UNIVERSITY OF TORONTO

Why Is My Classifier Discriminatory?

Irene Y. Chen, Fredrik D. Johansson, David Sontag

Presented by: Rui (Patrick) Xian, Rafael (Alex) Valencia
Mathematical Statistics I (STA2112) 12/2022

Outline

- Introduction and Motivation (Xian)
- Methods (Valencia)
- Results (Xian)
- Conclusions (Valencia)

Introduction – origin of discrimination in machine learning

$\{\mathbf{X}, \mathbf{y}\}$ setting:

Given features $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}\}$ and label $\mathbf{y}$

Find $\hat{Y} : \mathbf{X} \rightarrow \mathbf{y}$ that optimizes L

$\{\mathbf{X}, \mathbf{A}, \mathbf{y}\}$ setting:

What if $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}, \mathbf{A}\}$?

Does $\hat{Y} : \mathbf{X} \rightarrow \mathbf{y}$ that optimizes L also balances all $\hat{Y}|\mathbf{A}$'s?

* Not all features are created equal
* Sensitive attributes exist in the feature set
  - sex/gender
  - race/ethnicity
  - orientation
  - age group
  - socioeconomic status
  - religious belief
  - political leaning
  - ...

Introduction – origin of discrimination in machine learning

$\{\mathbf{X}, \mathbf{y}\}$ setting:

Given features $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}\}$ and label $\mathbf{y}$

Find $\hat{Y} : \mathbf{X} \rightarrow \mathbf{y}$ that optimizes L

$\{\mathbf{X}, \mathbf{A}, \mathbf{y}\}$ setting:

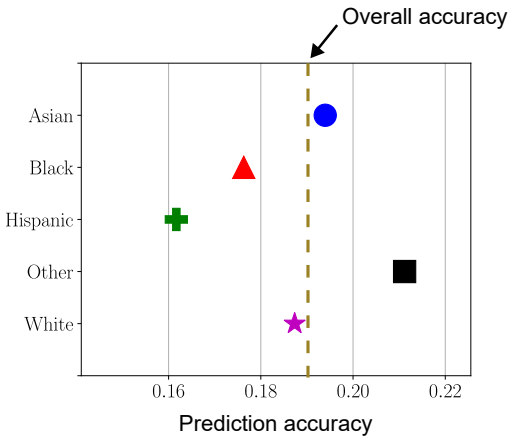
What if $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}, \mathbf{A}\}$?

Does $\hat{Y} : \mathbf{X} \rightarrow \mathbf{y}$ that optimizes L also balances all $\hat{Y}|\mathbf{A}$'s?

Mostly not! We need fair classifiers!

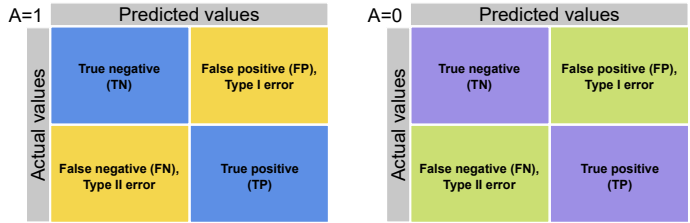
* Not all features are created equal
* Sensitive attributes exist in the feature set
  - sex/gender
  - race/ethnicity
  - orientation
  - age group
  - socioeconomic status
  - religious belief
  - political leaning
  - ...

Introduction – origin of discrimination in machine learning



* Not all features are created equal
* Sensitive attributes exist in the feature set
* Overall accuracy is optimized during training and testing
* Disaggregated evaluation of prediction models on subgroups shows discrepancies in accuracy
* Depends on model, dataset, training procedure, metrics/world views, etc.

Introduction – definition of group fairness



Equalized odds / statistical parity

$$\left| P(\hat{Y} = 1 | A = 1, Y = 0) - P(\hat{Y} = 1 | A = 0, Y = 0) \right| \leqslant \epsilon$$

$$\left| P(\hat{Y} = 1 | A = 1, Y = 1) - P(\hat{Y} = 1 | A = 0, Y = 1) \right| \leqslant \epsilon$$

Motivation and timeline

pre-2016
Dwork, Zemel, Hardt, Scheidegger, Venkata-subramanian, et al.

Definitions: group fairness, individual fairness, etc.

