Why is my classifier discriminatory?

Irene Chen, Fredrik Johansson, David Sontag - 2018

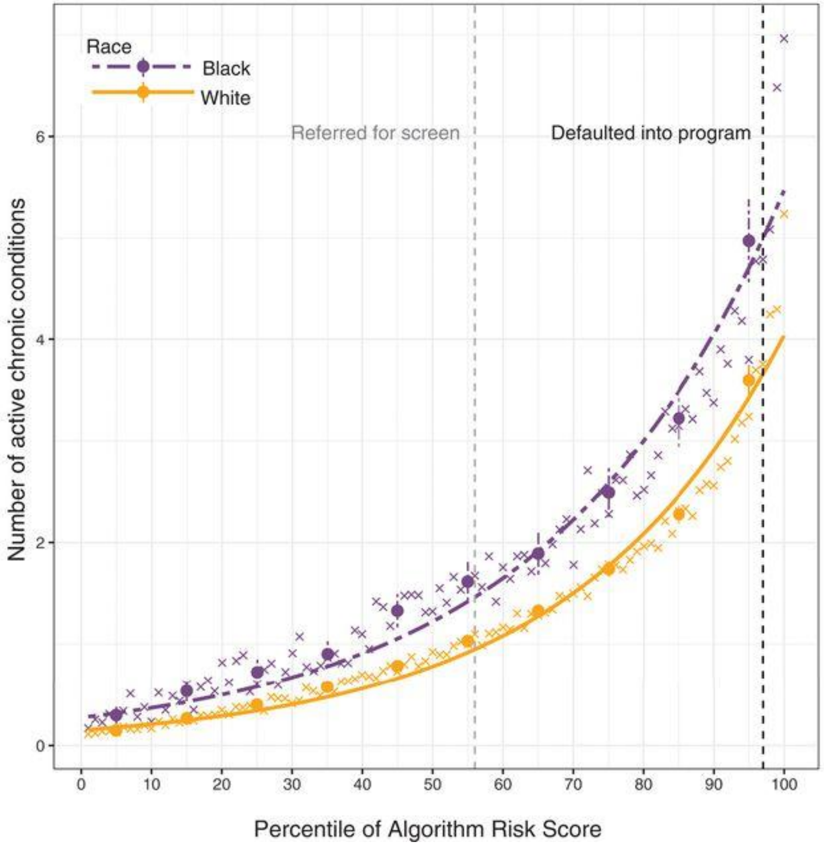
CSC2541HF - Caroline Malin-Mayor and Filip Miscevic

November 5, 2021

Rationale

- Classifiers can be biased towards protected groups (e.g. minorities)
  - Sample size
  - Class differences in noise
  - Model choice

Obermeyer et al., 2019.



Mathematical Definitions of Fairness

Equalized odds criterion: FPR and FNR equal across a binary class. For a given protected group a,

$$L_a(y, y') \in \begin{cases} FPR_a(\hat{Y}) := \mathbb{E}_X[\hat{Y} | Y = 0, A = a] \\ FNR_a(\hat{Y}) := \mathbb{E}_X[1 - \hat{Y} | Y = 1, A = a] \\ MSE = (y - y')^2, ZO = \mathbb{I}[y \neq y'] \end{cases}$$

Hard to measure/enforce in practice. Instead, measure *level of discrimination* on *loss functions:*

$$\Gamma^L := |L_0(\hat{Y}) - L_1(\hat{Y})|$$

Bias-Variance-Noise Decomposition of Discrimination

	Description
Bias	How well model fits data
Variance	How much sample size affects accuracy
Noise	Error independent of model class and sample size

Table from Irene Chen's Talk

Bias-Variance-Noise Decomposition of Discrimination

Define

$$\hat{y}_D = \text{predictor on training set D}$$
$$\tilde{y} = \text{average prediction over draws of training sets}$$
$$y^* = \text{Bayes optimal predictor (noise only)}$$

