s^{(1)}}$.

Contributions in fairness for ranking

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

See *r.h.s.* table for the problem distributions.
e.g. $H_s^{(0)} = \mathbb{P}\{s(X) \le t \mid Y=-1, Z=0\}$.

The ROC **curve** represents dissimilarity between dist. 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 .

Perf. measure: ROC_{H_s,G_s}: the true positive rate (TPR) for a false positive rate (FPR) for the test $Y=+1$ with $s(X) > t$.

Fairness measure: BNSP constraint $\mathrm{AUC}_{H_s,G_s^{(0)}} = \mathrm{AUC}_{H_s,G_s^{(1)}}$.

Our contributions:

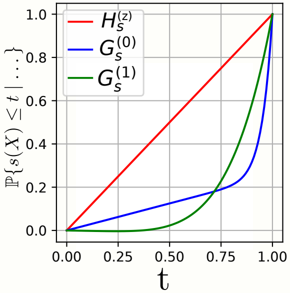
· A **general formulation** for AUC-based fairness constraints,
· A new, restrictive type of **constraint**: ROC-based constraints,
· A **gradient descent** for learning fair scores.

Illustrative Example

The problem distributions:

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

... represented:

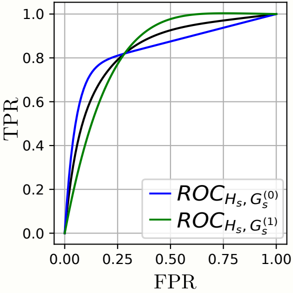


Illustrative Example

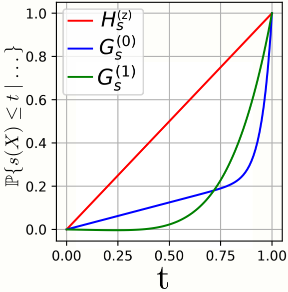
The problem distributions:

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

They satisfy an AUC constraint:



... represented:

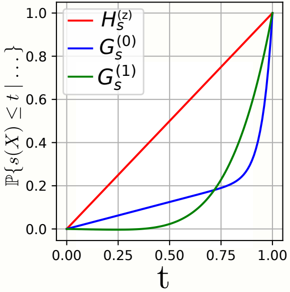


Illustrative Example

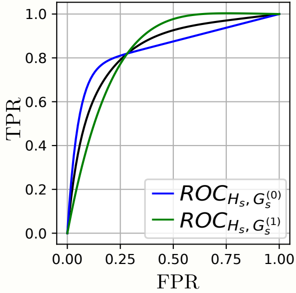
The problem distributions:

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

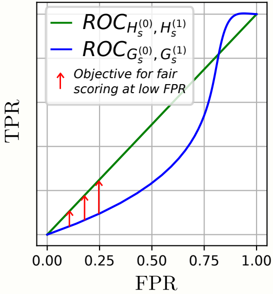
... represented:



They satisfy an AUC constraint:



... but are unfair in some situations:



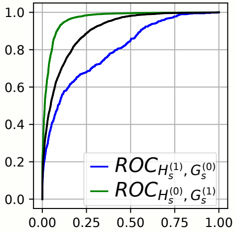
Practical Results

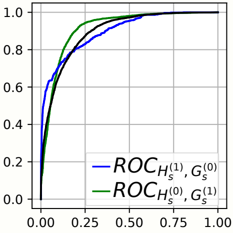
$z = 1$: man
$z = 0$: woman

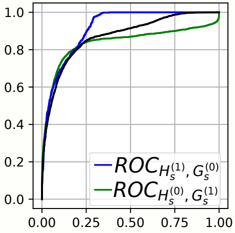
Adult, No constraint
AUC $= 0.91$

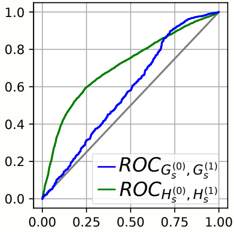
Adult, AUC constraint
AUC $= 0.89$

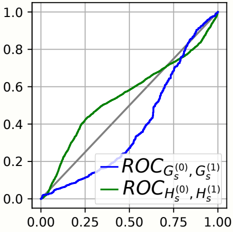
Adult, ROC constraint
AUC $= 0.87$

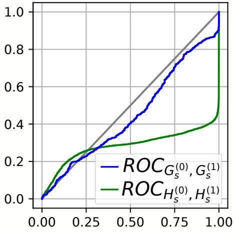












Thank you !

Come and see our poster !