Is Your Classifier Actually Biased?

Measuring Fairness under Uncertainty with Bernstein Bounds

Kawin Ethayarajh

Stanford University

ACL 2020

Background: Classification Bias

Given a protected attribute A (e.g., gender), how can a classifier be fair with respect to A? (Hardt et al., 2016)

Background: Classification Bias

Given a protected attribute A (e.g., gender), how can a classifier be fair with respect to A? (Hardt et al., 2016)

- Demographic Parity:

$$\Pr[\widehat{y} = 1 | A = 1] = \Pr[\widehat{y} = 1 | A = 0]$$

Background: Classification Bias

Given a protected attribute A (e.g., gender), how can a classifier be fair with respect to A? (Hardt et al., 2016)

- Demographic Parity:

$$\Pr[\hat{y} = 1 | A = 1] = \Pr[\hat{y} = 1 | A = 0]$$

- Equal Opportunity:

$$\Pr[\hat{y} = 1 | A = 1, y = 1] = \Pr[\hat{y} = 1 | A = 0, y = 1]$$

Background: Classification Bias

Given a protected attribute A (e.g., gender), how can a classifier be fair with respect to A? (Hardt et al., 2016)

- Demographic Parity:

$$\Pr[\widehat{y} = 1 | A = 1] = \Pr[\widehat{y} = 1 | A = 0]$$

- Equal Opportunity:

$$\Pr[\widehat{y} = 1 | A = 1, y = 1] = \Pr[\widehat{y} = 1 | A = 0, y = 1]$$

- Equalized Odds

Classification Bias in NLP

Measuring classification bias in NLP is difficult.

1. Most datasets are not annotated with protected attributes.
2. Standard fairness measures cannot be used without annotations.
3. Manually annotating a large dataset is slow and expensive.

Classification Bias in NLP

Why not create a small dataset ($<$ 5K examples) annotated with a protected attribute and use it to estimate the bias?

1. WinoGender (Rudinger et al., 2018)
2. WinoBias (Zhao et al., 2018)
3. Equity Evaluation Corpus (Kiritchenko and Mohammad, 2018)

Classification Bias in NLP

The fewer examples we annotate, the more uncertain we are that

$$\text{bias estimate } \bar{\delta} \approx \text{population-level bias } \delta$$

Classification Bias in NLP

The fewer examples we annotate, the more uncertain we are that

$$\text{bias estimate } \bar{\delta} \approx \text{population-level bias } \delta$$

How can we quantify our uncertainty about the bias estimate?

Bernstein-bounded Unfairness (BBU)

Given protected $\{(x_a, y_a)\}$, unprotected $\{(x_b, y_b)\}$, we can define a cost $c(y, \widehat{y})$ such that the bias is equal to the difference in expected cost:

$$\delta = \mathbb{E}_a[c(y_a, \hat{y}_a)] - \mathbb{E}_b[c(y_b, \hat{y}_b)]$$

where δ is the population-level bias.

different fairness measures $\iff$ different cost functions

Bernstein-bounded Unfairness (BBU)

Letting $f(x) = \{+1, -1, 0\}$ denote that x is protected / unprotected / neither, we *amortize* the bias:

$$\hat{\delta}(x_i, f; c) = \frac{c(y_i, \hat{y}_i) f(x_i)}{\Pr[f(x) = f(x_i)]}$$
$$\delta(f; c) = \mathbb{E}_x[\hat{\delta}(x)]$$

By averaging $\{\hat{\delta}(x_i)\}$, we get a Monte Carlo estimate $\bar{\delta}$ of the true bias δ.

Bernstein-bounded Unfairness (BBU)

The probability that δ is within a constant t of $\bar{\delta}$ (Bernstein's inequality):

$$\Pr[|\bar{\delta} - \delta| > t] \leq 2 \exp\left(\frac{-nt^2}{2\sigma^2 + \frac{2C}{3\gamma} t} \right)$$

for n examples with max cost C, where γ is the frequency of the smaller group and γ is the variance of the amortized bias.

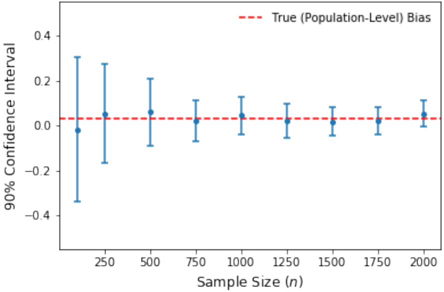
Bernstein-bounded Unfairness (BBU)

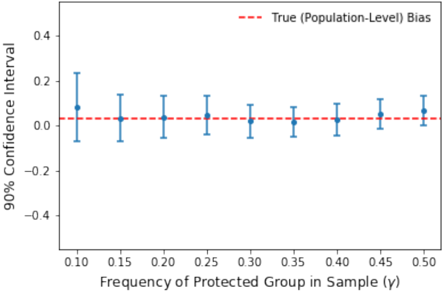
For a desired confidence level $\rho \in [0, 1]$, we can express our uncertainty about $\bar{\delta}$ as a confidence interval $[\bar{\delta} - t, \bar{\delta} + t]$.

$$\text{more uncertainty} \iff \text{higher } t \iff \text{wider confidence interval}$$

Experiments

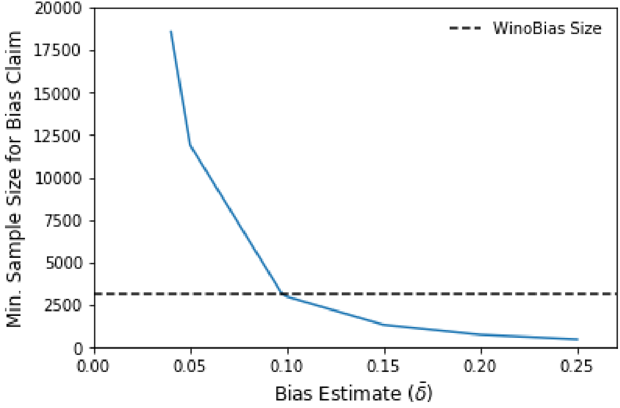
The bounds grow tighter as the sample size (left) and frequency of protected group (right) increases.





Experiments

We suspect that a co-reference resolution system is gender-biased. How much data do we need to be 95% confident about our bias claim?



Experiments

We suspect that a co-reference resolution system is gender-biased. How much data do we need to be 95% confident about our bias claim?

- To make a 95% confidence claim with WinoBias, system would need to be 9.75 points better on gender-stereotypical sentences.

Experiments

We suspect that a co-reference resolution system is gender-biased. How much data do we need to be 95% confident about our bias claim?

- To make a 95% confidence claim with WinoBias, system would need to be 9.75 points better on gender-stereotypical sentences.
- **We need larger bias-specific datasets!**

Takeaways

- It is possible to claim the *existence* of classification bias – with some level of confidence – without knowing the exact magnitude.
- Datasets currently used to estimate bias in NLP are too small to conclusively identify bias, except in the most egregious cases.

References

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems, pages 3315–3323.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 43–53.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20.