2016
Angwin et al.

Inherent fairness vs accuracy tradeoff

2017
Kleinberg, Mullainathan, Raghavan

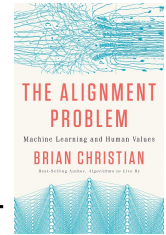
High-d models, intersectionality

2018
Buolamwini, Gebru

2019

OECD.AI Policy Observatory

> *AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.*
> *To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.*



2020
Christian

Motivation and timeline

pre-2016
Dwork, Zemel, Hardt, Scheidegger, Venkata-subramanian, et al.

Definitions: group fairness, individual fairness, etc.

2016
Angwin et al.

2017
Kleinberg, Mullainathan, Raghavan

Inherent fairness vs accuracy tradeoff

2018
Buolamwini, Gebru

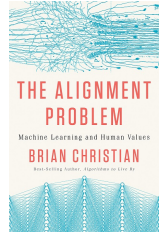
High-d models, intersectionality

Chen, Johansson, Sontag

2019

OECD.AI Policy Observatory

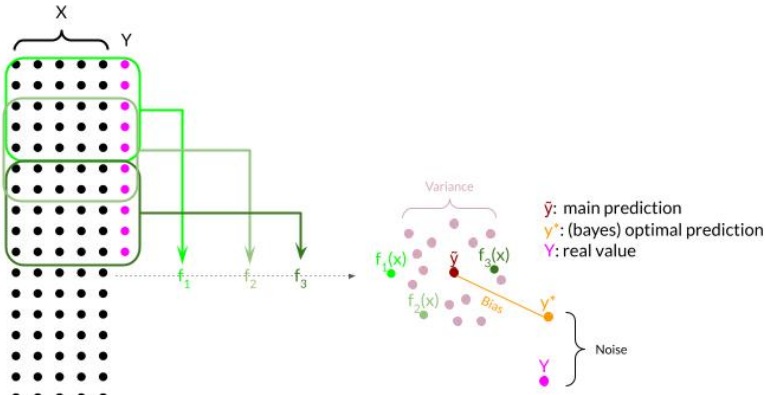
> *AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.*
> *To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.*



