Sensitive Information Disentanglement with Generative Model

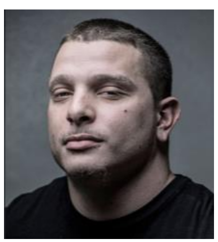
Taeuk Jang

24th, April, 2022

Purdue University

Preliminary

Biased decision making by algorithms



ML Classifier

	White	Black
High Risk, Didn't re-offend	23.5%	44.9%
Low Risk, Did re-offend	47.7%	28.0%

Make favorable decisions to white defendants

*Images from ProPublica

What is Fairness?

Individual Fairness

- Similar samples should be treated similarly.

Group Fairness

- Demographic Parity [Dwork et al. 2012]

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$$

- Equalized Odds [Hardt et al. 2016]

$$P(\hat{Y} = Y \mid A = 0) = P(\hat{Y} = Y \mid A = 1)$$

- Predictive Parity and more…

What is Fairness?

Achieving Group Fairness is non-trivial problem

- Removing sensitive information (Fairness through blindness) is not enough
- There are features that are highly correlated to sensitive information.

  e.g., ZIP code, graduated college, etc

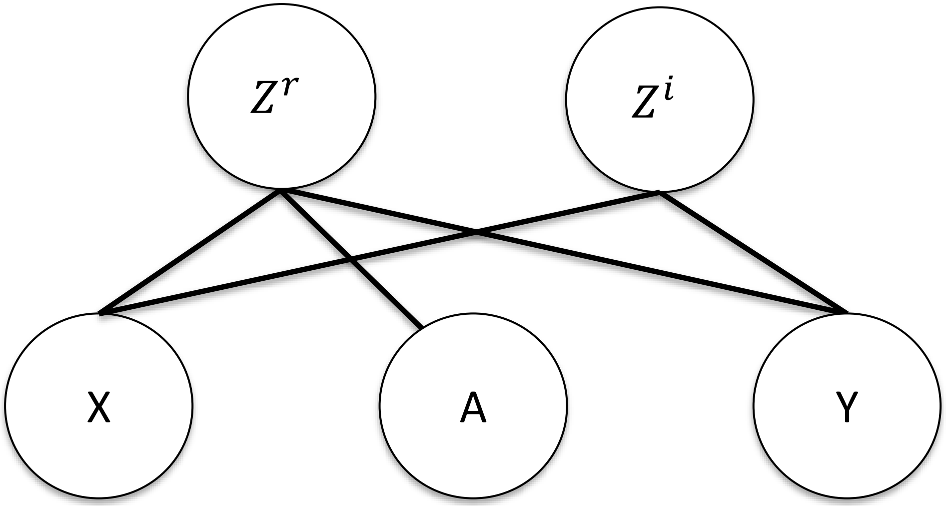
To achieve group fairness…

- Data Perspective
- Model Perspective
- Post processing

Problem Definition

Disentangle observed data into independent latent features

We assume $Z^i \perp Z^r$



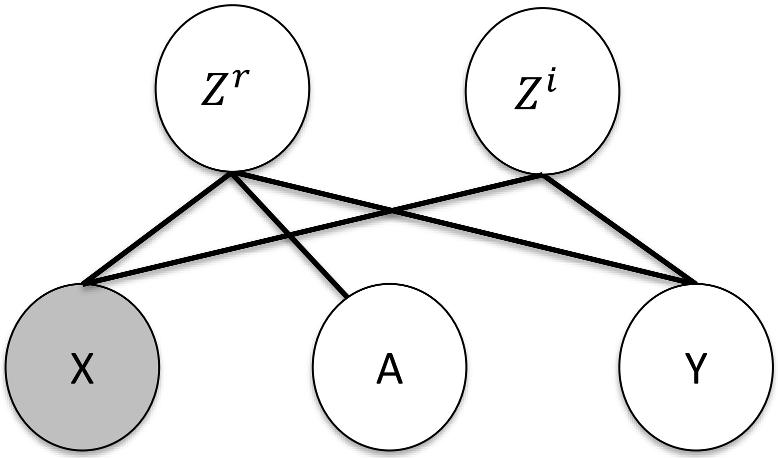
Z^r : sensitive relevant features

Z^i : sensitive irrelevant features

Problem Definition

Goal #1: Maximize $\log p_\theta(X)$

$$\begin{aligned}\log p_\theta(X) &\geq \mathcal{L}_{ELBO} \\ &= \mathbb{E}_{q_\phi(Z^r,Z^i|X)}\left[p_\theta(X|Z^r,Z^i)\right] - D_{KL}\big(q_\phi(Z^r,Z^i|X))||p(Z^r,Z^i)\big) \\ &= \mathbb{E}_{q_\phi(Z^r,Z^i|X)}\left[p_\theta(X|Z^r,Z^i)\right] - D_{KL}\big(q_\phi(Z^r|X))||p(Z^r)\big) \\ &\qquad - D_{KL}\big(q_\phi(Z^i|X))||p(Z^i)\big),\end{aligned}$$



