Why is My Classifier Discriminatory?

Irene Y. Chen, Fredrik D. Johansson, David Sontag
Massachusetts Institute of Technology

Introduction

- It is surprisingly easy to build a discriminatory algorithm, even unintentionally. Why? What might be **sources of unfairness**?
- We **decompose fairness metrics** into bias, variance, and noise to guide actions to reducing each error.
- In three experiments, we demonstrate the effect of **increasing data size** on fairness and **identifying subpopulations** for increasing feature size.

Background

Below we define our variables.

D	data	$\bar{\Gamma}$	unfairness
Y	outcome	y^*	Bayes optimal classifier
$\hat{Y}$	prediction	$\tilde{y}$	majority classifier
X	covariates	B_a	bias of group *a*
A	protected group	V_a	variance of group *a*
γ	loss (e.g. 0-1 error, false pos, false neg)	N_a	noise of group *a*

Prior work has focused on improving fairness through the model, but the training data is also important.

Model considerations
- Regularization [1,2]
- Loss function constraints [3]
- Representation learning [4]
- Post-hoc corrections [5]

Data considerations
- Processing [6]
- Cohort selection
- Sample size
- Number of features

References

[1] Bechavod, Yahav and Ligett, Katrina. Learning fair classifiers: A regularization-inspired approach. arXiv preprint arXiv:1707.00044, 2017.
[2] Kamishima, Toshihiro, Akaho, Shotaro, and Sakuma, Jun. Fairness-aware learning through regularization approach. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 643–650. IEEE, 2011.
[3] Kamiran, Faisal, Calders, Toon, and Pechenizkiy, Mykola. Discrimination aware decision tree learning. In Data Mining (ICDM), 2010 IEEE 10th International Conference on, pp. 869–874. IEEE, 2010.
[4] Zemel, Richard S, Wu, Yu, Swersky, Kevin, Pitassi, Toniann, and Dwork, Cynthia. Learning fair representations. ICML (3), 28:325–333, 2013.
[5] Hardt, Moritz, Price, Eric, Srebro, Nati, et al. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, pp. 3315–3323, 2016.
[6] Hajian, Sara and Domingo-Ferrer, Josep. A methodology for direct and indirect discrimination prevention in data mining. IEEE transactions on knowledge and data engineering, 25(7):1445– 1459, 2013.
[7] Domingos, Pedro. A unified bias-variance decomposition. In Proceedings of 17th International Conference on Machine Learning, pp. 231–238, 2000.

Methodology

We consider fairness through loss definitions.

One example of loss is false positive loss.

$$\gamma_a(\hat{Y}, Y, D) := P_D(\hat{Y} = 1 \mid A = a, Y = 0)$$

We define unfairness as the difference between losses.

$$\Gamma := |\gamma_1 - \gamma_0|$$

Theorem 1: We can decompose both loss $\bar{\gamma}_a$ and unfairness Γ.

$$\bar{\gamma}_a(\hat{Y}) = \bar{B}_a(\hat{Y}) + \bar{V}_a(\hat{Y}) + \bar{N}_a$$

$$\bar{\Gamma} = |\underbrace{(\bar{B}_1 - \bar{B}_0)}_{\text{diff in bias}} + \underbrace{(\bar{V}_1 - \bar{V}_0)}_{\text{diff in variance}} + \underbrace{(\bar{N}_1 - \bar{N}_0)}_{\text{diff in noise}}|$$

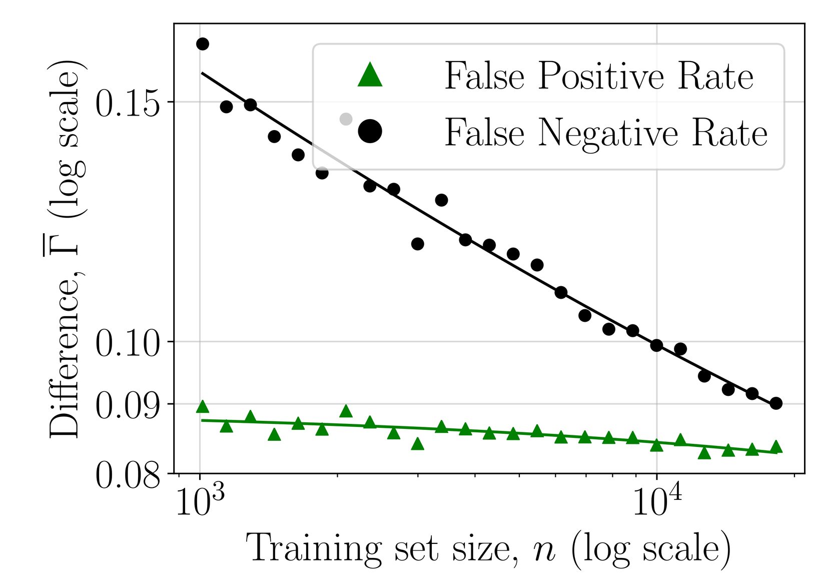
Proposition 1: If $\bar{N}_0 \neq \bar{N}_1$, no model can be 0-discriminatory in expectation.