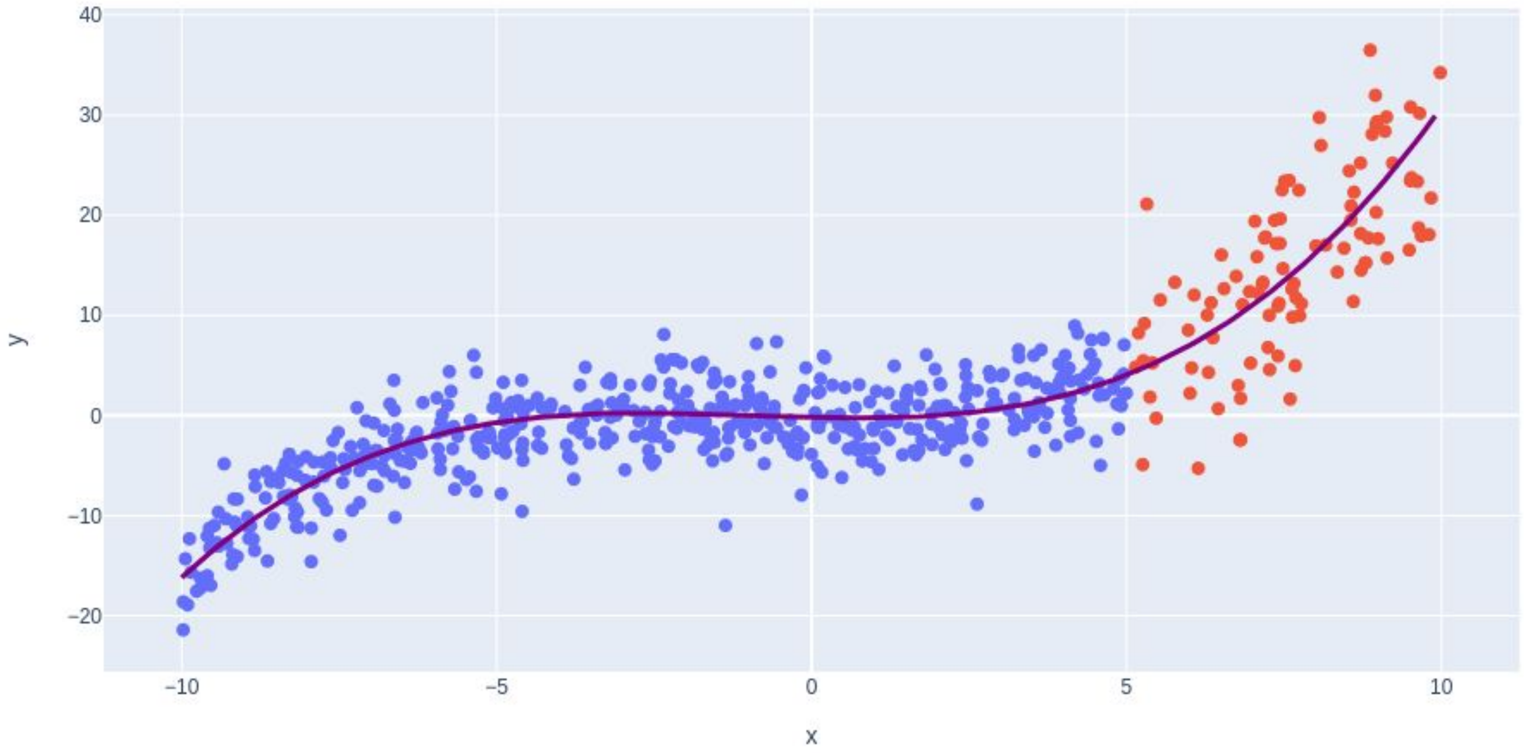
For a given class and training set D, define Bias, Variance and Noise as:

$$B(\hat{Y}) = L(y^*, \tilde{y}), V(\hat{Y}) = \mathbb{E}_D[L(\tilde{y}, \hat{y}_D)], N = \mathbb{E}_Y[L(y^*, Y)]$$

