Bias in facial recognition
A fair scoring proposition

Robin Vogel

Doctorant CIFRE IDEMIA / Télécom Paris

23/06/2020

Outline

A controversial technology

Face recognition (FR) is controversial, historically for privacy issues, now also for bias issues.

The New York Times

A.I. Experts Question Amazon's Facial-Recognition Technology

03/04/2019

The New York Times

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

19/12/2019

media coverage of:

FRVT FACE RECOGNITION VENDOR TEST

TC

IBM ends all facial recognition business as CEO calls out bias and inequality

09/06/2020

THE WALL STREET JOURNAL.

Amazon Suspends Police Use of Its Facial-Recognition Technology

Move comes after IBM said it was curtailing its facial-recognition activities amid widespread concerns about bias

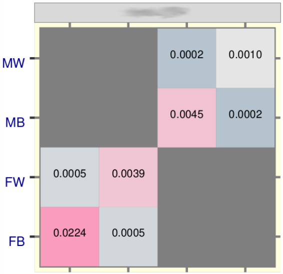
10/06/2020

Findings of NIST FRVT

Introduction to biometrics - 1:1 identification:
Set the face of a person x and a passport image x',
their similarity is quantified as $s(x, x')$.

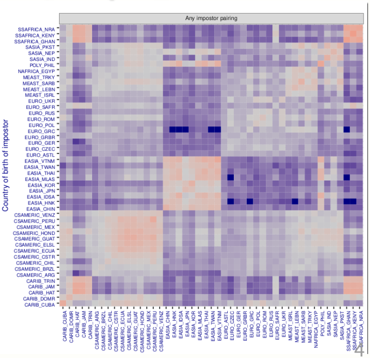
Given a threshold t:
$\rightarrow s(x, x') > t$ is a match (if wrong, false match), $\rightarrow$ more critical !
$\rightarrow s(x, x') \leq t$ is a reject (if wrong, false reject).

Now evaluate averages on a database to get rates.



False Match Rate
set to 0.001 on MW

MW= Male white
MB = Male black
FW = Female white
FB = Female black



Related research

Fairness of facial recognition technologies is a recent topic:
$\rightarrow$ lack of practical propositions to correct for unfairness.

However, public databases have been proposed recently.

FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age

Kimmo Kärkkäinen, UCLA
Jungseock Joo, UCLA

Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network

Mei Wang[1], Weihong Deng[1*], Jiani Hu[1], Xunqiang Tao[2], Yaohai Huang[2]
[1]Beijing University of Posts and Telecommunications, [2]Canon Information Technology (Beijing) Co., Ltd
[1]{wangmei1, whdeng, jnhu}@bupt.edu.cn, [2]{taoxunqiang, huangyaohai}@canon-ib.com.cn

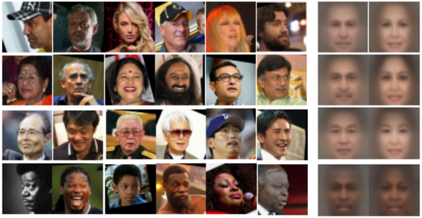


Figure 1. Examples and average faces of RFW database. In rows top to bottom: Caucasian, Indian, Asian, African.

Outline

Fairness in algorithmic decisions

Algorithmic decisions are increasingly used in many domains:

Banking (*e.g.* loans)	Recruting (*e.g.*, hiring)
Insurance (*e.g.* cars)	Judiciary (*e.g.*, bail)

Recently, the fairness of algorithms has gathered lots of attention.
e.g. May 2016: The COMPAS system assesses the likelihood of recidivism of a defendant for U.S. courts.

While algorithms are usually designed for the interest of some user, fair algorithms suggests confronting those to the law.

"Predictive models are really just opinions embedded in math."

Cathy O'Neil.

Fairness in ML - Literature review

Fairness of ML has gathered lots of attention recently.

In binary classification:

· A flexible approach for relaxed constraints [Zafar et al., 2019],
· ERM guarantees [Donini et al., 2018].

Other notable works:

· Textbook (WIP) on fairness in ML [Barocas et al., 2019],
· "Adversarially fair" representations [Madras et al., 2018].

Fairness in ranking became only recently a research topic, mostly tackled by the information retrieval (IR) community.

Some authors:

· modify a fixed score to induce a notion of fairness [Zehlike et al., 2017, Biega et al., 2018],
· introduce fairness in exposure over several rankings [Singh and Joachims, 2018, Singh and Joachims, 2019],
· use a notion of fairness based on the AUC [Borkan et al., 2019, Beutel et al., 2019].