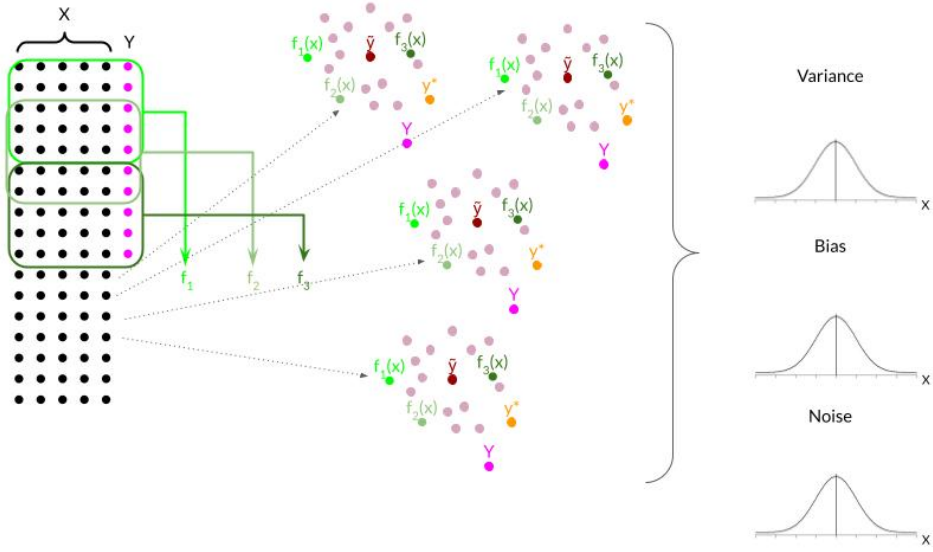
2020
Christian

UNIVERSITY OF TORONTO

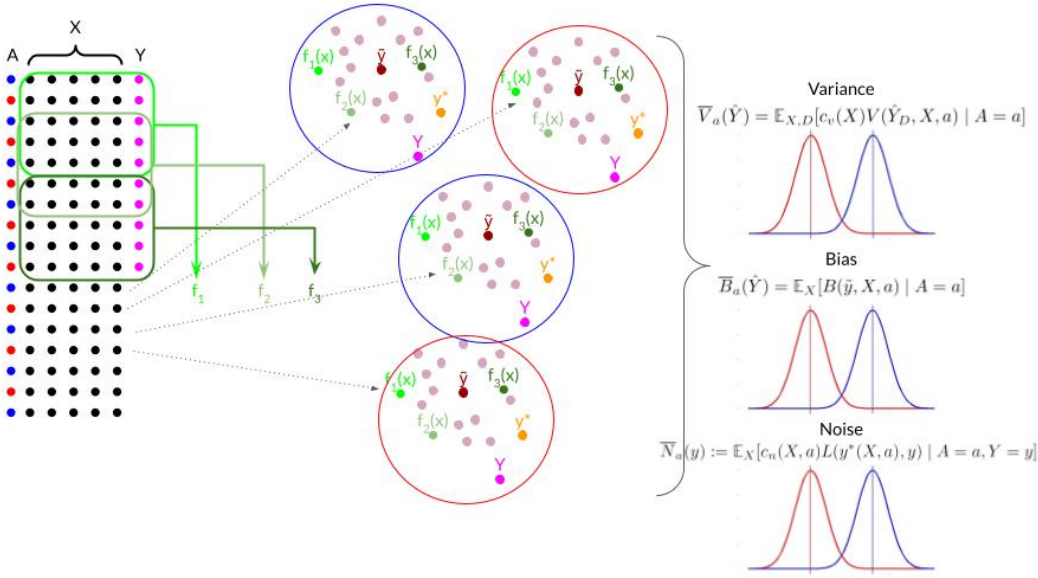
Methods - Setting up the framework



Methods - Setting up the framework



Methods - Setting up the framework



Methods - Defining discrimination

The measurement of discrimination is based on cost functions $\gamma_a \in \{FPR, FNR, ZO, MSE\}$:

- False Positive Rate $\mathbf{FPR}_a(\hat{Y}) := \mathbb{E}_X[\hat{Y} \mid Y = 0, A = a]$
- False Negative Rate $\mathbf{FNR}_a(\hat{Y}) := \mathbb{E}_X[1 - \hat{Y} \mid Y = 1, A = a]$
- Zero-One loss $\mathbf{ZO}_a(\hat{Y}) := \mathbb{E}_X[\mathbb{1}[\hat{Y} \neq Y] \mid A = a]$
- Mean-squared errors $\mathbf{MSE}_a(\hat{Y}) := \mathbb{E}_X[(\hat{Y} - Y)^2 \mid A = a]$

Expected level of discrimination

A model is fair if the expected level of discrimination is low.

$$\bar{\Gamma}(\hat{Y}) := \left| \mathbb{E}_D \left[\gamma_0(\hat{Y}_D) \right] - \mathbb{E}_D \left[\gamma_1(\hat{Y}_D) \right] \right|$$

Methods - Decomposing discrimination

Components of the discrimination

For the cost functions listed above (**ZO**,**FPR**,**FNR**,**MSE**); the expected cost $\bar{\gamma}_a$ and the expected discrimination level $\bar{\Gamma}$ admit a decomposition of the form:

$$\bar{\gamma}_a(\hat{Y}) = \underbrace{\overline{N}_a}_{\text{Noise}} + \underbrace{\overline{B}_a(\hat{Y})}_{\text{Bias}} + \underbrace{\overline{V}_a(\hat{Y})}_{\text{Variance}}$$

$$\bar{\Gamma}(\hat{Y}) = \left|(\overline{N}_0 - \overline{N}_1) + (\overline{B}_0(\hat{Y}) - \overline{B}_1(\hat{Y})) + (\overline{V}_0(\hat{Y}) - \overline{V}_1(\hat{Y}))\right|$$

Methods - Sources of discrimination

- Bias $\overline{B}_0(\hat{Y}) - \overline{B}_1(\hat{Y})$
  Model is better for one group and is not flexible enough for both.
- Variance $\overline{V}_0(\hat{Y}) - \overline{V}_1(\hat{Y})$
  Difference in sample sizes or variance of features is dependent on the protected attribute $V(X|A)$ and possible overfitting
- Noise $\overline{N}_0 - \overline{N}_1$
  Discrimination is partially unrelated to model or training set size and the conjecture is that it can only be reduced with additional features.

Methods - Reducing discrimination through data collection

- Increasing training set size

Assumption: Learning curves

Considering a training set of size n, the expected costs $\bar{\gamma}(\hat{Y}, n)$, $\bar{\gamma}_0(\hat{Y}, n)$, $\bar{\gamma}_1(\hat{Y}, n)$, behave asymptotically as inverse power-law curves with parameters (α, β, δ).

$$\bar{\gamma}(\hat{Y}, n) = \alpha n^{-\beta} + \delta \quad \text{and} \quad \forall a \in \mathcal{A} : \bar{\gamma}_a(\hat{Y}, n_a) = \alpha_a n_a^{-\beta_a} + \delta_a$$