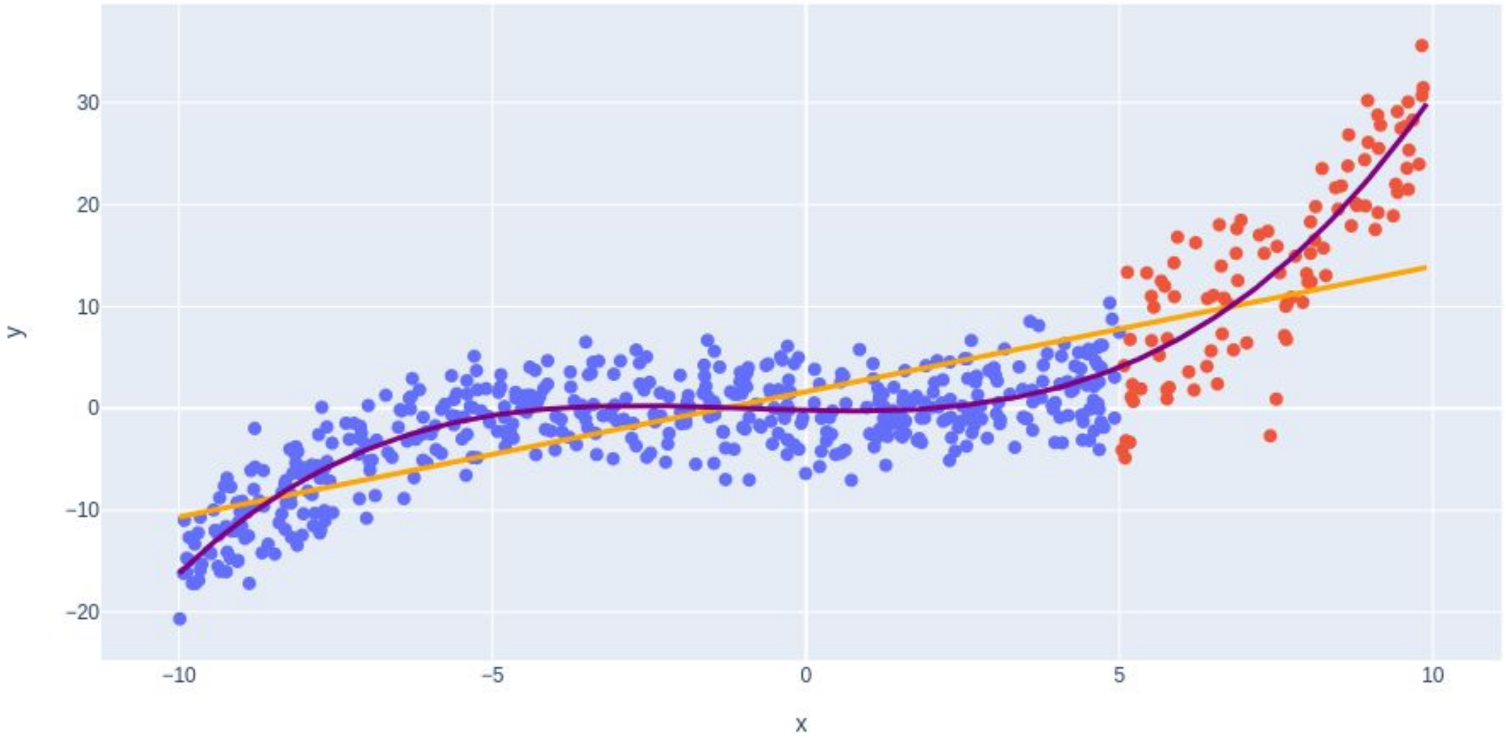
Now, for groups 0 and 1:

