Unfair AI: It Isn’t Just the Data
Chowdhury M R Haider, Chris Clifton
{chaider, clifton}@purdue.edu

Summary

Conventional wisdom: biased training data leads to biased models.

We show,
- Machine learning can be expected to introduce types of bias not found in the training data.
- Different group-wise optimal models with unequal accuracy leads to unfair optimal accuracy joint model w.r.t disparate impact.
- Likely occurrence due to systemic bias.
- De-biasing training data is insufficient to ensure machine learning fairness.

Example Scenario

- College admission prediction
- Test score best classifies Caucasians,
- GPA best classifies the non-Caucasians, but less accurate than test-score on Caucasian

Problem Formulation

- Dataset, \( D = \{x^{(k)}, y^{(k)}, s^{(k)}\}_{k=1}^N \)
- Feature vector, \( x^{(k)} = [x_1^{(k)}, x_2^{(k)}, ..., x_n^{(k)}] \)
- Class labels, \( y_i \in \{+, -\} \)
- Sensitive attribute, \( s_i \in \{p, u\} \)
- Base rate, \( \alpha = \mathbb{P}(y^+ | s) = \mathbb{P}(y^+) \)
- Ratio of groups, \( \beta = \mathbb{P}(p)/\mathbb{P}(u) \)
- \( x_i^{sy} \sim N(\mu_i^{sy}, \sigma_i^{sy}) \) with \( \sigma_i^{sy} = \sigma_i^s \perp s, y \)
- Number of redlining features is \( 2r \)
- The rest \( n - 2r \) are independent

Disparate Impact of dataset \( D \),
\[
DI(D) = \frac{\mathbb{P}(y^+ | u)}{\mathbb{P}(y^+ | p)}
\]

Disparate Impact of model \( \theta \) (\( y \) prediction),
\[
DI(\theta) = \frac{\mathbb{P}(y^+ | u)}{\mathbb{P}(y^+ | p)}
\]

Assumptions

- \( r = 1, n = 2, \beta = 1 \)
- \( \mu_i^{p+} - \mu_i^{p-} = \mu_i^{u+} - \mu_i^{u-} = \delta \)
- \( \mu_i^{p+} - \mu_i^{p-} = \mu_i^{u+} - \mu_i^{u-} = 0 \)
- \( p \) is more separable than \( u \), \( \Rightarrow \sigma_1 < \sigma_2 \)

Result Summary

- True Positive Rate is higher for \( p \)
- False Positive Rate is higher for \( u \)

In other words,

- Favorable for the privileged
- Unfavorable for the unprivileged

Furthermore, with \( \alpha < 0.5 \),
\[
\mathbb{P}(y^+ | p) > \mathbb{P}(y^+ | u)
\]

Similarly, \( \alpha > 0.5 \) shows,
\[
\mathbb{P}(y^+ | p) < \mathbb{P}(y^+ | u)
\]

The joint optimal model is expected to induce disparate impact even when the training data with imbalanced base rates is free from such bias.

Discussion

- Systemic bias resulting from lack of diversity in feature design
- Dataset repair doesn’t always work
- Joint optimization of fairness and accuracy is a step towards right direction
- Effect of other systemic bias, i.e., disparity in noise rate, missing value rate, representation, etc., needs further study

Reference