Understanding and Mitigating Accuracy Disparity in Regression

Jianfeng Chi[§], Yuan Tian[§], Geoffrey J. Gordon[†], Han Zhao[‡]
[§]**University of Virginia,** [†]**Carnegie Mellon University,**
[‡]**University of Illinois Urbana-Champaign**
{jc6ub,yuant}@virginia.edu
ggordon@cs.cmu.edu, hanzhao@illinois.edu

ICML International Conference On Machine Learning

Overview

Accuracy Disparity Problem Exists in Regression Models:

Classifier	Metric	All	F	M
MSFT	PPV(%)	93.7	89.3	97.4
	Error Rate(%)	6.3	10.7	2.6
	TPR (%)	93.7	96.5	91.7
	FPR (%)	6.3	8.3	3.5
Face++	PPV(%)	90.0	78.7	99.3
	Error Rate(%)	10.0	21.3	0.7
	TPR (%)	90.0	98.9	85.1
	FPR (%)	10.0	14.9	1.1
IBM	PPV(%)	87.9	79.7	94.4
	Error Rate(%)	12.1	20.3	5.6
	TPR (%)	87.9	92.1	85.2
	FPR (%)	12.1	14.8	7.9

Error gaps between different demographic groups are too large!

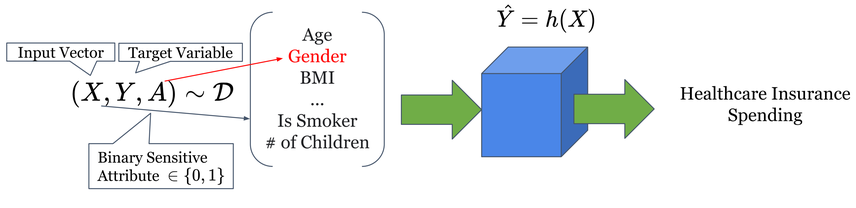
Figure 1: The above table is taken from (Buolamwini and Gebru, 2018)

Questions:

- *How does the accuracy disparity problem arise in regression?*
- *Are there any algorithmic interventions to reduce the disparity gap between different demographic subgroups in the regression setting?*

Preliminaries

- *Error Gap:* $\Delta_{\text{Err}} := |\text{Err}_{\mathcal{D}_0} - \text{Err}_{\mathcal{D}_1}|$.
- $\Delta_{\text{Err}} = 0$ implies *accuracy parity*.
- $\mathcal{D}_a^y$: the conditional distribution of $\mathcal{D}$ given $A = a$ and $Y = y$.



Main Results

Error Decomposition Theorem

Boundedness Assumption: There exists $M > 0$, such that for any hypothesis $\mathcal{H} \ni h : \mathcal{X} \to \mathcal{Y}$, $\|h\|_\infty \le M$ and $|Y| \le M$.

Theorem: If the above boundedness assumption holds, then for $\forall h \in \mathcal{H}$, let $\widehat{Y} = h(X)$, the following inequality holds:

$$\Delta_{\text{Err}}(h) \le 8M^2 \underbrace{d_{\text{TV}}(\mathcal{D}_0(Y), \mathcal{D}_1(Y))}_{\text{TV distance between label distributions across groups}} + 3M \underbrace{\min\{\mathbb{E}_{\mathcal{D}_0}[|\mathbb{E}_{\mathcal{D}_0^y}[\widehat{Y}] - \mathbb{E}_{\mathcal{D}_1^y}[\widehat{Y}]|],\ \mathbb{E}_{\mathcal{D}_1}[|\mathbb{E}_{\mathcal{D}_0^y}[\widehat{Y}] - \mathbb{E}_{\mathcal{D}_1^y}[\widehat{Y}]|]\}}_{\text{discrepancy between conditional predicted distributions across groups}}.$$

Implication:

- If the label distributions are highly imbalanced across groups, then the error gap could be potentially large.
- If we can minimize the second term on the right side, we then have a model that is free of accuracy disparity when the label distribution is well aligned.

Algorithmic Interventions

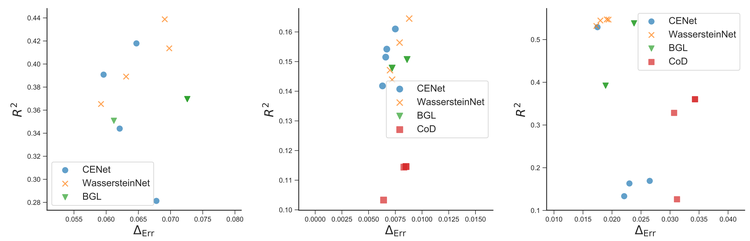
- Given a Markov chain $X \xrightarrow{g} Z \xrightarrow{h} \widehat{Y}$, we learn group-invariant *joint* representations between $\mathcal{D}_0(Z = g(X), Y)$ and $\mathcal{D}_1(Z = g(X), Y)$ via adversarial representation learning using a discriminator.
- We prove that the equilibria of the objective functions below are attained when the distances between *conditional* predicted distributions $\mathcal{D}_0^Y(Z = g(X))$ and $\mathcal{D}_1^Y(Z = g(X))$ are minimized.

$$\min_{h,g} \max_{f \in \mathcal{F}} \text{MSE}_{\mathcal{D}}(h(g(X)),\ Y) - \lambda \cdot \text{CE}_{\mathcal{D}}(A \,\|\, f(g(X), Y)) \tag{1}$$

$$\min_{h,g,Z_0 \sim g_\sharp \mathcal{D}_0, Z_1 \sim g_\sharp \mathcal{D}_1} \max_{f : \|f\|_L \le 1} \text{MSE}_{\mathcal{D}}(h(g(X)),\ Y) + \lambda \cdot \left| f(Z_0, Y) - f(Z_1, Y) \right|. \tag{2}$$

Experiments

- **Datasets:** (1) Adult, (2) COMPAS, (3) Communities and Crime, (4) Law School, and (5) Medical Insurance Cost.
- **Baselines:** (1) Bounded group loss (BGL) (Agarwal et al., 2019), (2) Coefficient of determination (CoD) (Komiyama et al., 2018).



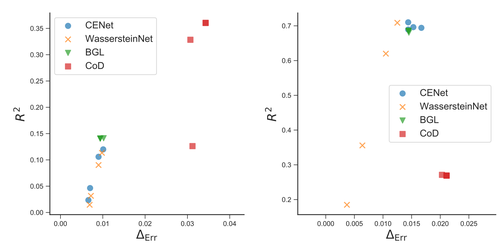


Figure 2: Overall results: R^2 regression scores and error gaps of different methods in five datasets. Results shown from left to right, top to bottom are from Adult, COMPAS, Crime, Law School, and Insurance datasets.

Conclusion: Trade-offs between regression performance and accuracy parity exist in all datasets. Our proposed methods achieve the best trade-offs in Adult, COMPAS, Crime and Insurance datasets.

Reference

- Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In International conference on machine learning, pages 2737–2746, 2018.
- Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction- based algorithms. In International Conference on Machine Learning, pages 120–129, 2019.
- Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency. PMLR, 2018.