Why is My Classifier Discriminatory?

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

NeurIPS 2018
Spotlight Presentation

As narrated by Dany Haddad, Alan Gee

The Cost of Fairness

- Most research has suggested sacrificing model accuracy for the sake of fairness
- Often, sacrificing predictive accuracy is difficult to justify
- Almost too obvious: This work suggests additional data collection as a strategy to improve a model's fairness rather than constraining model

Where does unfairness come from?

Best to understand where the discrimination may originate from so a proper solution can be applied...but prior work* has focused mainly on models

Modeling Considerations	Data Considerations
Loss function constraints*	Pre-processing*
Modeling the Data*	Population/Group Diversity
Regularization*	Feature Selection
Trade-offs*	Sample Size

*See paper for references

Notation

$A = a$, is the protected attribute

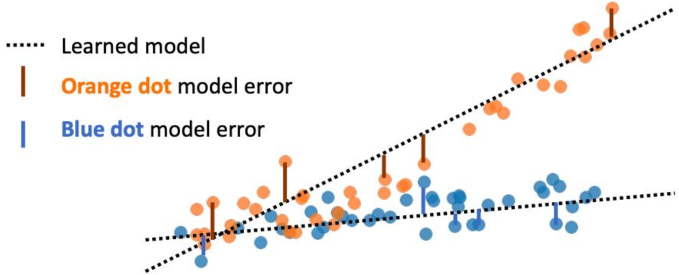
$\hat{Y}_d := h(X, A)$ are predictions learned from dataset d

$\bar{\cdot} := E_D[\cdot]$

main prediction $\tilde{y}(x, a) = \arg\min_{y'} \mathbb{E}_D[L(\hat{Y}_D, y') \mid X = x, A = a]$

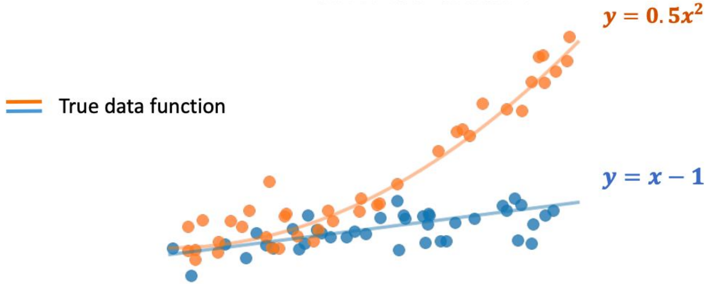
(Bayes) optimal prediction $y^*(x, a) = \arg\min_{y'} \mathbb{E}_Y[L(Y, y') \mid X = x, A = a]$

Decomposition of Error



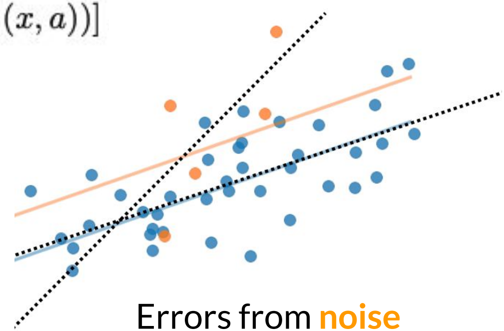
Errors from variance

$$V(\hat{Y}, x, a) = \mathbb{E}_D[L(\tilde{y}(x,a), \hat{y}_D(x,a))]$$



Errors from bias

$$B(\hat{Y}, x, a) = L(y^*(x,a), \tilde{y}(x,a))$$



Errors from noise

$$N(x, a) = \mathbb{E}_Y[L(y^*(x,a), Y) \mid X = x, A = a]$$

Chen et al. NeurIPS 2018 Slides

Estimating Bias, Variance and Noise

$$V(\hat{Y}, x, a) = \mathbb{E}_D[L(\tilde{y}(x, a), \hat{y}_D(x, a))]$$

$$N(x, a) = \mathbb{E}_Y[L(y^*(x, a), Y) \mid X = x, A = a]$$

$$B(\hat{Y}, x, a) = L(y^*(x, a), \tilde{y}(x, a))$$

Definitions of Discrimination Level

$$\mathrm{FNR}_a(\hat{Y}) := \mathbb{E}_X[1 - \hat{Y} \mid Y = 1, A = a]$$

$$\mathrm{FPR}_a(\hat{Y}) := \mathbb{E}_X[\hat{Y} \mid Y = 0, A = a]$$

$$\mathrm{ZO}_a(\hat{Y}) := \mathbb{E}_X[\mathbb{1}[\hat{Y} \neq Y] \mid A = a]$$

$$\gamma_a \in \{\mathrm{ZO}, \mathrm{FPR}, \mathrm{FNR}\}$$

$$\Gamma^{\gamma}(\hat{Y}) := \left|\gamma_0(\hat{Y}) - \gamma_1(\hat{Y})\right|$$ Level of discrimination