Fairness in ML - Classification example

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, …

Fairness without ground truth: Parity in …

- Treatment: $g(X, Z) = g(X)$ almost surely.
  i.e. the decision does not depend on the sensitive attribute.
- Impact: $\mathbb{P}\{g(X) = +1 | Z = 0\} = \mathbb{P}\{g(X) = +1 | Z = 1\}$.

Fairness with ground truth: 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, …

Outline

Bipartite ranking (1/2)

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 $\{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P$.

Objective: Order new elements $X_1', \dots, X_m'$ by relevance,
i.e. by decreasing posterior probability $\eta(x) := \mathbb{P}\{Y = +1 \mid X = x\}$.

Perf. measure: The ROC curve: the true positive rate (TPR) for any false positive rate (FPR) for testing $Y = +1$ with $s(X) > t$.

Introduce the distributions (cdf) of $s(X)|Y = -1$ and $s(X)|Y = +1$ as:

$$H_s(t) = \mathbb{P}\{s(X) \le t \mid Y = -1\} \quad \text{and} \quad G_s(t) = \mathbb{P}\{s(X) \le t \mid Y = +1\}.$$

In the context of **fairness**, we denote by:
· $H_s^{(z)}$ the cdf of $s(X)|Y = -1, Z = z$,
· $G_s^{(z)}$ the cdf of $s(X)|Y = +1, Z = z$,
for any $z \in \mathcal{Z}$ with $\mathcal{Z} = \{0, 1\}$.

Bipartite ranking (2/2)

Let $\bar{F} = 1 - F$ and define the pseudo-inverse of F as:

$$F^{-1} : u \mapsto \inf\{t \mid F(t) > u\}.$$

The FPR (resp. TPR) of s at threshold t is equal to $\overline{H}_s(t)$ (resp. $\overline{G}_s(t)$).
Formally, the ROC and AUC write:

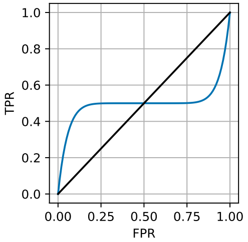
$$\mathrm{ROC}_{H_s,G_s}(\alpha) = \overline{G}_s \circ \overline{H}_s^{-1}(\alpha) \quad \text{and} \quad \mathrm{AUC}_{H_s,G_s} = \int_0^1 \mathrm{ROC}_{H_s,G_s}(\alpha)\, d\alpha.$$

The ROC **measures the difference** between two cdfs in $\mathbb{R}$.

Specifically, given two distributions F, F' on $\mathbb{R}$:

$$\forall \alpha \in [0,1], \quad \mathrm{ROC}_{F,F'}(\alpha) = \alpha \quad \Leftrightarrow \quad F = F'.$$

The AUC is a **scalar summary** of the ROC.



Our contributions

In [Vogel et al., 2020] with Aurélien Bellet and Stephan Clémençon, we focus on fairness for bipartite ranking, and provide:

- A **general formulation** for AUC-based fairness constraints,
- **Guarantees** for learning under AUC-based constraints,
- A gradient descent (GD) **method** for learning w/ AUC constraints,
- A new, restrictive type of **constraint**: ROC-based constraints,
- **Guarantees** and a **GD method** for learning with ROC constraints.

A general AUC constraint

Intra-group pairwise and BNSP AUC fairness ([Borkan et al., 2019]):

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

and many many more...

Introduce all relevant distributions as $D(s) = (H_s^{(0)}, H_s^{(1)}, G_s^{(0)}, G_s^{(1)})$.
Any known AUC constraint writes as:

$$\mathrm{AUC}_{\alpha^\top D(s), \beta^\top D(s)} = \mathrm{AUC}_{\alpha'^\top D(s), \beta'^\top D(s)},$$

with $\alpha, \alpha', \beta, \beta' \in [0, 1]^4$ and any of those sums to 1.

Learning with AUC constraints

Let $\mathcal{S}$ be a proposal family of scores. With an example constraint, integrating the constraint as a penalty gives, where $\lambda > 0$ is fixed:

$$\max_{s \in \mathcal{S}} L_\lambda(s) \quad \text{with} \quad L_\lambda(s) = \mathrm{AUC}_{H_s, G_s} - \lambda \left| \mathrm{AUC}_{H_s^{(0)}, G_s^{(0)}} - \mathrm{AUC}_{H_s^{(1)}, G_s^{(1)}} \right|,$$

and its solution is written s_λ^*.

