Inherent Tradeoffs in Learning Fair Representations

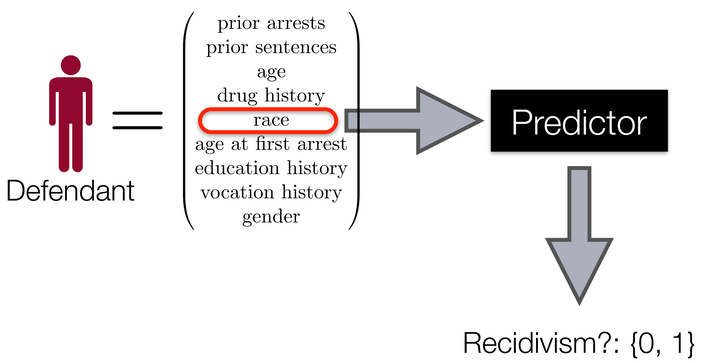
Han Zhao[†], Geoffrey J. Gordon[†,‡]

[†]Carnegie Mellon University, [‡]Microsoft Research Montreal

han.zhao@cs.cmu.edu, geoff.gordon@microsoft.com

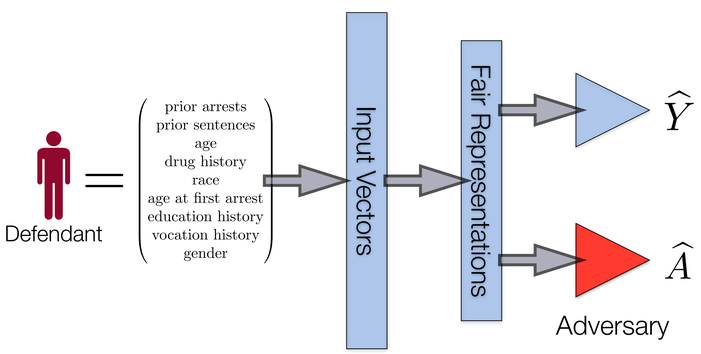
Overview

Recidivism Prediction:



- Joint distribution $\mathcal{D}$ over input $X \in \mathbb{R}^d$, label $Y \in \{0, 1\}$ and sensitive attribute $A \in \{0, 1\}$.
- A (randomized) classifier $\widehat{Y} = h(g(X)) \in \{0, 1\}$.
- Feature mapping $g : \mathcal{X} \to \mathcal{Z}$.
- A classifier is said to be *fair*, or satisfy *statistical parity* if: $\Pr_{\mathcal{D}}(\widehat{Y} = 1 \mid A = 0) = \Pr_{\mathcal{D}}(\widehat{Y} = 1 \mid A = 1) \Leftrightarrow \widehat{Y} \perp A$.
- Base rate: $\Pr(Y = 1 \mid A = a), a \in \{0, 1\}$. $\Delta_{\mathrm{BR}} := |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$.

Learning Fair Representations:



- Representations compete with the adversary to confuse the latter.

Question:

What is the fundamental tradeoff between fairness and utility? How fair representations affect utility and fairness of the predictor?

Our Answer: When the base rates differ, any fair classifier has to make a large error on at least one of the demographic group.

Preliminary

Demographic Parity Gap (DP Gap):

$$\Delta_{\mathrm{DP}}(\widehat{Y}) := |\mathcal{D}_0(\widehat{Y} = 1) - \mathcal{D}_1(\widehat{Y} = 1)|.$$

- $\mathcal{D}_a(\cdot) = \Pr_{\mathcal{D}}(\cdot \mid A = a), \forall a \in \{0, 1\}$.
- $\Delta_{\mathrm{DP}}(Y) = \Delta_{\mathrm{BR}}$, by definition.

Accuracy Parity: $\mathrm{Err}_{\mathcal{D}_0}(\widehat{Y}) = \mathrm{Err}_{\mathcal{D}_1}(\widehat{Y})$.

- $\mathrm{Err}_{\mathcal{D}}(\widehat{Y}) = \mathbb{E}_{\mathcal{D}}[Y \neq \widehat{Y}]$.

f-divergence: Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability distributions over the same space and assume $\mathcal{P}$ is absolutely continuous w.r.t. $\mathcal{Q}$ ($\mathcal{P} \ll \mathcal{Q}$). Then for any convex function $f : (0, \infty) \to \mathbb{R}$ that is strictly convex at 1 and $f(1) = 0$, the f-divergence of $\mathcal{Q}$ from $\mathcal{P}$ is defined as

$$D_f(\mathcal{P} \,||\, \mathcal{Q}) := \mathbb{E}_{\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right].$$

The function f is called the *generator function* of $D_f(\cdot \,||\, \cdot)$.