Discrimination Level Decomposition

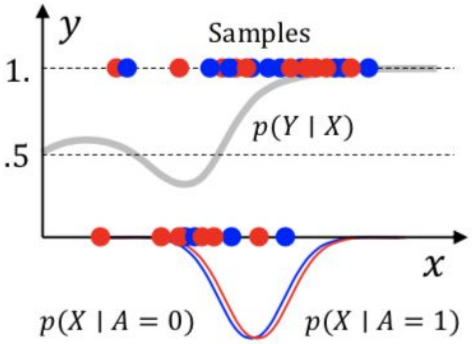
$$\overline{\gamma}_a(\hat{Y}) = \underbrace{\overline{N}_a}_{Noise} + \underbrace{\overline{B}_a(\hat{Y})}_{Bias} + \underbrace{\overline{V}_a(\hat{Y})}_{Variance}$$

The discrimination level decomposes as:

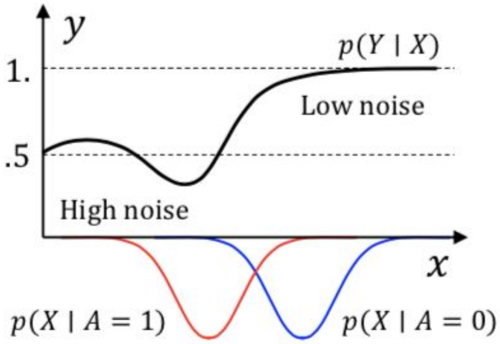
$$\overline{\Gamma} = \left|(\overline{N}_0 - \overline{N}_1) + (\overline{B}_0 - \overline{B}_1) + (\overline{V}_0 - \overline{V}_1)\right|$$

- Test for statistical significance of discrimination using a two-tailed z-test
    - The class specific error is approximately normally distributed for a large number of samples

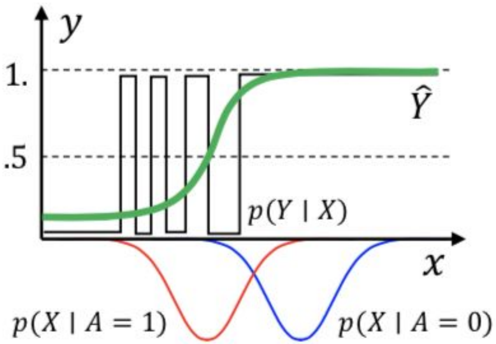
Error Decomposition and Discrimination



Groups are identically distributed wrt features, X. Discrimination is only due to **predictor variance.**



Groups are NOT identically distributed.
Difference in **noise** across values of X leads to discrimination.



One group may be harder to predict for than another. Errors due to **bias** will affect one group more than another.

Discrimination Level Decomposition

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

- The magnitude of each difference shows the sources of discrimination due to modeling error
- $(\overline{N}_0 - \overline{N}_1)$ Reduce by measuring additional features
- $(\overline{B}_0 - \overline{B}_1)$ Reduce by selecting a more appropriate model class
- $(\overline{V}_0 - \overline{V}_1)$ Reduce by increasing the training set size

Implications

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

- If the noise N_a differs across the protected attribute, a, then:
  - No classifier can have 0 discrimination, must have bias or variance larger than the Bayes optimal classifier
- Otherwise:
  - Noise is homoskedastic
  - Discrimination is only a result of the Bias and Variance terms

Mitigation of Discrimination Through Data

- Model performance as function of samples *n* behave like inverse power-law curves (a.k.a. *Type II learning curves*):

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

asymptotic bias and Bayes error

- Can be used to extrapolate discrimination learning curve:

$$\overline{\Gamma}(\hat{Y}, n) := |\overline{\gamma}_0(\hat{Y}, n) - \overline{\gamma}_1(\hat{Y}, n)|$$

Mitigation of Discrimination Through Data

- When discrimination $\overline{\Gamma}(\hat{Y}, n)$ is dominated by a difference in noise, $(\overline{N}_0 - \overline{N}_1)$ fairness may not be improved through model selection

- If the variance in outcomes within a cluster is not explained by the available feature set, additional variables may be used to further distinguish its members.

$$\rho_a^{\mathrm{ZO}}(c) := \mathbb{E}_X[\mathbb{1}[\hat{Y} \neq Y] \mid A = a, C = c],$$

