Regression under demographic parity constraints via unlabeled post-processing

Gayane Taturyan, Evgenii Chzhen (NeurIPS 2024)
Presented by YunSeop Shin

September 28, 2025

Seoul national university, statistics, IDEA LAB

Introduction

- This paper considers the regression setting and proposes a post-processing procedure that, given a regressor $\eta(x) = \mathbb{E}[Y \mid X = x]$, adjusts its predictions so that they satisfy demographic-parity constraints.
- The post-processing is learned using unlabeled feature samples only, and the method assumes access to estimates of $\eta(x)$ and $\tau(x) = (\mathbb{P}(S = s \mid X = x))_{s=1}^{K}$.
- The authors discretize the output space and, on this finite grid, post-process the regressor into a randomized decision rule $\pi(\cdot \mid x)$ that behaves like a probabilistic classifier over grid points.
- They then minimize an entropy-regularized risk subject to demographic-parity constraints and can obtain a closed form for $\pi(\cdot \mid x)$.

Notation

- $(X, S, Y) \in \mathbb{R}^d \times [K] \times \mathbb{R}$: Feature, Sensitive attribute, Output.
- $\eta(x) := \mathbb{E}[Y \mid X = x]$ and assume $|\eta(X)| \leq B$.
- $\mathbf{p} := (p_s)_{s \in [K]}$, with $p_s := \mathbb{P}(S = s)$.
- $\tau(x) := (\tau_s(x))_{s \in [K]}$, with $\tau_s(x) := \mathbb{P}(S = s | X = x)$.
- $\pi(y|x) : \mathcal{B}(\mathbb{R}) \times \mathbb{R}^d \to [0, 1]$: A randomized prediction function.
- For any π define a random variable $\hat{Y}_\pi$ s.t.

$$\hat{Y}_\pi | X = x \sim \pi(\cdot | x).$$

- $\mathcal{R}(\pi) := \mathbb{E}[(\hat{Y}_\pi - \eta(X))^2]$: Risk of a prediction function π.
- $\mathcal{U}_s(\pi, \hat{y}) := |\mathbb{E}[\pi(\hat{y}|X)|S = s] - \mathbb{E}[\pi(\hat{y}|X)]|$: Measure of Unfairness.

Proposed methodology

- Discretization: For given integer $L \geq 0$ and real $B > 0$, a uniform grid
$$\hat{\mathcal{Y}}_L := \left\{ -B, -\frac{B(L-1)}{L}, \cdots, -\frac{B}{L}, 0, \frac{B}{L}, \cdots, \frac{B(L-1)}{L}, B \right\}.$$
- Author define Entrophic regularization as
$$\mathcal{R}_\beta(\pi) := \mathcal{R}(\pi) + \frac{1}{\beta}\mathbb{E}[\Psi(\pi(\cdot|X))]$$
where $\Psi(\mu) := \sum_{y \in \mathrm{supp}(\mu)} \mu(y) \log(\mu(y))$ and β is strength of regularization.
- Optimal discretized entropic-regularized fair estimator can be obtained by
$$\mathrm{argmin}_\pi \Big\{ \mathcal{R}_\beta(\pi) : \mathrm{supp}(\pi(\cdot|x)) = \hat{\mathcal{Y}}_L \text{ for } x \in \mathcal{R}^d, \\ \mathcal{U}_s(\pi, \hat{y}) \leq \epsilon_s \text{ for } \hat{y} \in \hat{\mathcal{Y}}_L,\ s \in [K] \}.$$

Closed form expression of the solution

- In this case, $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_m) : \mathbb{R}^m \to \mathbb{R}^m$ is the softmax function as $\sigma_j(\omega) = \frac{exp(\omega_j)}{\sum_{i=1}^{m} exp(\omega_i)}$ where $\omega = (\omega_1, \cdots, \omega_m)^t \in \mathbb{R}^m$ and $m = 2L + 1$.
- Author denote by $\mathrm{LSE}_\beta : \mathbb{R}^m \to \mathbb{R}$ the log-sum-exp function, defined as

$$\mathrm{LSE}_\beta(\omega) = \frac{1}{\beta} \log \left(\sum_{j=1}^{m} exp(\beta \omega_j) \right).$$