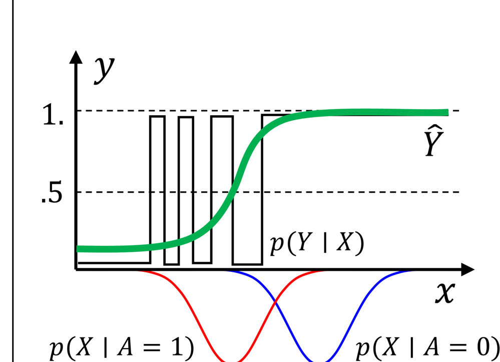
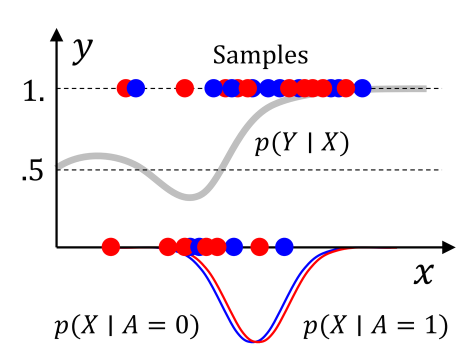
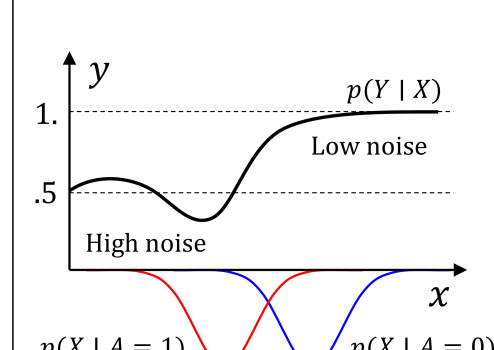
Experiments

Dataset	Task	Protected Group
Census income (UCI)	Predict over/under 50k	Gender
Clinical notes (MIMIC-III)	Predict hospital mortality	Race
Book reviews (Goodreads)	Predict book review score	Author gender

Experiment 1: Income Prediction

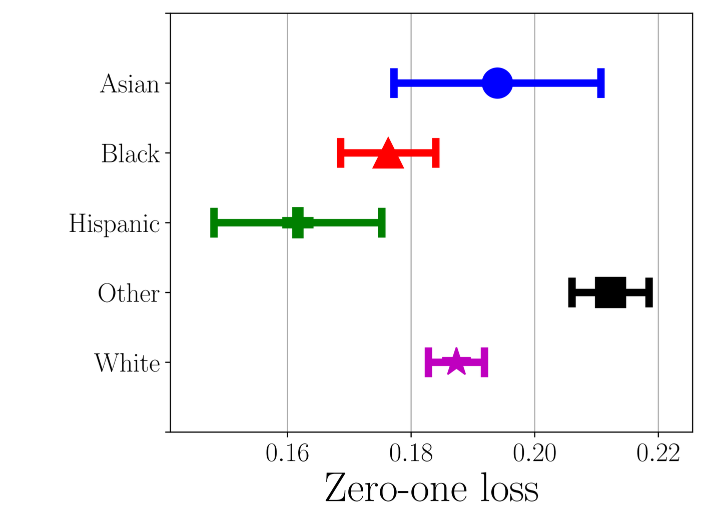
Differences in false positive rate and false negative rate decrease as we add more training data.



