Fairness without Demographics through Knowledge Distillation

Junyi Chai
chai28@purdue.edu

Taeuk Jang
jang141@purdue.edu

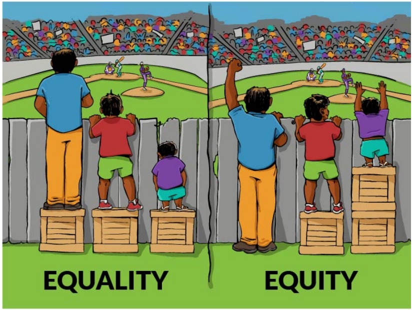
Xiaoqian Wang
joywang@purdue.edu

November 30, 2022

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Introduction

Machine learning has been widely adopted in real-world scenarios that has great social influence, and fairness in automatic decision-making systems has become an arising concern.



PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

In practice, however, due to legal or regulatory concerns, it is often infeasible to collect sensitive information, greatly limiting the usage of conventional methods on fairness.

Besides, in many real-world applications, we expect the decision-making systems to be fair w.r.t. multiple sensitive attributes.

Introduction

Current methods on fairness without demographics can be divided into two categories: fairness with proxy sensitive attribute and Max-Min fairness.

However, these methods can be too strict to improve fairness.

Our goal: knowledge distillation for improving fairness without accessing sensitive information.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Preliminary results

Label smoothing helps reduce disparity:

Label Assignment	Accuracy	Sensitive Attribute: Race		Sensitive Attribute: Gender	
		Dis. Impact	Eq. Odds	Dis. Impact	Eq. Odds
Binary	**85.19±0.17%**	11.80±0.48%	13.22±1.70%	17.74±0.56%	16.65±1.80%
Random	85.13±0.17%	12.21±1.37%	13.15±1.12%	17.64±0.95%	15.57±1.49%
Linear	85.14±0.25%	**10.75±1.15%**	**10.13±1.06%**	16.62±0.83%	**12.37±1.64%**
Softmax	**85.19±0.17%**	11.21±0.74%	10.67±1.24%	**16.37±0.58%**	13.34±1.89%

Table 1: Experimental results on new Adult dataset with race and gender as sensitive attribute, respectively. Fairness is evaluated using two metrics: Disparate Impact and Equalized Odds.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Method

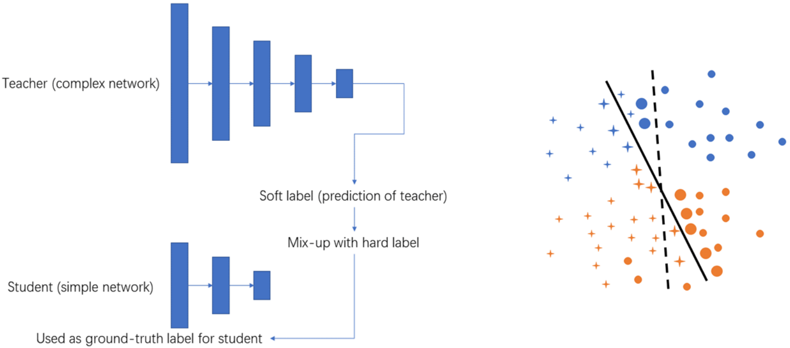


Figure: Demonstration of our knowledge distillation method.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Method

We consider two different mapping functions ϕ:

Softmax function:

$$\hat{y}_{ij}^{\mathrm{t}} = \phi_{\mathrm{softmax}}(z_i)_j = \frac{\exp\left(z_{ij}/T\right)}{\sum_k \exp\left(z_{ik}/T\right)}.$$

Linear function:

$$r_{ij'} = \frac{z_{ij'} - \min_{i \in S_j} z_{ij'}}{\max_{i \in S_j} z_{ij'} - \min_{i \in S_j} z_{ij'}} + 0.5,$$

$$r_{ik} = (1 - r_{ij'}) \frac{z_{ik}}{\sum_{k \neq \arg\max_l z_{il}} z_{ik}}, \forall k \neq \arg\max_l z_{il}.$$

