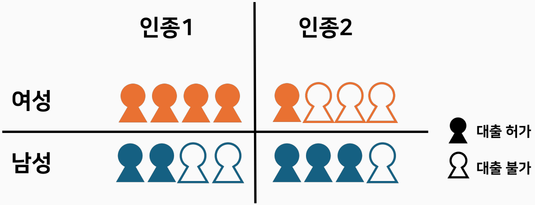
AICC 중간 발표 - Subgroup Fairness

Kyungseon Lee

August 26, 2025

Seoul National University

Motivation



- 하나의 민감속성으로만 공정성을 평가하면 다른 민감속성의 차별은 반영할 수 없음.
- 그림에서 성별에 대한 공정성은 도달했지만 (여성, 인종2) 그룹은 훨씬 적게 대출을 허가함.

Subgroup이란?

- 데이터: $(x_1, y_1, s_1), \ \dots \ , (x_n, y_n, s_n)$
  - $s_i = (s_{i1}, \dots, s_{iq})^\top$,는 q개의 민감속성벡터 $(s_{il} \in \{0, 1\})$
- Subgroup: 특정힌 민감속성 조합에 해당하는 데이터

$$\mathcal{D}_v = \{i : s_i = v\}, \ \ v \in \{0, 1\}^q$$

  - Number of subgroups $= 2^q$

Motivation

Subgroup fairness

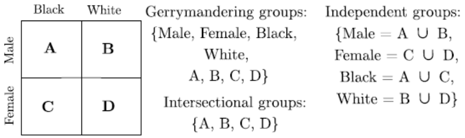


Figure 1: Definitions of group fairness [Yang *et al.*, 2020a].

- 모든 subgroup 단위에서 예측의 공정성을 보장하는 프레임워크
- 유한개의 데이터에서는 많은 subgroup이 아주 작은 관측치만 갖음 (data sparsity problem)

연구에서 추구하는 목표

- **대부분**의 subgroup에 대한 예측분포의 편차를 최소화
  ㄴ 뒤에서 설명 예정

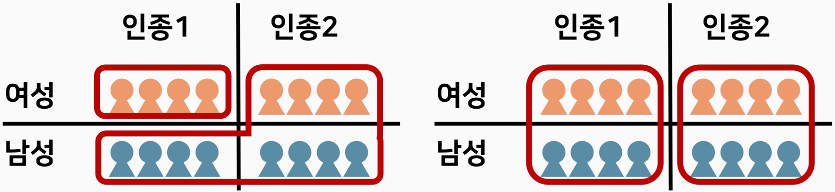
기존의 문제점. 샘플 수가 적은 subgroup의 경우, fairness level을 control하기 어려움.

- Tian et al., 2025
  - 부족한 subgroup의 sample을 imputation: 민감속성들의 샘플을 가중합하여 새로운 샘플을 생성함.
    - ▶ 문제: 만들어진 sample이 실제 sample을 대체한다고 보기 어려움.
- Molina Loiseau, 2022
  - Subgroup fairness를 marginal fairness로 upper bound 시켜서 marginal fairness를 조절함.
    - ▶ 문제: Subgroup fairness에 대한 조정이 잘 되지 않음.

Notation

- $S = \{0,1\}^q$: 모든 민감속성 조합을 원소로 가지는 집합
- $S_1, \ldots, S_M$: 사전에 정의된 M개의 S의 부분집합 (called "subgroup-subsets")
- $\mathcal{V} = \{S_1, \ldots, S_M\}$

아이디어: $\mathcal{V}$에 있는 subgroup-subset들에서만 예측값 분포를 비슷하게 만들자.



- 다음의 unfairness measure를 최소화 하는 예측모형 f중에서 가장 좋은 예측모형을 찾음

$$\mathrm{supIPM}(f) = \sup_{S \in \mathcal{V}} \mathrm{IPM}(\mathbb{P}_{f,S}, \mathbb{P}_{f,S^c})$$

이고 여기서 $\mathbb{P}_{f,S}$는 $\mathcal{D}_S = \cup_{s \in S}$에 포함되는 데이터만을 이용한 f의 확률분포

연구 아이디어1 - Partial subgroup fairness

$\mathcal{V}$를 선택하는 방법: 표본수가 일정 수준 이상인 모든 subgroup-subsets:

$$\text{All } S' \subset S \quad \text{such that } |\mathcal{D}_{S'}| \geq n_{low}$$

이고 여기서 n_{low}는 사전에 주어진 최소표본수

- 문제점: $|\mathcal{V}|$이 큰 경우 계산량이 폭증함
- 왜냐하면,

$$\mathrm{supIPM}(f) = \sup_{S \in \mathcal{V}} \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{z \sim \mathbb{P}_{f,S}}[g(z)] - \mathbb{E}_{z \sim \mathbb{P}_{f,S^c}}[g(z)] \right|$$

이고 여기서 $\mathcal{G}$는 discriminator class이기 때문에 $\mathrm{supIPM}(f)$를 계산하기 위해서는 $|\mathcal{V}|$만큼의 discriminator 를 학습해야함.