- Measuring additional variables

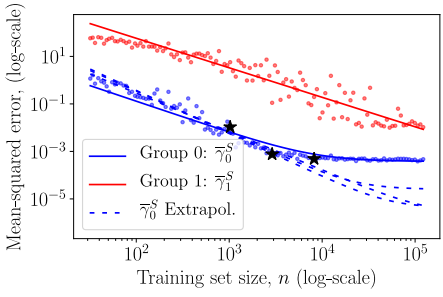
Clustering w.r.t expected prediction cost

Defining $\rho_a^{\gamma}(c) := \mathbb{E}_X[\gamma(\hat{Y}) \mid A = a, C = c]$. Clusters c for which $|\rho_0(c) - \rho_1(c)|$ is large identify groups of individuals where discrimination is above average and could serve as an indicator for collection additional information.

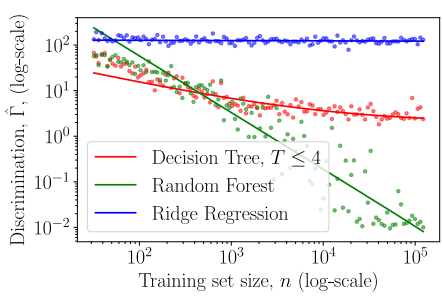
Results – offsetting discrimination in simulation

Simulation setting: $X \sim \mathcal{N}(\mu_A, \sigma_A^2)$, $Y = 2X^2 - 2X + 0.1 + \epsilon X^2$, $\epsilon \sim \mathcal{N}(0, 1)$
Model: various

a



b

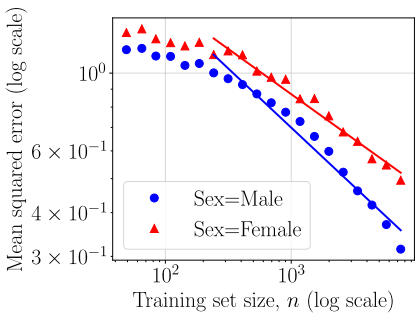


Results – book rating prediction

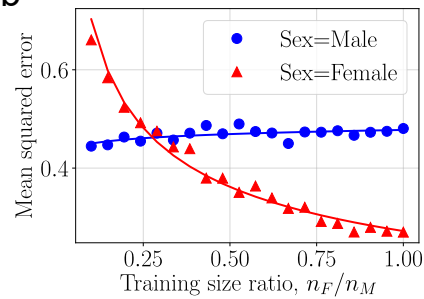
Problem setting: $\mathbf{X}$ = text embedding from online book reviews (Goodreads),
$\mathbf{y}$ = book rating, $\mathbf{A}$ = sex/gender of the author

Model: random forest

a



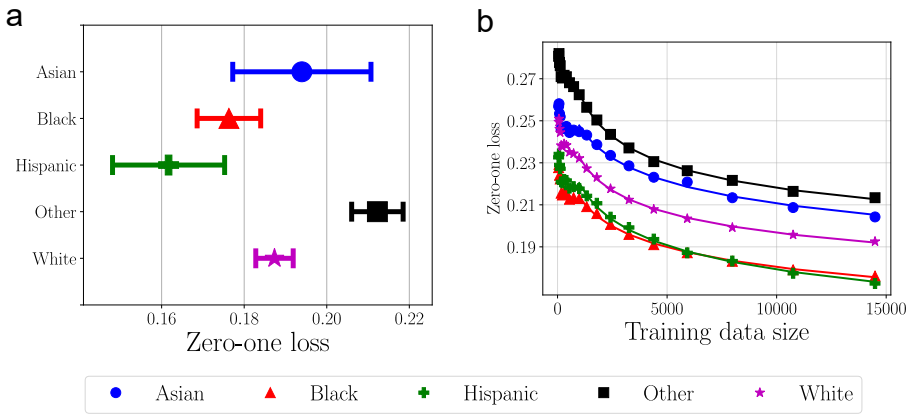
b



Results – intensity care unit mortality prediction

Problem setting: $\mathbf{X}$ = text embedding from clinical notes (MIMIC-III),
$\mathbf{y}$ = mortality after 48 hours, $\mathbf{A}$ = race of the patient

Model: logistic regression



Conclusions

- Discrimination can be decomposed into Bias, Variance and Noise.
- Each component calls for different actions to fix it.
- The best way to decrease discrimination without affecting accuracy is to get more information.
- Why Is My Classifier Discriminatory? .... because of your data and assumptions!

Thank you for listening!
Questions?