$$\Gamma := |(B_0 - B_1) + (V_0 - V_1) + (N_0 - N_1)|$$

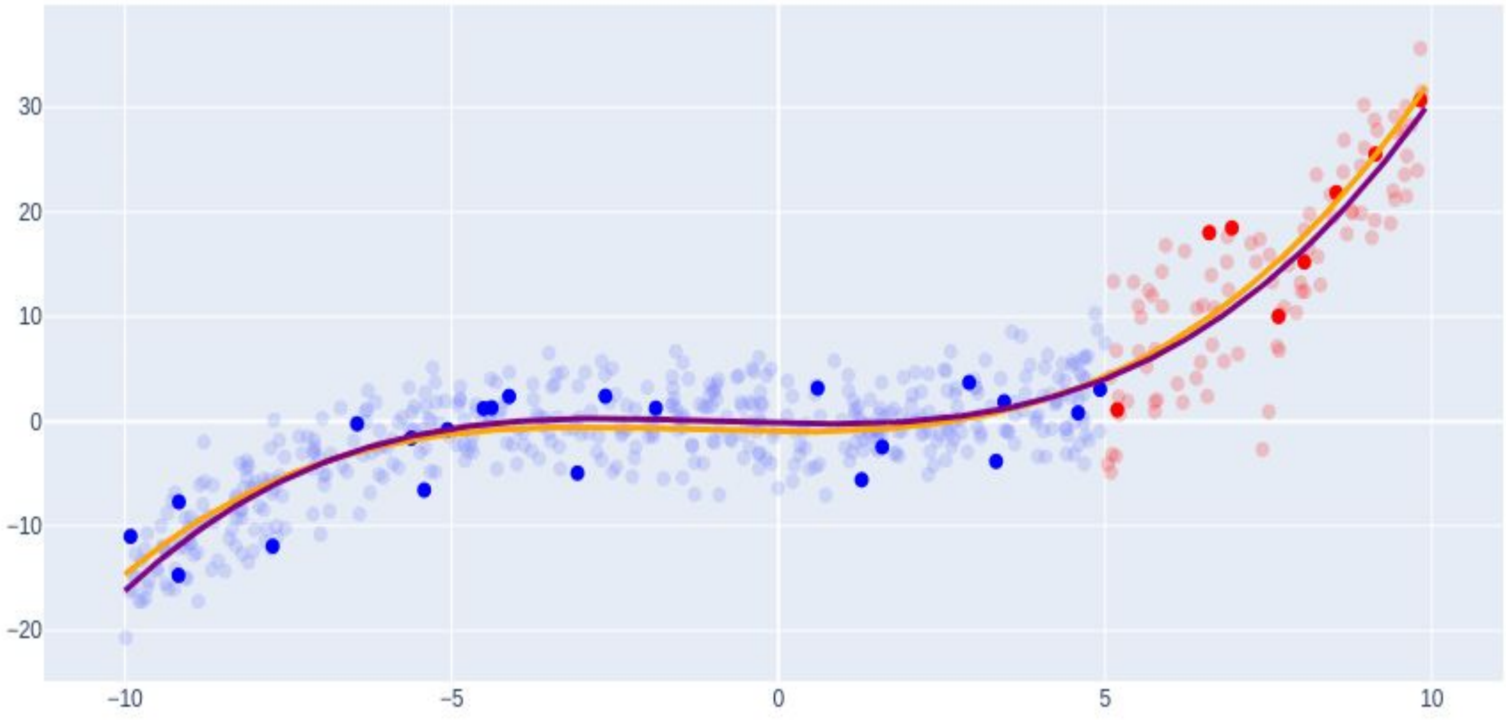
Discrimination Due to Bias



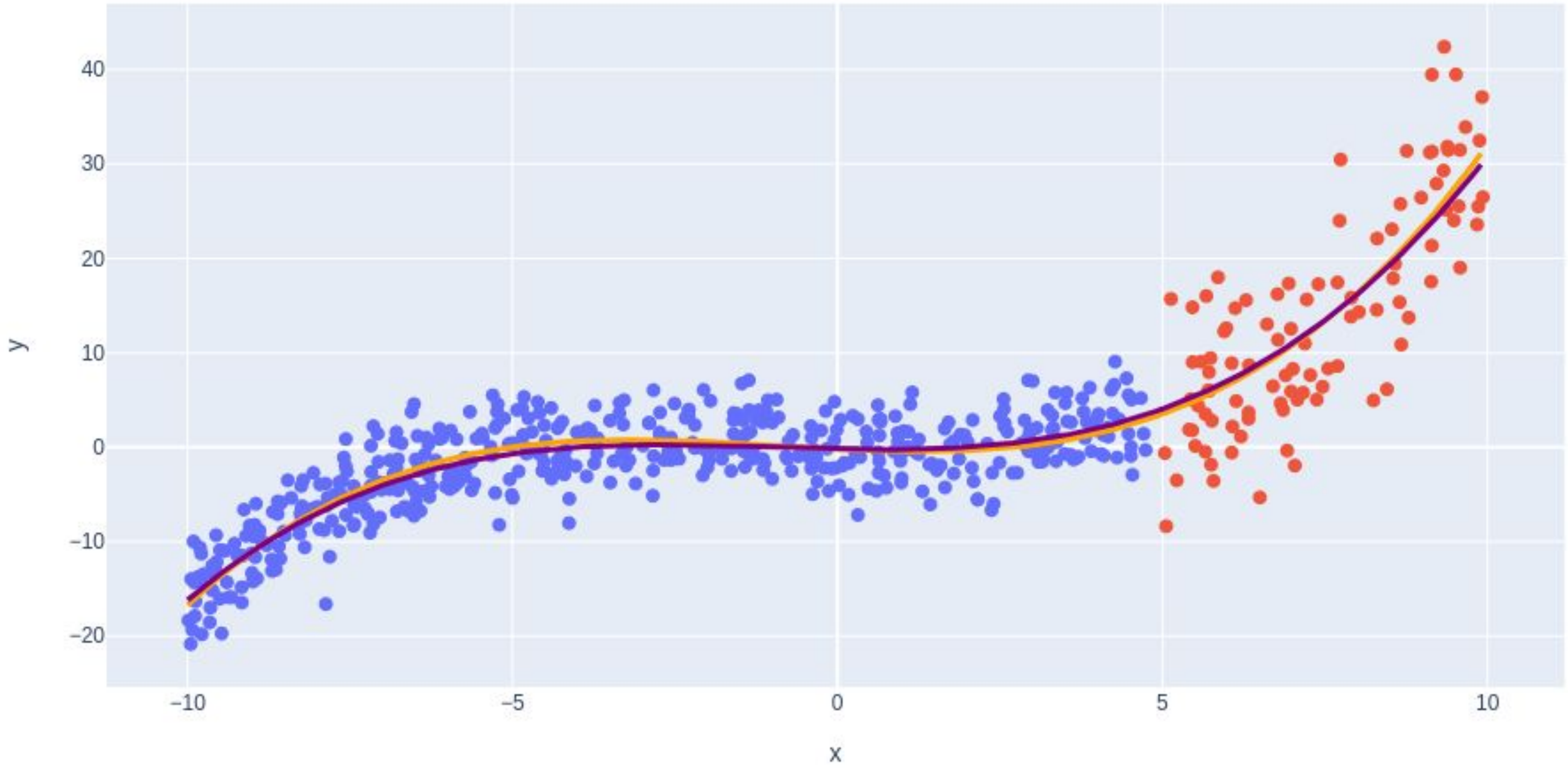
Discrimination Due to Bias



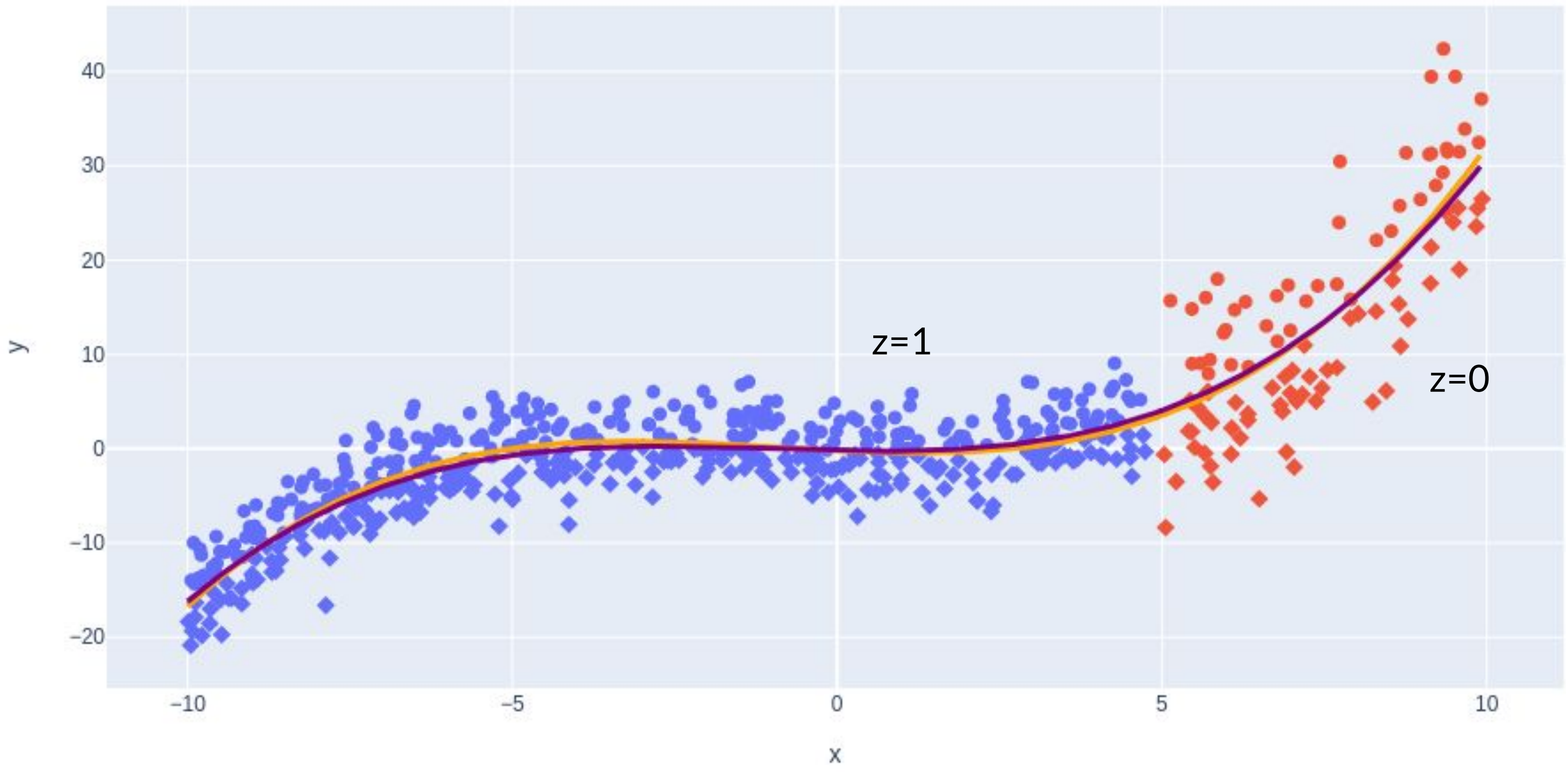
Discrimination Due to Variance



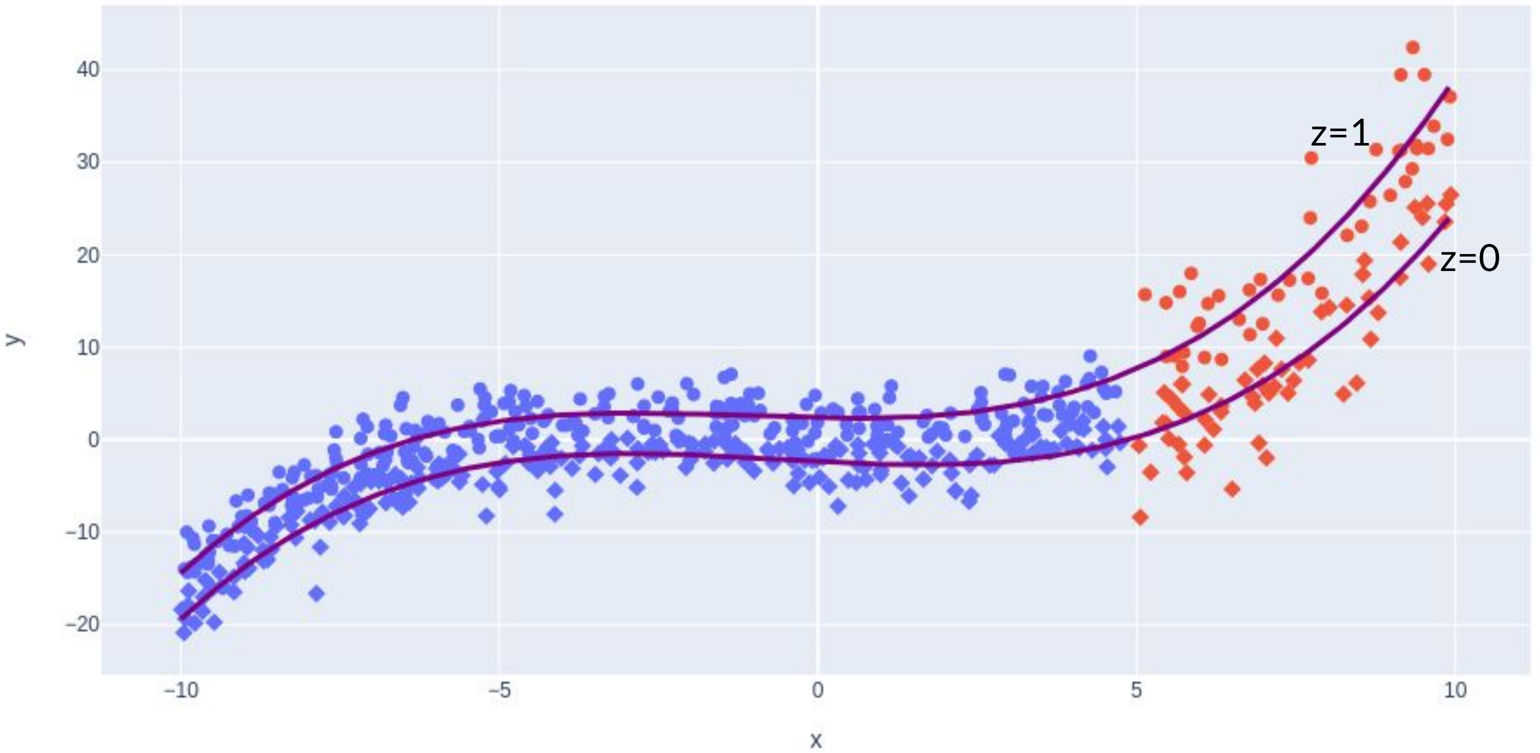
Discrimination Due to Noise



Discrimination Due to Noise



Discrimination Due to Noise



Experiment: Income Prediction