연구 아이디어2 - Doubly regressing approach

- 문제: supIPM 방법은 M개의 판별기가 필요하므로 computation cost가 높음.
- 해결 방법: 하나의 discriminator만으로 partial subgroup fairness를 달성할 수 있는 New and novel adversarial learning을 제안함

연구 아이디어2 - Doubly regressing approach

- $|\mathcal{S}| = M$이고 $\mathcal{V} = \{S_1, \ldots, S_M\}$
- $c_i \in \{0,1\}^M$이고 $c_{im} = \mathbb{I}(s_i \in S_m)$.
- Let

$$R^2(w, g : f) = 1 - \frac{\sum_{i=1}^n \left(\mathrm{w}^\top \mathrm{c}_i - g(f(\mathrm{x}_i, \mathrm{s}_i))\right)^2}{\sum_{i=1}^n \left(\mathrm{w}^\top \mathrm{c}_i - \mu_w\right)^2},$$

  where $\mu_w = \sum_{i=1}^n w^\top c_i / n$.
- 제안하는 new and novel unfairness measure

$$DR(f) = \sup_{g \in \mathcal{G}, \mathrm{w} \in \mathcal{W}} R^2(w, g : f)$$

  where $\mathcal{W}$ is the simplex on $\mathbb{R}^M$.
- $DR(f)$가 작은 예측모형 중에서 정확한 모형을 찾자!
- $DR(f)$를 계산하기 위해서는 하나의 discriminator (그리고 하나의 m차원 벡터 w)만 사용됨

Theorem (1)

There exists a positive constant $C > 0$ *such that for any given* $\delta \geq 0$ *and* $f \in \mathcal{F}$, $DR(f) \leq \delta$ *impies* $\sup IPM(f) \leq C\delta$.

Theorem (2)

There exists a positive constant $C' > 0$ *such that for any given* $\delta \geq 0$ *and* $f \in \mathcal{F}$, $\sup IPM(f) \leq \delta$ *impies* $DR(f) \leq C'\delta$.

- 결론: DR을 조절해서 $\sup IPM$을 조절할 수 있음.

Algorithm

- 목적함수

$$\min_{\theta} \min_{w,\phi} \frac{1}{n} \sum_{i=1}^{n} \mathrm{CE}(f_\theta(x_i, s_i), y_i) + \lambda DR(w, g_\phi : f_\theta)$$

  - λ: 분류–공정성 trade-off hyperparameter

Algorithm 1: Gradient descent를 사용하여 update

Input: 학습 데이터 $\{(x_i, s_i, c_i, y_i)\}_{i=1}^{n}$, 학습률 (분류, 판별기, 가중치)
$\eta_{\text{cls}}, \eta_g, \eta_w$, 반복 횟수 T, trade-off hyperparameter λ

Output: θ, ϕ, w

파라미터 초기화: $\theta \leftarrow \theta_0,\ \phi \leftarrow \phi_0,\ w \leftarrow w_0$

for $t = 1$ **to** T **do**

$\hat{y}_i \leftarrow f_\theta(x_i, s_i) \quad \forall i$

$L_{\text{cls}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \mathrm{CE}\big(\hat{y}_i,\ y_i\big)$

$\widehat{\mathrm{DR}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \big(w^\top c_i - g_\phi(\hat{y}_i)\big)^2$

$\phi \leftarrow \phi + \eta_g \nabla_\phi \widehat{\mathrm{DR}}$

$\tilde{w} \leftarrow w + \eta_w \nabla_w \widehat{\mathrm{DR}}, \quad w \leftarrow$ project $(\tilde{w})$ onto $\mathcal{W}$

$\theta \leftarrow \theta - \eta_{\text{cls}} \nabla_\theta L_{\text{cls}} - \lambda \cdot \eta_{\text{cls}} \nabla_\theta \widehat{\mathrm{DR}}$

return θ, ϕ, w

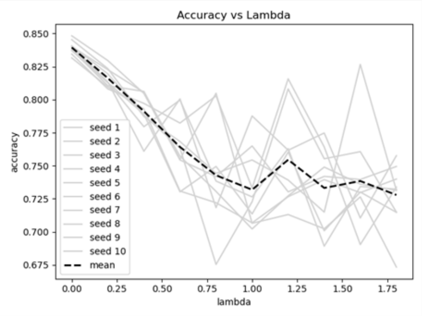
실험: Doubly regressing 만족 $\rightarrow$ *supIPM* $\leq C \cdot DR$

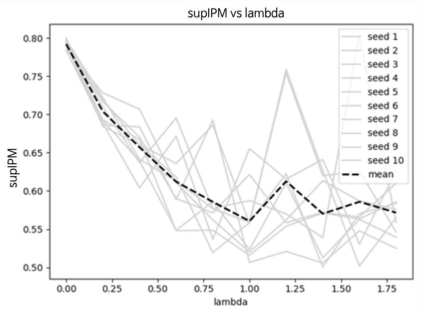
Theorem (1)

For any given $\delta \geq 0$, $DR(\phi) \leq \delta$ *implies* $\sup IPM(\phi) \leq C\delta$ *for some constant* $C > 0$.

- 데이터셋: UCI Adult 데이터셋 (14개 변수: 나이, 학력, 주당 근로시간 등)
- 민감 속성 실험:
    - 기본 실험: `race`, `sex`
    - 점진적 확장: 범주형 변수 8개 전체
- 평가 지표: 분류 성능 (Accuracy), Subgroup Fairness (supIPM)

Experiment





- λ 를 0부터 2까지 0.25씩 증가시키면서 실험한 그래프
- 첫 번째 그래프는 λ 가 상승함에 따라 Accuracy 하락을 보여주고 있음.
- 두 번째 그래프는 λ 가 상승함에 따라 supIPM의 하락(공정성 개선)을 확인함.

Future Work

- Representation learning.
- Fairness metric 확장(Demographic Parity -> Equalized Odds, Equality of Opportunity, ...).
- Writing papers