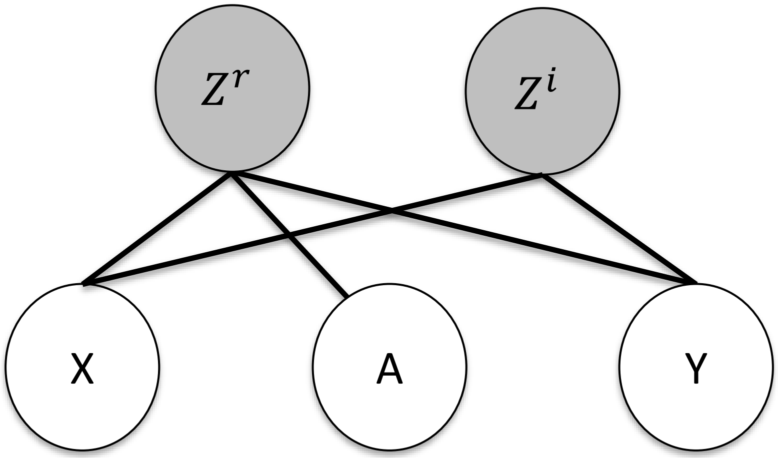
Z^r : sensitive relevant features

Z^i : sensitive irrelevant features

Problem Definition

Goal #2: Maximize $\log p_\theta(A|Z^r)$ and $\log p_\theta(Y|Z^i, Z^r)$

We minimize $\mathcal{L}_{CE}(Y, C_{\omega 1}(Z^r \oplus Z^i)) + \mathcal{L}_{CE}(A, C_{\omega 2}(Z^r))$



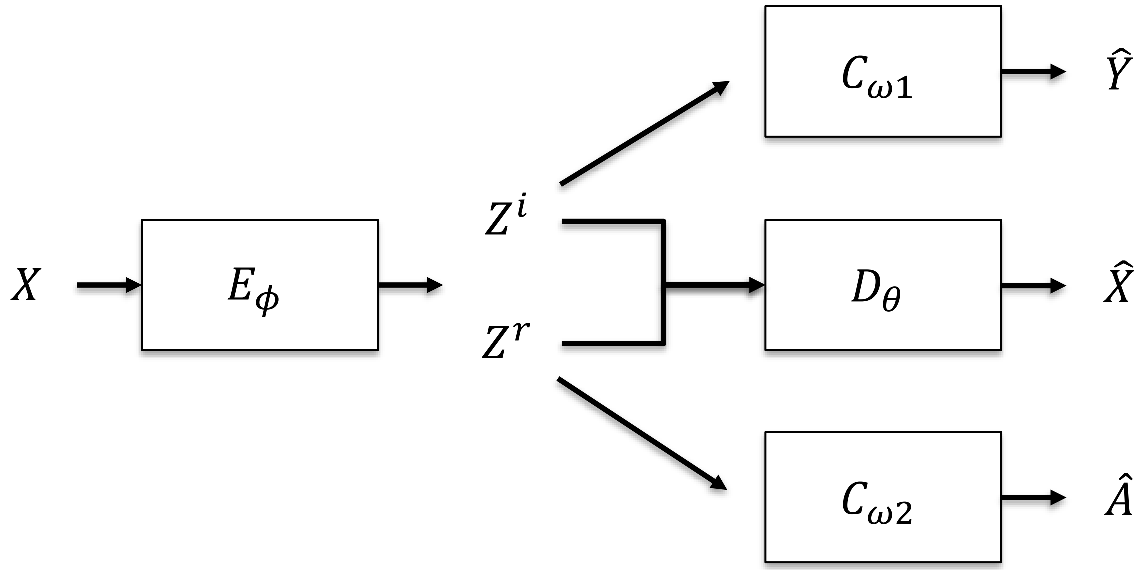
Z^r : sensitive relevant features

Z^i : sensitive irrelevant features

Problem Definition

Final Objective :