Theorem 1

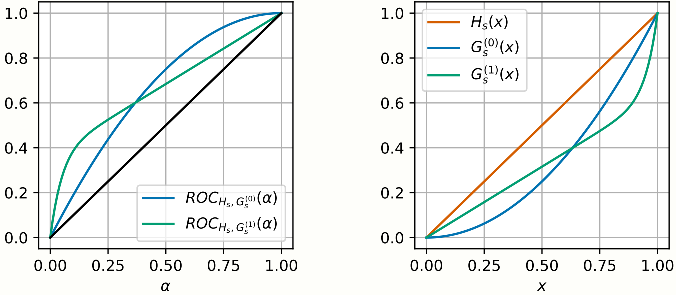
Assume that $\mathcal{S}$ is VC-major with VC-dim $V < +\infty$, and there exists $\epsilon > 0$, $\epsilon \leq \mathbb{P}\{Y = y, Z = z\}$ for any $y \in \mathcal{Y}, z \in \mathcal{Z}$. Then, for any $\delta > 0$ and $n > 1$, with probability $\geq 1 - \delta$:

$$\epsilon^2 \left[L_\lambda(s_\lambda^*) - L_\lambda(\widehat{s}_\lambda) \right] \leq C \sqrt{\frac{V}{n}} + (8\lambda + 2\epsilon) \sqrt{\frac{\log(13/\delta)}{n-1}} + O(n^{-1}).$$

To learn with **gradient descent:**
We relax $\widehat{L}_\lambda$ by replacing $x \mapsto \mathbb{I}\{x \geq 0\}$ by a sigmoid function σ and replace the abs $|\cdot|$ by a multiplication by $c \in [-1, +1]$.

Limitations of AUC constraints

In the example below, with $s \in [0,1]$, an AUC constraint is verified.
However, $\sup_{t\in[0,1]} |G_s^{(0)}(t) - G_s^{(1)}(t)| \approx 0.10$.



Let h, g, h', g' cdfs on $\mathbb{R}$ s.t. $\mathrm{ROC}_{h,g}$ and $\mathrm{ROC}_{h',g'}$ are continuous.
If $\mathrm{AUC}_{h,g} = \mathrm{AUC}_{h',g'}$, $\exists \alpha \in (0,1)$ s.t. $g \circ h^{-1}(\alpha) = g' \circ h'^{-1}(\alpha)$.

Conclusion: An AUC constraint imposes a “pointwise constraint”.

Learning with pointwise constraints

To measure the difference between cdfs for $Z = 0$ and $Z = 1$, let:

$$\Delta_{H,\alpha}(s) = \mathrm{ROC}_{H_s^{(0)},H_s^{(1)}}(\alpha) - \alpha \quad \text{and} \quad \Delta_{G,\alpha}(s) = \mathrm{ROC}_{G_s^{(0)},G_s^{(1)}}(\alpha) - \alpha.$$

We introduce a sum of m_H pointwise constraints for $\Delta_{H,\cdot}$ and m_G for $\Delta_{G,\cdot}$ as a penalization, and maximize L_Λ in $\mathcal{S}$, where:

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

which gives the score s_Λ^*.

The empirical counterpart of L_Λ is $\widehat{L}_\Lambda$, its maximizer is $\widehat{s}_\Lambda$.

Theorem 2

Assume that $\exists M, \kappa > 0$ s.t. $M \leq D_k'(s) \leq M \cdot \kappa$ for all $k \in [\![1, 4]\!], s \in \mathcal{S}$. Under the assumptions of Theorem 1,

$$\epsilon^2 \cdot [L_\Lambda(s_\Lambda^*) - L_\Lambda(\widehat{s}_\Lambda)] \leq C_{\lambda,\epsilon,\kappa}\sqrt{\frac{V}{n}} + C'_{\Lambda,\epsilon,\kappa}\sqrt{\frac{\log(19/\delta)}{n-1}} + O(n^{-1}).$$

Outline

The COMPAS database

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions.
$\rightarrow$ recidivism prediction.

