Developing Fair ML Models from Theoretical Aspects

Yijun Bian

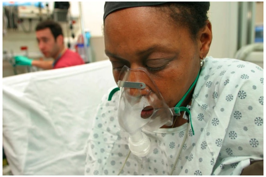
Machine Learning Section
Department of Computer Science
University of Copenhagen

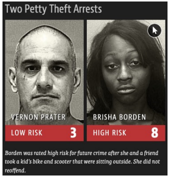
August 2024

My research

1. **Yijun Bian**[⋆], Kun Zhang, Anqi Qiu, and Nanguang Chen. ”Increasing fairness via combination with learning guarantees”. In: arXiv preprint arXiv:2301.10813 (2023). In Revision.
2. **Yijun Bian**[#⋆] and Yujie Luo[#]. ”Does Machine Bring in Extra Bias in Learning? Approximating Fairness in Models Promptly”. In: arXiv preprint arXiv:2405.09251 (2024). In Revision.
3. **Yijun Bian**[#⋆], Yujie Luo[#⋆], and Ping Xu. ”Approximating Discrimination within Models When Faced With Several Non-Binary Sensitive Attributes”. (2024). Under Review.

Examples [1,2,3]



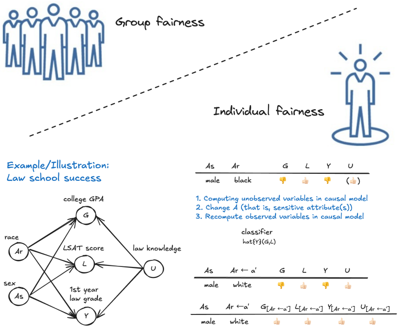


[1] AI detectors were more likely to flag writing by international students (i.e., non-native speakers) as AI-generated (Weixin Liang et al. “GPT detectors are biased against non-native English writers”. In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023)

[2] When people of color have complex medical needs, they are less likely to be referred to programmes that provide more individualised care (Linda Nordling. “A fairer way forward for AI in health care”. In: *Nature* 573.7775 [2019], S103–S103)

[3] Black defendants were mislabelled as high risk more often than white defendants (Lorenzo Belenguer. “AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry”. In: *AI and Ethics* 2.4 [2022], pp. 771–787)

- Fairness estimation based on a finite sample
- Insufficient data
- Fairness measures/metrics [4]
- The trade-off between fairness and accuracy

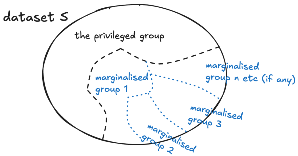


Our motivation

[4] *Group fairness* focuses on statistical/demographic equality among groups defined by sensitive attributes, while *individual fairness* follows a principle that "similar individuals should be evaluated or treated similarly."

Appendix

Our current research up to the present



Fairness metric/measure

Discriminative risk (DR)

Pro: it works for several sensitive attributes (SAs) with multiple values

Increasing fairness via combination with learning guarantees

Harmonic fairness via manifolds (HFM)

built upon a concept of distance between sets

HFM for one SA with binary values

Distance between sets for one SA with binary values

HFM for one/several SA(s) with multiple values

Distance between sets for one SA with multiple values
Distance between sets for several SAs with multiple values

maximal / average HFM

Does machine bring in extra bias in learning?
Approximating discrimination within models quickly

Discriminative risk (DR) —from an individual aspect

Following the principle of individual fairness, the fairness quality of one hypothesis[5] $f(\cdot)$ could be evaluated by

sensitive attribute(s)

general attributes

the indicator function

sensitive attribute(s) that are slightly disturbed

$$
\begin{aligned}
\ell_{\text{fair}}(f, \boldsymbol{x}) &= \mathbb{I}\big(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big) \\
&= \mathbb{I}\big(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big)\,, \qquad (1)
\end{aligned}
$$

model prediction on the raw instance

model prediction when only sensitive attribute(s) are changed

similarly to the 0/1 loss. Note that Eq. (1) is evaluated on only one instance with sensitive attributes $\boldsymbol{x}$.

[5]The hypothesis used in this equation could indicate an individual classifier or an ensemble classifier.

Discriminative risk (DR) —from a group aspect

To describe this characteristic of the hypothesis on multiple instances (aka. from a group level), then the empirical discriminative risk on one dataset S is expressed as

$$\hat{\mathcal{L}}_{\text{fair}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell_{\text{fair}}(f,\boldsymbol{x}_i)}_{\text{discriminative risk of } f(\cdot) \text{ on one instance}} \,, \tag{2}$$

and the true discriminative risk[6] of the hypothesis over a data distribution is

$$\mathcal{L}_{\text{fair}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\, \underbrace{\ell_{\text{fair}}(f,\boldsymbol{x})}_{\text{discriminative risk of } f(\cdot) \text{ on one instance}} \,] \,, \tag{3}$$

respectively.

[6]The instances from S are independent identically distributed (i.i.d.) drawn from an input/feature-output/label space $\mathcal{X}\times\mathcal{Y}$ according to an unknown distribution $\mathcal{D}$.

Cancellation-of-bias effect[7] in ensemble combination

- Inspired by existing work for error rates and oracle bounds
- First- and second-order oracle bounds concerning fairness
- Similarly to the *cancellation-of-errors* effect in ensemble combination

The DR of one ensemble can be bounded by a constant times the DR of one individual classifiers

[7] Yijun Bian et al. "Increasing fairness via combination with learning guarantees". In: *arXiv preprint arXiv:2301.10813* (2023).