Obtaining Fairness using Optimal Transport Theory

ICML 2019, Long Beach

E. del Barrio[a], F. Gamboa[b], **P. Gordaliza**[a,b] and JM. Loubes[b]

a : IMUVa (Universidad de Valladolid)
b : IMT (Université de Toulouse)

INSTITUT de MATHEMATIQUES de TOULOUSE

imUVa

Framework for achieving Fairness

Target class

$$Y = \begin{cases} 0 & failure \\ 1 & success \end{cases}$$

Visible attributes

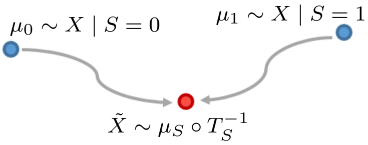
$$X \in \mathbb{R}^d,\ d \geqslant 1,$$

Protected attribute

$$S = \begin{cases} 0 & unfavored \\ 1 & favored \end{cases}$$

Goal: Replace X by $\tilde{X}$ such that for all $g \in \mathcal{G}$
$\mathcal{L}\left(g(\tilde{X}) \mid S = 0\right) = \mathcal{L}\left(g(\tilde{X}) \mid S = 1\right)$

Methodology: Find $\tilde{X} = T_S(X)$ such that
$\mathcal{L}\left(T_0(X) \mid S = 0\right) = \mathcal{L}\left(T_1(X) \mid S = 1\right)$



Questions:

a) Best choice for the distribution $\tilde{X} \sim \nu$?

b) Optimal way of transporting μ_1 and μ_0 to ν?

Reasonable and feasible solutions: T_S optimal transport map carrying μ_S towards their Wasserstein barycenter μ_B with weights $\pi_0 = P(S = 0)$ and $\pi_1 = P(S = 1)$:

$$\mu_{S\sharp} T_S = \mu_B$$

$$\mu_B \in \text{argmin}_{\nu \in \mathcal{P}_2} \left\{ \pi_0 W_2^2(\mu_0, \nu) + \pi_1 W_2^2(\mu_1, \nu) \right\}$$

Our proposal

1. **Justification for repair with Wasserstein Barycenter:** under some regularity conditions, $\mathcal{E}(\tilde{X}) := R_B(\tilde{X}) - R_B(X,S)$

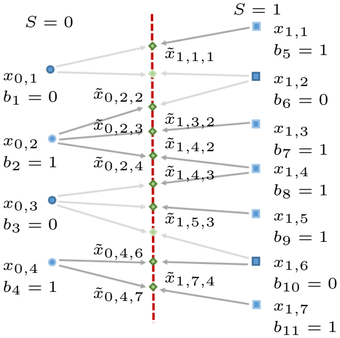
$$\mathcal{E}(\tilde{X}) \leq 2\sqrt{2}K \left(\sum_{s=0,1} \pi_s \mathcal{W}_2^2(\mu_s, \mu_{s\sharp}T_s) \right)^{\frac{1}{2}}, K > 0$$

2. **Multidimensional extension for Total Repair**
3. **New Partial Repair method: Random Repair** (Improvement of Geometric Repair)

For $s \in \{0,1\}$,

$$\tilde{\mu}_{s,\lambda} = \mathcal{L}(BT_s(X) + (1-B)X \mid S = s),$$
$$B \sim \mathcal{B}(\lambda) \text{ with } \lambda \in (0,1) \text{ level of repair}$$

$\lambda = 0$ Accuracy of $g(\tilde{X})$	$\longleftarrow$ **Trade-off** $\longrightarrow$	$\lambda = 1$ Blurring of S



Related work

E. del Barrio and J.-M. Loubes. (2019) Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, **47**, 926–951.

E. del Barrio, P. Gordaliza and J.-M. Loubes. (2019) A central limit theorem on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA.*

Thanks for the attention!