SIPM

21 March 2023 09:22

Learning Fair Representation using a Parametric Integral Probability Metric

ICML '22

X : Non Sensitive random input vector.
S : Binary sensitive random input vector.
Y : Output variable.

$h : \mathcal{X} \times \{0,1\} \to \mathcal{Z}$ encoding function
$Z := h(X, S)$ representation.
$f_D : \mathcal{Z} \to \mathcal{X} \times \{0,1\}$ decoding function.

Focus: Unsupervised Fair Representation Learning:

loss to min : $L_{reconstruction}(f_D \circ h) + \lambda\, d(P_0^h, P_1^h)$ where $h(X,S)|S=s \sim P_s^h$

Prior: $d(P_0^h, P_1^h) = \sup_{v \in \mathcal{V}} E[\ell(S, v \circ h(X,S))]$

Computation issues.
Can't say that $d(P_0^h, P_1^h) \le \epsilon \Rightarrow$ bound on predictor's fairness.

ϕ- fairness of prediction model :

$$DP_\phi(f) = | E[\phi \circ f(X,S) | S=0] - E[\phi \circ f(X,S) | S=1] |$$

Proposed method: IPM-based deviance.

$$d_{\mathcal{V}}(P_0, P_1) = \sup_{v \in \mathcal{V}} \left| E_{P_0}[v(z)] - E_{P_1}[v(z)] \right|$$

If $d_{\mathcal{V}}(P_0, P_1) \le \epsilon$,

$$DP_\phi(f \circ h) = \left| E_{P_0}[\phi \circ f \circ h(X,S)] - E_{P_1}[\phi \circ f \circ h(X,S)] \right|$$
$$\le \epsilon \text{ if } \phi \circ f \in \mathcal{V}$$

Proposed parametric family for $\mathcal{V}$

$$\mathcal{V}_{sig} = \{ \sigma(\theta^T x + \mu) : \theta \in \mathbb{R}^m, \mu \in \mathbb{R} \}$$

where σ is the sigmoid function.

Lemma 1. $d_{\mathcal{V}_{sig}}(P_0, P_1) < \epsilon \Rightarrow \sup_{a \in \mathbb{R}^m} \sup_{t \in \mathbb{R}} | P(a^T U_0 \le t) - P(a^T U_1 \le t) | < \epsilon$

Let $\eta = \frac{1}{1 + e^{-1/\delta}}$

$$\eta | P(a^T U_0 \le t) - P(a^T U_1 \le t) | = | 1 - \eta P(a^T U_0 \le t) - (1 - \eta P(a^T U_1 \le t)) |$$

$$= \left| \begin{matrix} 1 - \eta P(a^T U_0 \in [t-\delta, t]) - \eta P(a^T U_0 \le t - \delta) - \\ (1 - \eta P(a^T U_1 \in [t-\delta, t]) - \eta P(a^T U_1 \le t - \delta)) \end{matrix} \right|$$

$$\le \eta \sum_{s \in \{0,1\}} P(a^T U_s \le t - \delta) + | 1 - \eta P(a^T U_0 \le t - \delta) - (1 - \eta P(a^T U_1 \le t - \delta)) |$$

$$= \eta \sum_{s \in \{0,1\}} P(a^T U_s \le t - \delta) + \left| \begin{matrix} \sum_{s \in \{0,1\}} \left(1 - \eta P(a^T U_s \le t - \delta) \right) - E\left[\sigma\left(\frac{a^T U_1 - t}{\delta^2} \right) \right] \\ - E\left[\sigma\left(\frac{a^T U_0 - t}{\delta^2} \right) \right] + E\left[\sigma\left(\frac{a^T U_0 - t}{\delta^2} \right) \right] \end{matrix} \right|$$

$$= \eta \sum_{s \in \{0,1\}} P(a^T U_s \le t - \delta) + \left| \begin{matrix} \sum_{s \in \{0,1\}} \left[E\left[\sigma\left(\frac{a^T U_1 - t}{\delta^2} \right) \right] - (1 - \eta P(a^T U_s \le t - \delta)) \right] \\ - E\left[\sigma\left(\frac{a^T U_0 - t}{\delta^2} \right) \right] \end{matrix} \right|$$

$$\le \eta \sum_{s \in \{0,1\}} P(a^T U_s \le t - \delta) + \left| \sum_{s \in \{0,1\}} \left[E\left[\sigma\left(\frac{a^T U_1 - t}{\delta^2} \right) \right] - (1 - \eta P(a^T U_s \le t - \delta)) \right] \right| + \left| E\left[\sigma\left(\frac{a^T U_1 - t}{\delta^2} \right) \right] - E\left[\sigma\left(\frac{a^T U_0 - t}{\delta^2} \right) \right] \right|$$

$$\le \eta\, 2\epsilon + 4\epsilon + \epsilon$$

$$= \frac{2\epsilon}{1 + e^{-1/\delta}} + 5\epsilon$$