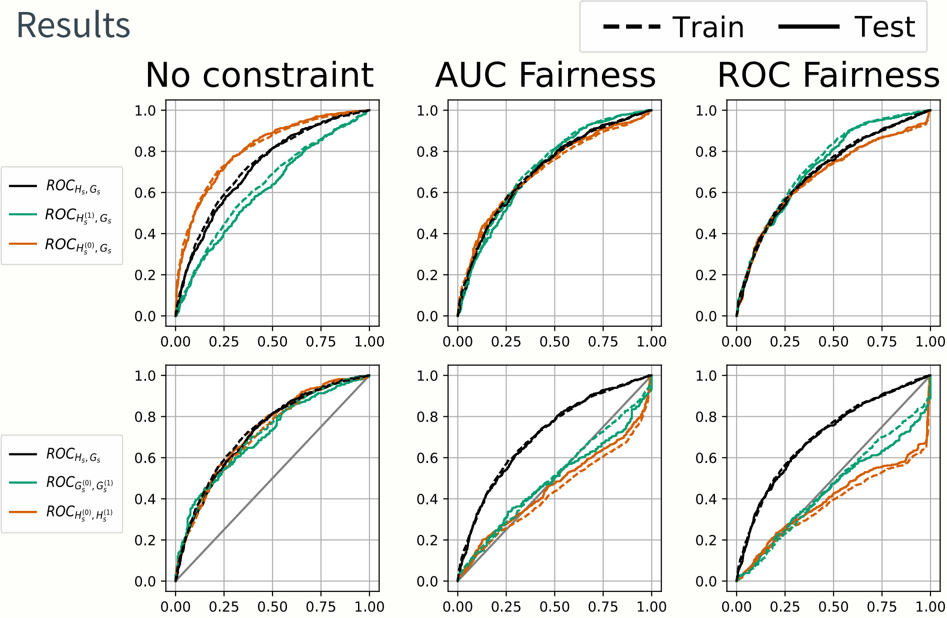
Problem distributions:

	Non-recidivist ($Y = -1$)	Recidivist ($Y = 1$)
Other ($Z = 0$)	$H_s^{(0)}$	$G_s^{(0)}$
African-american ($Z = 1$)	$H_s^{(1)}$	$G_s^{(1)}$

We run:
$\rightarrow$ Baseline: optimize the ranking performance,
$\rightarrow$ Ranking and AUC -based constraint,
$\rightarrow$ Ranking and ROC -based constraint.

[Vogel et al., 2020] confirms our results on 4 different tabular DBs.

Results



AUC cons. $\rightarrow$ top col. fig. / ROC cons: $|\Delta_{H,1/8}(s)|$, $|\Delta_{H,1/4}(s)|$, $|\Delta_{G,1/8}(s)|$, $|\Delta_{G,1/4}(s)|$.

Outline

Conclusion

Extension to similarity learning:

We proposed ways to balance a **score function** for sensitive groups, using **explicit fairness constraints**

Similarity learning in biometrics is **scoring on a product space**, see [Vogel et al., 2018].

Hence, one can build on [Vogel et al., 2020] to **tackle bias in FR**.

Open questions:

Loss function for similarity learning,

Generalization to many sensitive classes, *i.e.* $Z \in \mathbb{N}$.

Empirical performance for FR tasks.

References I

- Barocas, S., Hardt, M., and Narayanan, A. (2019).
  Fairness and Machine Learning.
  fairmlbook.org.
  `http://www.fairmlbook.org`.
- 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 *KDD*.
- Biega, A. J., Gummadi, K. P., and Weikum, G. (2018).
  Equity of attention: Amortizing individual fairness in rankings.
  In *SIGIR*.
- Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019).
  Nuanced metrics for measuring unintended bias with real data for text classification.
  arXiv:1903.04561.
- Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018).
  Empirical risk minimization under fairness constraints.
  In *NeurIPS*.
- Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. (2018).
  Learning adversarially fair and transferable representations.
  ICML, abs/1802.06309.

References II

Singh, A. and Joachims, T. (2018).
Fairness of exposure in rankings.
In *KDD*.

Singh, A. and Joachims, T. (2019).
Policy learning for fairness in ranking.
In *NeurIPS*.

Vogel, R., Bellet, A., and Clémençon, S. (2018).
A probabilistic theory of supervised similarity learning for pointwise ROC curve optimization.
In *ICML*. PMLR.

Vogel, R., Bellet, A., and Clémençon, S. (2020).
Learning fair scoring functions: Fairness definitions, algorithms and generalization bounds for bipartite ranking.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019).
Fairness constraints: A flexible approach for fair classification.
Journal of Machine Learning Research, 20(75):1–42.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017).
FA*IR: A Fair Top-k Ranking Algorithm.
In *CIKM*.