Trustworthy Machine Learning

Fairness in Learning 2

Sangdon Park

POSTECH

Contents from

Learning Fair Representations

Richard Zemel ZEMEL@CS.TORONTO.EDU
Yu (Ledell) Wu WUYU@CS.TORONTO.EDU
Kevin Swersky KSWERSKY@CS.TORONTO.EDU
Toniann Pitassi TONI@CS.TORONTO.EDU
University of Toronto, 10 King's College Rd., Toronto, ON M6H 2T1 CANADA

Cynthia Dwork DWORK@MICROSOFT.COM
Microsoft Research, 1065 La Avenida Mountain View, CA. 94043 USA

- ICML 2013

Why Representation Learning for Fairness?

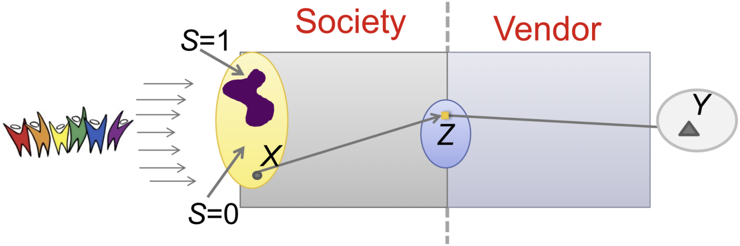


Image Credit: Richard Zemel

Goal:

- Find a representation Z that removes information about sensitive attributes
- Then, vendors can do whatever they want!
- *i.e.*, Separate the fairness responsibility of the (trusted) society and (untrusted) vendors.

Representation Learning via Prototypes

Prototype representation

$$\hat{x}_i := \sum_{k=1}^{K} \mathbb{P}\{z = k \mid x_i\} v_k$$

- $\mathcal{X}_n := \{x_1, \ldots, x_n\}$: examples
- K: the number of prototypes
- v_k: the k-th prototype
- $\mathbb{P}\{z = k \mid x\}$: the weight of the k-th prototype for x, *i.e.*,

$$\mathbb{P}\{z = k \mid x\} := \frac{e^{-d(x, v_k)}}{\sum_{k=1}^{K} e^{-d(x, v_k)}},$$

  where $d(x, v)$ is a distance, *e.g.*, $d(x, v) := \sum_{d=1}^{D} \alpha_d |x_d - v_d|^2$, where D is the dimension of examples.
- learnable parameters: v_k, α_d

Information Loss

information loss

$$L_x := \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$

- An reconstructed example $\hat{x}_i$ from prototypes is similar to the original example x_i.

Fairness Constraint

statistical parity

$$L_z := \sum_{k=1}^{K} \left| \frac{1}{|\mathcal{X}^+|} \sum_{x^+ \in \mathcal{X}^+} \mathbb{P}\{z = k \mid x^+\} - \frac{1}{|\mathcal{X}^-|} \sum_{x^- \in \mathcal{X}^-} \mathbb{P}\{z = k \mid x^-\} \right|$$

- $\mathcal{X}^+ \subseteq \mathcal{X}_n$: examples with sensitive attributes
- $\mathcal{X}^- \subseteq \mathcal{X}_n$: examples with non-sensitive attributes
- Demographic parity?

Definition (demographic parity)

$$\mathbb{P}\left\{\widehat{Y} = 1 \,\middle|\, A = 0\right\} = \mathbb{P}\left\{\widehat{Y} = 1 \,\middle|\, A = 1\right\}$$

Classification Loss

binary cross-entropy

$$L_y := \sum_{i=1}^{N} -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)$$

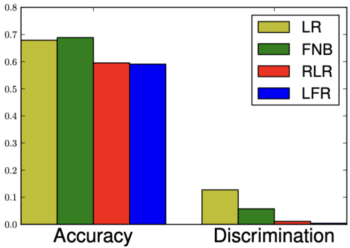
- A binary classifier: $\hat{y}_i = \sum_{k=1}^{K} \mathbb{P}\{z = k \mid x_i\} w_k$
- Learnable parameters: $w_k \in [0, 1]$

Learning Objective

objective function

$$\min_{v_k, \alpha_d, w_k} \lambda_x L_x + \lambda_z L_z + \lambda_y L_y$$

Results



- Discrimination (measuring statistical parity):

$$\left| \frac{\sum_{i:x_i \in \mathcal{X}^+} \hat{y}_i}{|\mathcal{X}^+|} - \frac{\sum_{i:x_i \in \mathcal{X}^-} \hat{y}_i}{|\mathcal{X}^-|} \right|$$

- Achieves high accuracy while satisfying the fairness constraint