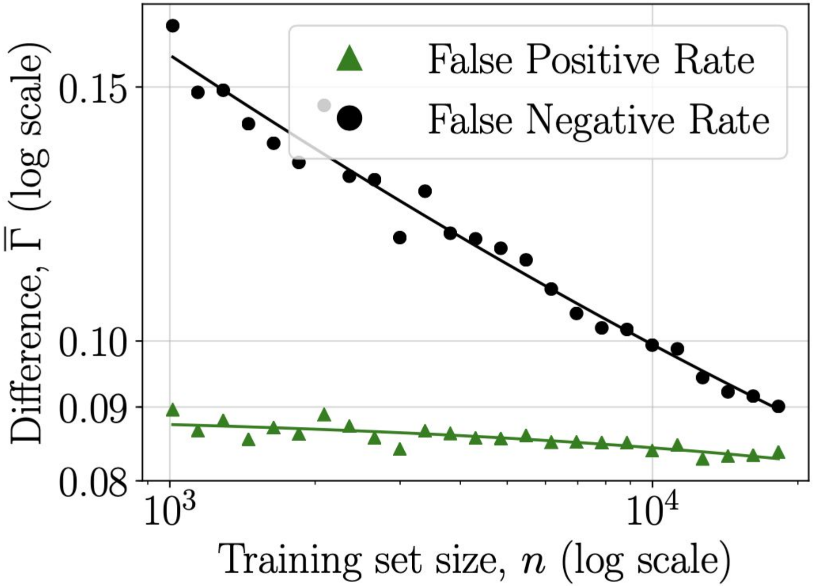
- Goal: Predict if income >$50,000 from census data such as education, age, and marital status.
- Protected attribute: gender
- Dataset*: 32,561 samples, 12 categorical and continuous features - preprocessing generates 105 input features
- Model: random forest

* [UCI Machine Learning Repository Adult Dataset]

Results: Income Prediction

$$\Gamma^{ZO}(\hat{Y}) = .085 \pm .069$$

	FPR	FNR
M	0.111 ± 0.011	0.388±0.026
F	0.033 ± 0.008	0.448±0.064



Income Prediction: Clustering

	Executive/ Managerial	Other
M	0.157	0.461
F	0.412	0.543

False Negative Rate, Clustered by Occupation Category

Improving Discrimination: Takeaways

	Description	**How to Fix**
Bias	How well model fits data	Change model class
Variance	How much sample size affects accuracy	Increase training data size ***Collect more data from smaller groups**
Noise	Error independent of model class or sample size	Increase number of features ***Cluster to find subgroups with high discrimination, add additional features**

Limitations

- Does not apply to custom post-hoc loss functions for maximizing error equality
    - Treats it as increasing variance for one group to improve fairness
- Hard to determine sources of error in practice
    - Variance can be estimated through bootstrapping
    - Bias and noise hard to measure directly

Questions?