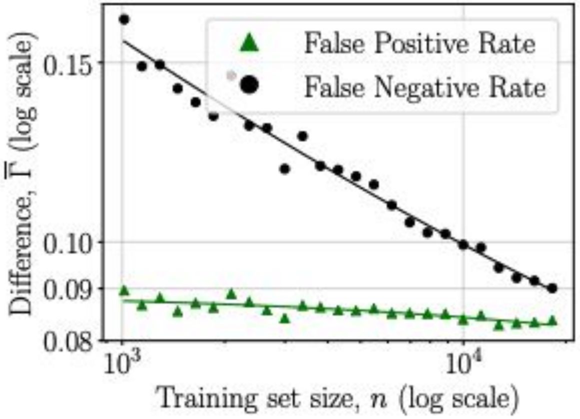
$$\Rightarrow \quad |\rho_0(c) - \rho_1(c)|$$

Experiments

Dataset	Objective	Protected Group
UCI's Census Income	Predict Income over/under 50k	Gender
MIMIC III's Clinical Notes	Predict Mortality	Race
Goodread's Book Reviews	Predict Review Score	Author Gender

- Analyze the level of discrimination for the full data
- Estimate the value of increasing training set size by fitting Type II learning curves
- Use clustering to identify subgroups where discrimination is high.

Experimental Results: Income Prediction

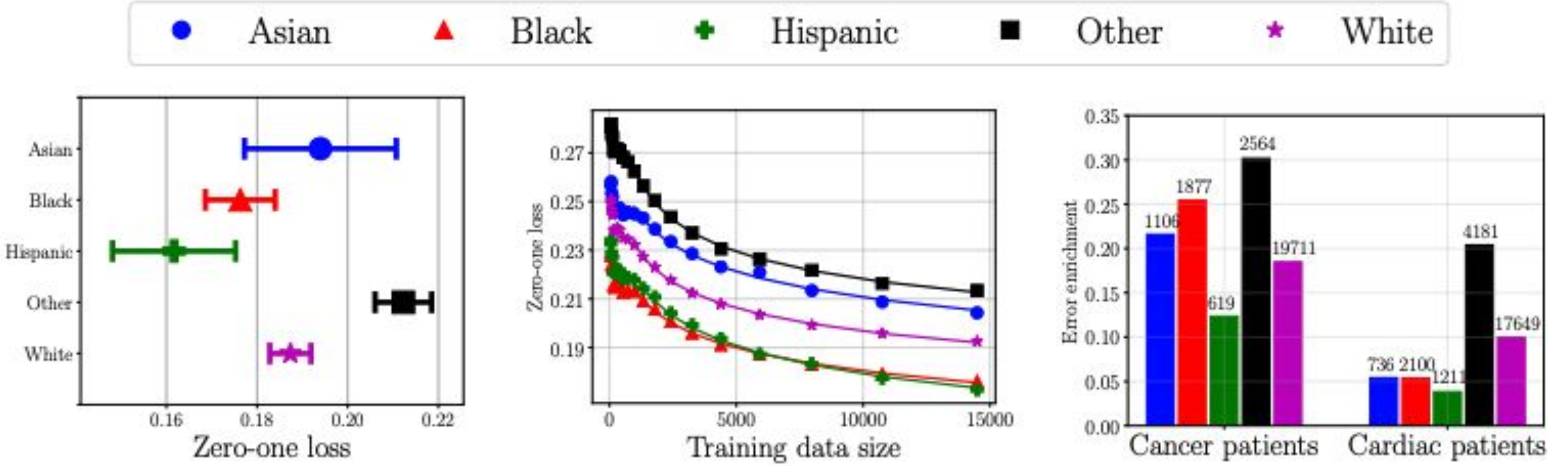


Key Takeaway: Differences in false negative rate (discrimination) decreases as training set size increases.

<u>Identifying Discrimination in Sub-groups</u>

- Income prediction at managerial level
  $FNR_{Women} = 0.412 > FNR_{men} = 0.157$
- For other positions:
  $FNR_{Women} = 0.543 > FNR_{men} = 0.461$

Experimental Results: Mortality Prediction



Statistically significant racial differences in zero-one loss

Shows benefit of fitting more data to reduce variance (discrimination levels decrease significantly)

Identify sub-groups were more features would help reduce noise (data-augmentation)

Paper Contributions

(1) Decompose unfairness into three categories: bias, variance, and noise.

(2) Show how to estimate these quantities.

(3) Experimentally show their methods can help identify subpopulations experiencing discrimination and suggest steps to counter the unfairness without sacrificing model accuracy.