Name	$D_f(\mathcal{P} \,||\, \mathcal{Q})$	$f(t)$	Symm.	Tri.
Kullback-Leibler	$D_{\mathrm{KL}}(\mathcal{P} \,||\, \mathcal{Q})$	$t \log t$	✗	✗
Reverse-KL	$D_{\mathrm{KL}}(\mathcal{Q} \,||\, \mathcal{P})$	$-\log t$	✗	✗
Jensen-Shannon	$D_{\mathrm{JS}}(\mathcal{P}, \mathcal{Q})$	$t \log t - (t+1) \log(\frac{t+1}{2})$	✓	✗
Squared Hellinger	$H^2(\mathcal{P}, \mathcal{Q})$	$(1 - \sqrt{t})^2/2$	✓	✗
Total Variation	$d_{\mathrm{TV}}(\mathcal{P}, \mathcal{Q})$	$\lvert t - 1\rvert/2$	✓	✓

Tradeoff between Utility and Fairness

Theorem: Let $\widehat{Y} = h(g(X))$ be the predictor. If $\widehat{Y}$ satisfies statistical parity, then $\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq \Delta_{\mathrm{BR}}(\mathcal{D}_0, \mathcal{D}_1)$.

- If $A = Y$ a.s., then $\Delta_{\mathrm{BR}}(\mathcal{D}_0, \mathcal{D}_1) = 1$ and $\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq 1$.
- If $A \perp Y$, then $\Delta_{\mathrm{BR}}(\mathcal{D}_0, \mathcal{D}_1) = 0$ and $\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq 0$, i.e., no constraint on utility.
- The lower bound is tight, in the sense that there exists problem instances where the equality holds.

Implication: When the base rates differ between different demographic groups, then any fair algorithm has to make an error of at least $\Delta_{\mathrm{BR}}(\mathcal{D}_0, \mathcal{D}_1)/2$ on one of the groups.

Tradeoff in Fair Representation Learning

Induced distribution by feature mapping g:

Given a feature mapping $g : \mathcal{X} \to \mathcal{Z}$, define $g_\sharp\mathcal{D} := \mathcal{D} \circ g^{-1}$ to be the induced distribution (pushforward) of $\mathcal{D}$ under g, i.e., for any event $E' \subseteq \mathcal{Z}$, $g_\sharp\mathcal{D}(E') := \mathcal{D}(g^{-1}(E')) = \mathcal{D}(\{x \in \mathcal{X} \mid g(x) \in E'\})$.

Theorem: Assume $d_{\mathrm{JS}}(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1) \leq d_{\mathrm{JS}}(\mathcal{D}_0(Y), \mathcal{D}_1(Y))$ and $H(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1) \leq H(\mathcal{D}_0(Y), \mathcal{D}_1(Y))$, then the following hold:

- Total variation lower bound:

$$\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq d_{\mathrm{TV}}(\mathcal{D}_0(Y), \mathcal{D}_1(Y)) - d_{\mathrm{TV}}(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1).$$

- Jensen-Shannon lower bound:

$$\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq \left(d_{\mathrm{JS}}(\mathcal{D}_0(Y), \mathcal{D}_1(Y)) - d_{\mathrm{JS}}(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1)\right)^2/2.$$

- Hellinger lower bound:

$$\mathrm{Err}_{\mathcal{D}_0}(h \circ g) + \mathrm{Err}_{\mathcal{D}_1}(h \circ g) \geq \left(H(\mathcal{D}_0(Y), \mathcal{D}_1(Y)) - H(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1)\right)^2/2.$$

Fair Representations Lead to Accuracy Parity:

Theorem: For any hypothesis $\mathcal{H} \ni h : \mathcal{X} \to \mathcal{Y}$, the following inequality holds:

$$|\mathrm{Err}_{\mathcal{D}_0}(h) - \mathrm{Err}_{\mathcal{D}_1}(h)| \leq n_{\mathcal{D}_0} + n_{\mathcal{D}_1} + d_{\mathrm{TV}}(\mathcal{D}_0(X), \mathcal{D}_1(X)) + \min\left\{\mathbb{E}_{\mathcal{D}_0}[|h_0^* - h_1^*|], \mathbb{E}_{\mathcal{D}_1}[|h_0^* - h_1^*|]\right\}.$$

- $n_{\mathcal{D}_a},\ a \in \{0, 1\}$: the noise over distribution $\mathcal{D}_a$.
- $h_a^*,\ a \in \{0, 1\}$: the optimal predictor over distribution $\mathcal{D}_a$.

Experiments

Income prediction on the Adult dataset, sensitive attribute: gender.