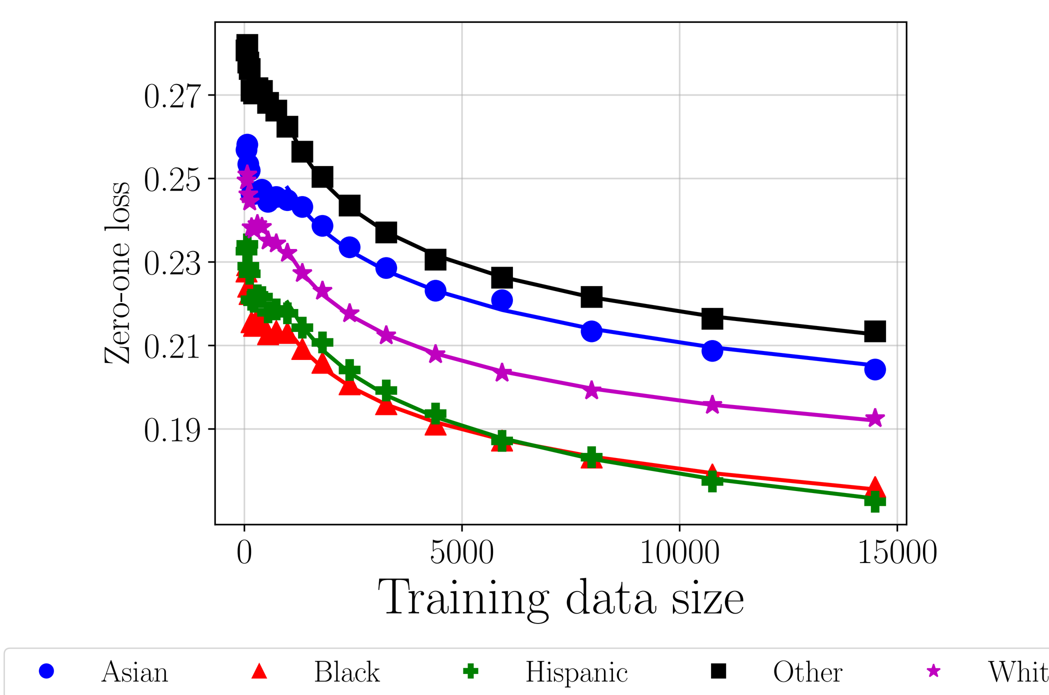
	Description	Definition [7] and impact	Illustrative plot	How to detect	How to fix
Bias	How well the model fits the data	$B_a(\hat{Y}, x, a) = L(y^*(x,a), \tilde{y}(x,a))$ One choice of model may be better suited for one group, causing differences in expected bias.		Experiment with model complexity	Change model class
Variance	How much sample size affects accuracy	$V_a(\hat{Y}, x, a) = E_D[L(\tilde{y}(x,a), \hat{y}_D(x,a))]$ For identically distributed groups, bias and noise are equal in expectation. Perceived discrimination is only from variance.		Fit inverse power laws from subsampling	Increase training data size
Noise	Irreducible error independent of sample size and model	$N(x,a) = E_Y[L(y^*(x,a)) \mid X, A]$ Differences in noise between two groups may contribute to discrimination if protected groups are not identically distributed		Estimate Bayes error using distance metrics	Increase number of features

Experiment 2: ICU Mortality

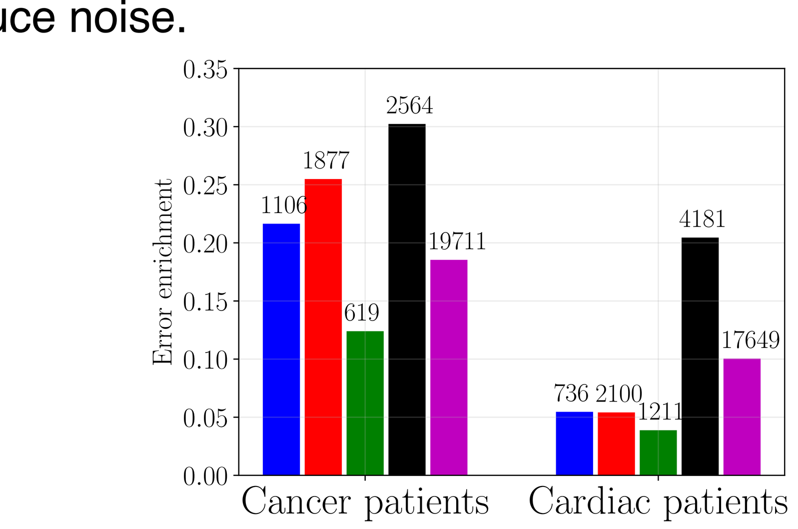
We find statistically significant racial differences for zero-one loss.



Subsampling training data follows inverse power laws. The infinite data limit reveals error with no variance, only noise and bias.



Topic modeling reveals subpopulations with high differences in error to guide feature augmentation, to reduce noise.



Conclusion

1. For **accurate and fair models** deployed for real world applications, both the model and the algorithm must be considered.
2. We provide **easily implementable fairness tools** to evaluate bias, variance, and noise in an algorithm, which can guide further efforts to reduce unfairness.