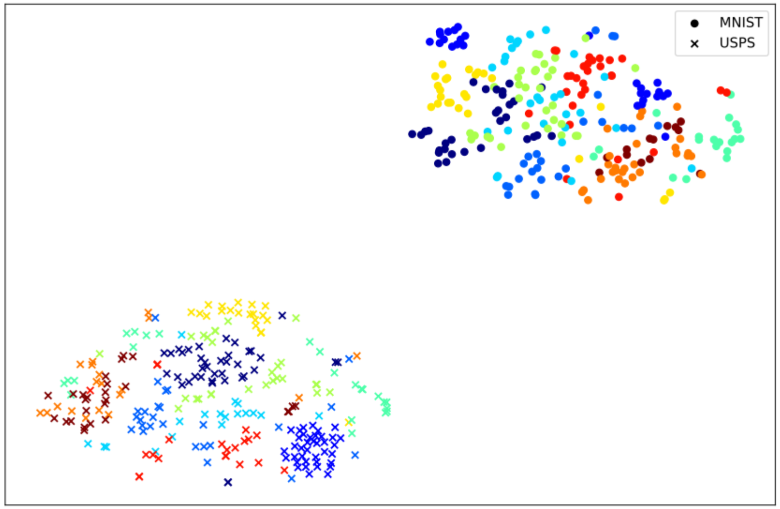
$$\arg\min_{\theta,\phi,\omega} \mathcal{L}_{MSE}(X, D_\theta(Z^r \oplus Z^i)) + D_{KL}\big(q_\phi(Z^r|X)||p(Z^r)\big)$$
$$+ D_{KL}\big(q_\phi(Z^i|X)||p(Z^i)\big) + \mathcal{L}_{CE}(Y, C_{\omega 1}(Z^r \oplus Z^i)) + \mathcal{L}_{CE}(A, C_{\omega 2}(Z^r))$$



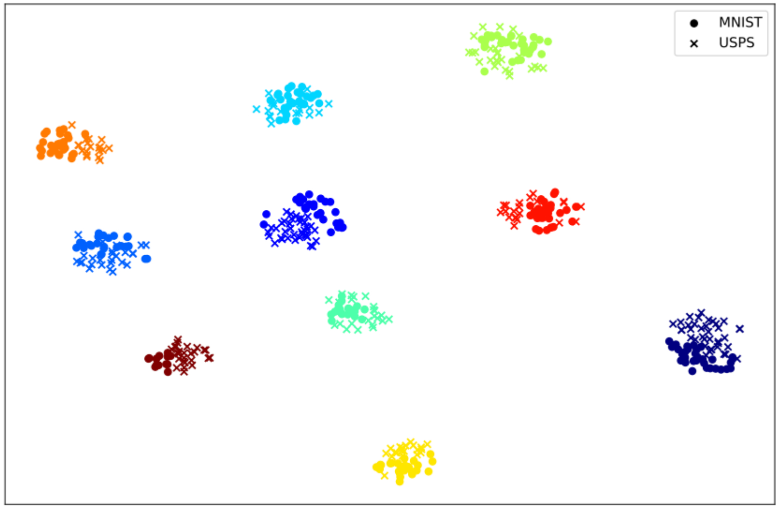
PURDUE
UNIVERSITY

Experimental Result

Multi-class Classification (MNIST-USPS)



Z^r



Z^i

Tabular Dataset (Adult Income Dataset)

	Acc	Acc diff	EOp	EOd	DP
Baseline	0.853	0.108	0.119	0.098	0.186
SD-VAE	0.838	0.099	0.047	0.050	0.155

Table 1: Comparison of SD-VAE with Logistic regression on Adult dataset.

PURDUE UNIVERSITY

Experimental Result

CelebA Dataset

$\hat{X} = D_\theta([Z^i, Z^r])$

$\hat{X}' = D_\theta([Z^i, \tilde{Z}^r])$

$\hat{X}'' = D_\theta([\tilde{Z}^i, Z^r])$

X

Experimental Result

CelebA Dataset

$\hat{X}' = D_\theta([Z^i, \tilde{Z}^r])$

$\hat{X}'' = D_\theta([\tilde{Z}^i, Z^r])$

PURDUE UNIVERSITY