Learning fair representation with a parametric integral probability metric

D.Kim [1,2] and K.Kim [3] and I.Kong [3] and I.Ohn [4] and Y.Kim [3]

Speaker : Dongha Kim

[1]Department of Statistics, Sungshin Women's University

[2]Data Science Center, Sungshin Women's University

[3]Department of Statistics, Seoul National University

[4]Department of Statistics, Inha University

Learning fair representation (LFR)

- Learning a function $h(X, S) : \mathcal{X} \times \{0, 1\} \rightarrow \mathcal{Z} \subset \mathbb{R}^m$ that maps data including the sensitive information to feature vectors so that the distributions for each sensitive group are similar.
- Adversarial training scheme is a popular approach for LFR.
- But, most methods suffered from
  - lacking their theoretical properties.
  - unstability, so they are not appealing to practitioners.
- Need to develop a new method that
  - has concrete theoretical guarantees.
  - gives stable results for various initializations.

Sigmoid IPM (sIPM)

- We consider a IPM measure with the sigmoid function to learn fair representations:

$$d_{\mathcal{V}_{sig}}(\mathbb{P}, \mathbb{Q}) = \sup_{v \in \mathcal{V}} \left| \int v(\mathbf{z})(d\mathbb{P}(\mathbf{z}) - d\mathbb{Q}(\mathbf{z})) \right|,$$

where $\mathbb{P}, \mathbb{Q}$ are two probability measures and

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

  - $\sigma(\cdot)$: sigmoid function.

sIPM for LFR (sIPM-LFR)

- We use the sIPM as the regularization term.

1. Supervised learning

$$L(f \circ h) + \lambda d(\mathbb{P}_0^h, \mathbb{P}_1^h)$$

   - $\mathbb{P}_s^h$: distribution of $h(X, S)|S = s$
   - $f : \mathcal{Z} \to \mathbb{R}$: prediction function
   - L: classification loss function

2. Unsupervised learning

$$L_{recon}(f_D \circ h) + \lambda d(\mathbb{P}_0^h, \mathbb{P}_1^h)$$

   - $f_D : \mathcal{Z} \to \mathcal{X} \times \{0, 1\}$: decoder
   - L_{recon}: reconstruction loss function

Theoretical properties

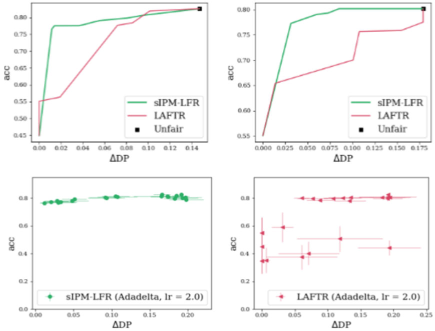
- We characterize the relationship between the level of the sIPM and the fairness of the final prediction model: $f \circ h$:

$$\sup_{f \in \mathcal{F}} DP_{\phi}(f \circ h) \leq \rho \left\{ d_{\mathcal{V}}(\mathbb{P}_0^h, \mathbb{P}_1^h) \right\},$$

  where DP_{ϕ} is the demographic parity and ρ is a non-decreasing function.
- Thus, we can achieve a fair prediction model by forcing representations to be fair with the sIPM.

Performance evaluation (Adult data set)

- Our method provides
  - better trade-offs between accuracy and fairness.
  - much more stable results.



danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
thank you
gracias
tapadh leat
mochchakkeram
bedankt
dziękuję
hvala
mauruuru
obrigado
sukriya
kop khun krap
go raibh maith agat
arigatō
dakujem
terima kasih
grazie
merci
мерси
감사합니다
xièxie
ευχαριστώ
dank je
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
mamnun
chokrane
djiere dieuf
tau
дякую
trugarez
shukriya
merce
dhanyavadagalu
diolch
takk
taiku
tanemirt
rahmet
sulpay
gratias ago
gracies
chnorakaloutioun
najis tuke
kam sah hamnida
rahmat
ありがとう
sagolun
didi madloba
mèsi
dėkui
sobodi
enkosi
bayarlalaa
gracie
köszönöm
dankon
aćiū
akun
nanni
nandri
kiitos
dankie
dhanyavad
faafetai lava
vinaka
спасибі
blagodaram
mersi
kia ora
barka
welalin
tack
spas