Closed form expression of the solution

- Let $\mathbf{t}(x) := 1 - \mathrm{Diag}(\mathbf{p})^{-1}\tau(x), \epsilon := (\epsilon_s)_s, \lambda_l = (\lambda_{ls})_s, \nu_l = (\nu_{ls})_s$ be length K vectors and $r_l(x) := (\eta(x) - \frac{lB}{L})^2$. For $L \in \mathbb{N}$ and $\beta > 0$, optimal discretized entropic-regularized fair estimator is given by

$$\pi_{\hat{\mathbf{\Lambda}},\hat{\mathbf{V}}}\left(\frac{Bl}{L}|x\right) = \sigma_l\left(\beta(\langle\,\hat{\lambda}_{l'} - \hat{\nu}_{l'}, \mathbf{t}(x)\rangle - r_{l'}(x))_{l' \in [[L]]}\right) \text{ for } l \in [[L]]$$

where $\hat{\mathbf{\Lambda}} = (\hat{\lambda}_{ls})_{l,s}$ and $\hat{\mathbf{V}} = (\hat{\nu}_{ls})_{l,s}$ matrices are solutions to

$$\mathrm{argmin}_{\mathbf{\Lambda},\mathbf{V}}\big\{F(\mathbf{\Lambda},\mathbf{V}) = \mathbb{E}[\mathrm{LSE}_\beta(\langle\,\lambda_l - \nu_l, \mathbf{t}(X)\rangle - r_l(X))_{l \in [[L]]}] + \sum_{l \in [[L]]}\langle\,\lambda_l + \nu_l, \epsilon\rangle\big\}.$$

Post-processing algorithm

- Gradient of $F(\mathbf{\Lambda}, \mathbf{V})$ is

$$\nabla_{\square_{ls}} F(\mathbf{\Lambda}, \mathbf{V}) = \Delta\, \mathbb{E}\Big[\sigma_l\Big(\beta\left(\langle\, \lambda_{l'} - \nu_{l'},\, \mathbf{t}(X)\rangle - r_{l'}(X)\right)_{l' \in [[L]]}\Big)\ t_s(X)\Big] + \epsilon_s$$

  where $\square \in \lambda, \nu$ and $\Delta = 1$ if $\square = \lambda$ and $\Delta = -1$ otherwise.
- Using the approximation of this gradient, we perform T steps of stochastic gradient descent to obtain the estimates $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{V}}$.

Theoretical guarantees

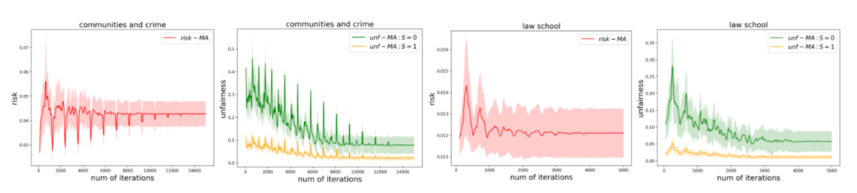
- Main theoretical guarantee is that both unfairness and risk decrease at the rate $\frac{1}{\sqrt{T}}$ i.e.,

$$\mathbb{E}^{1/2}\left[\sum_{\ell \in [[L]]} \sum_{s \in [K]} \left(\mathcal{U}_s\left(\pi_{\hat{\mathbf{\Lambda}},\hat{\mathbf{V}}}, \frac{B\ell}{L}\right) - \epsilon_s\right)_+^2\right] = O\left(\frac{1}{\sqrt{T}}\right) \text{ and}$$

$$\mathcal{E}(\pi_{\hat{\mathbf{\Lambda}},\hat{\mathbf{V}}}) = O\left(\frac{1}{\sqrt{T}}\right)$$

where $\mathcal{E}(\pi_{\hat{\mathbf{\Lambda}},\hat{\mathbf{V}}})$ is the excess risk, which is defined as the difference between the risk of $\pi_{\hat{\mathbf{\Lambda}},\hat{\mathbf{V}}}$ and that of the Bayes estimator.

Experiment



- Risk and unfairness of author's estimator on Communities and Crime and Law School datasets.
- Authors Split the data 0.4/0.4/0.2. The first 40% (with labels and the sensitive attribute) trains η and τ. The next 40% (features only) uses them to learn the post-processing $\hat{\pi}$. Finally, the last 20% is for testing.