Estimating Bias, Variance and Noise

Variance can be estimated through bootstrapping. Bias and noise, however, are in practice not easy to measure directly. Can still measure differences in discrimination level between two models using Monte-Carlo sampling:

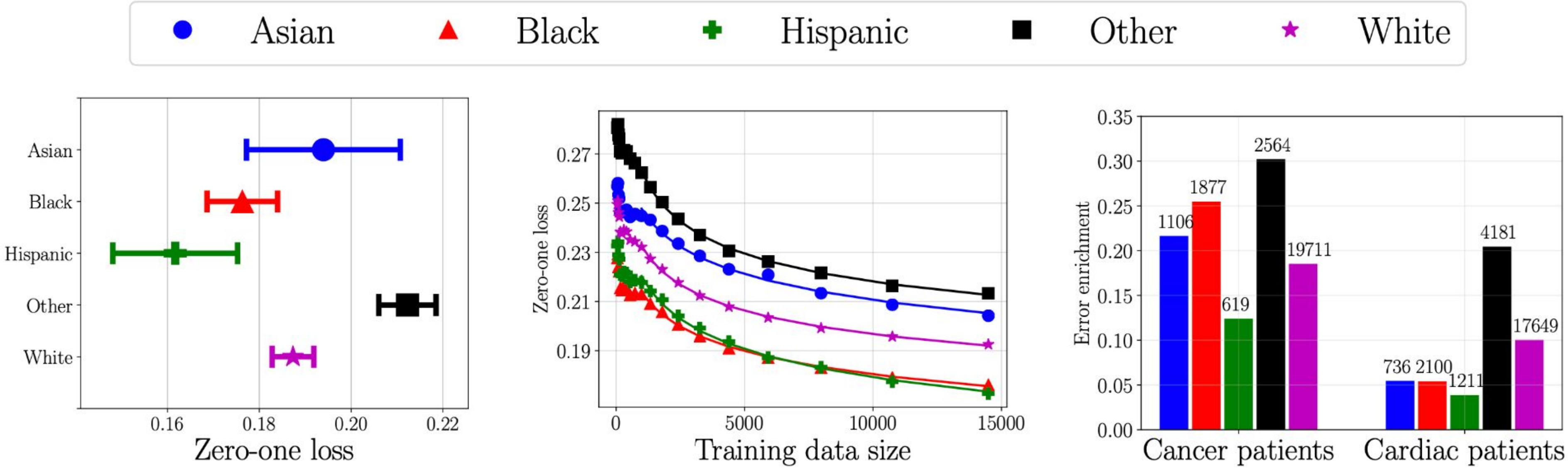
$$\gamma_a(\hat{Y}) = \frac{1}{\sum_i \mathbb{I}[a_i = a]} \sum_i L(y_i, \hat{y}_i)\mathbb{I}[a_i = a] \qquad \Gamma^\gamma := |\gamma_0(\hat{Y}) - \gamma_1(\hat{Y})|$$

$$Z_\alpha := \alpha(\Gamma(\hat{Y}) - \Gamma(\hat{Y}')), \alpha \in \{-1, 1\}$$

Experiment 2: Mortality Outcomes

- Goal: Predict hospital mortality from clinical notes
- Protected attribute: Self-reported ethnicity - Asian (2.2%), Black (8.8%), Hispanic (3.4%), White (70.8%), and Other (14.8%)
- Dataset: MIMIC-III - 25,879 patients, TF-IDF of 10,000 most frequent words
- Model: L1 Logistic Regression

Results: Hospital Mortality



Experiment 3: Book Reviews

- Goal: Predict rating (1-5) from text of review
- Protected attribute: Author gender
- Dataset: Goodreads reviews - 13,244 reviews, TF-IDF with 5,000 most common words - 18% female
- Model: Random forest

Results: Book Reviews

MSE: Male - 0.224 Female - 0.358