(NeurIPS 2022)
Bounding and Approximating Intersectional Fairness through Marginal Fairness

April 7, 2025

Seoul National University

Introduction

- Problem:

  In classification tasks with multiple protected attributes

  $$A = (A_1, ..., A_d),$$

  intersectional unfairness is defined as

  $$u^* = \sup_{(y,a,a') \in \mathcal{Y} \times \mathcal{A}^2} \left| \log \frac{\Pr(\hat{Y} = y \mid A = a)}{\Pr(\hat{Y} = y \mid A = a')} \right|$$

  but the number of subgroups grows exponentially (i.e., 2^d), leading to data sparsity issues,

  $\hat{Y} = h(X)$, h:classifier, X is features vector, Y is the label in $\mathcal{Y}$

Introduction

- Objective:

  To leverage **marginal unfairness**

$$u_k^* = \sup_{(y, a_k, a_k')} \left| log \frac{Pr(\hat{Y} = y | A_k = a_k)}{Pr(\hat{Y} = y | A_k = a_k')} \right|$$

  in order to better evaluate and understand μ^* via:

  - Bounding: Provide a probabilistic upper bound $\mu^* \leq \epsilon$
  - Approximation: Partition the protected attributes to approximate μ^*

Bounding

Probabilistic Upper Bound via Marginal Fairness

- Upper Bound:

$$u^* \leq \epsilon$$

where $\epsilon = 2\sqrt{2}\, s^* \sqrt{\delta} + \sup_{y \in \mathcal{Y}} \left\{ \sum_{k=1}^{d} \sup_{(a_k, a_k') \in \mathcal{A}_k^2} u_k(y, a_k, a_k') \right\}$,

with $Pr(U > \epsilon) \leq \delta$, $\epsilon \geq 0$, $\delta \in [0, 1]$ s.t. $U = u(\hat{Y}, A, A')$

  - $u_k(y, a_k, a_k^{'}) = \left| log \frac{Pr(\hat{Y}=Y|A_k=a_k)}{Pr(\hat{Y}=Y|A_k=a_k^{'})} \right|$
  - $s^* = (\sigma^{2/3} + \sigma_y^{2/3})^{3/2}$ with

$$L = \log \frac{p_A(A)}{\prod_{k=1}^{d} p_{A_k}(A_k)} \quad \text{and} \quad L_y = \log \frac{p_{A|\hat{Y}}(A \mid \hat{Y})}{\prod_{k=1}^{d} p_{A_k|\hat{Y}}(A_k \mid \hat{Y})}.$$

Approximation

Partitioning to Approximate Intersectional Fairness

- Direct estimation of $p_{\hat{Y}|A}$ is difficult due to data sparsity in the joint space.
- Partition A into groups using a partition $q = \{t_1, ..., t_m\}$ (partition of $\{1, ..., d\}$)
- For each group t, define the group-specific marginal unfairness:

$$u_t(y, a_t, a_t^{'}) = \left| log \frac{Pr(\hat{Y} = Y | A_t = a_t)}{Pr(\hat{Y} = Y | A_t = a_t^{'})} \right|$$

  where $A_t = (A_k)_{k \in t},\ t \subset \{1, 2, 3, ...m\}$
- The overall approximation is given by:

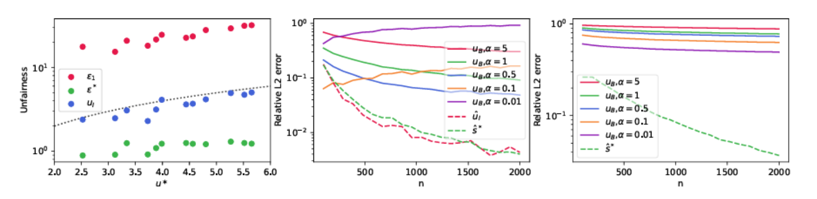
$$u_I(q) = \sup_{y \in \mathcal{Y}} \left\{ \sum_{t \in q} \sup_{(a_t, a_t') \in \mathcal{A}_t^2} u_t(y, a_t, a_t') \right\}.$$

Approximation

Algorithm 1 Greedy Partition Finder

input: Protected attributes data and occurrences of $\hat{Y}$
require: The partition of singletons is feasible
$q^* \leftarrow$ the partition of singletons
repeat
  $\mathcal{M} = \{\{t_1 \cup t_2\} \cup q^* \setminus (\{t_1\} \cup \{t_2\}), (t_1, t_2) \in q^{*2}, t_1 \neq t_2\}$
  $s^*_{\text{min}} \leftarrow +\infty$
  for q in $\mathcal{M}$ **do**
    if q is feasible and $s^*(q) < s^*_{\text{min}}$ **then**
      $(s^*_{\text{min}}, q^*) \leftarrow (s^*(q), q)$
    end if
  end for
until $\mathcal{M} = \emptyset$ or $s^*_{\text{min}} = \infty$ (Nothing possible to merge)
return: q^*

Experiment



Unfairness
10^1
10^0
ε_1
ε^*
u_l
2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
$u*$
Relative L2 error
10^0
10^{-1}
10^{-2}
$u_B, \alpha = 5$
$u_B, \alpha = 1$
$u_B, \alpha = 0.5$
$u_B, \alpha = 0.1$
$u_B, \alpha = 0.01$
$\hat{u}_l$
$\hat{s}^*$
500 1000 1500 2000
n
Relative L2 error
10^0
10^{-1}
$u_B, \alpha = 5$
$u_B, \alpha = 1$
$u_B, \alpha = 0.5$
$u_B, \alpha = 0.1$
$u_B, \alpha = 0.01$
$\hat{s}^*$
500 1000 1500 2000
n