Elmore Family School of Electrical and Computer Engineering

Theoretical analysis

Change in training loss:

$$
\begin{aligned}
L_{soft} - L_{hard} &= (y - y')\log(f(x)) - (y - y')\log(1 - f(x)) \\
&= \alpha(y - \hat{y}^t)\log\left(\frac{\frac{\exp(z_1)}{\exp(z_0)+\exp(z_1)}}{1 - \frac{\exp(z_1)}{\exp(z_0)+\exp(z_1)}}\right), \\
&= \alpha(y - \hat{y}^t)(z_1 - z_0).
\end{aligned}
$$

In terms of reweighing:

$$
w(x) = (1 - \tau(x))[\alpha\hat{y}^t + (1 - \alpha)y] + \tau(x).
$$

Hard samples **correctly** classified are assigned with **higher** weight.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Theoretical analysis

Theorem

Consider a classifier $f : \mathcal{X} \to [0,1]$ *for binary classification. Denote the classification loss as* $L_{soft} = -y' \log(f(x)) - (1-y') \log(1 - f(x))$ *with soft label* $y' = \alpha \hat{y}^t + (1-\alpha)y$*, where* $\hat{y}^t \in [0,1]$ *is the predicted label from teacher model,* $y \in \{0,1\}$ *is the binary label, and* α *is the balance parameter. The equal odds fairness metrics w.r.t. classifier* f *is upper bounded by* L_{soft}*.*

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Experiments

Method	Accuracy	Disparate impact	Equalized odds
Teacher	84.41%	20.27%	39.64%
Student (with hard label)	64.13±0.32%	23.27±2.43%	38.34±3.37%
DRO (Hashimoto et al., 2018)	62.67±0.73%	21.41±2.19%	30.43±3.24%
ARL (Lahoti et al., 2020)	63.23±0.47%	21.37±3.46%	29.46±1.74%
FairRF (Zhao et al., 2022)	63.26±0.83%	21.47±1.76%	25.67±2.63%
Student (with softmax label)	63.47±0.44%	19.52±2.46%	21.32±1.97%
Student (with linear label)	63.34±0.46%	20.27±2.34%	20.31±2.62%

Table 2: Results on COMPAS dataset with sensitive attribute *race*.

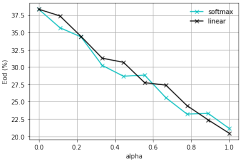
Method	Accuracy	Disparate impact	Equalized odds
Teacher	84.41%	19.42%	34.41%
Student (with hard label)	64.13±0.32%	19.17±2.33%	20.25±2.53%
DRO (Hashimoto et al., 2018)	62.67±0.73%	19.62±2.27%	18.75±2.18%
ARL (Lahoti et al., 2020)	63.23±0.47%	18.87±3.32%	19.14±2.56%
FairRF (Zhao et al., 2022)	63.26±0.83%	17.23±1.84%	18.74±2.21%
Student (with softmax label)	63.37±0.44%	16.63±1.67%	14.32±2.47%
Student (with linear label)	63.34±0.46%	16.14±1.83%	15.13±2.34%

Table 3: Results on COMPAS dataset with sensitive attribute *sex*.

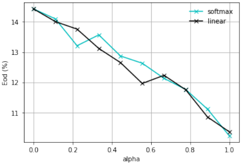
PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Experiments

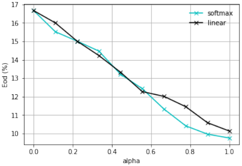
Parameter analysis:



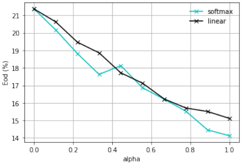
(a) COMPAS (race)



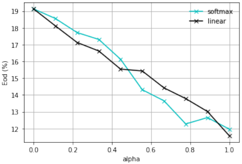
(b) New Adult (race)



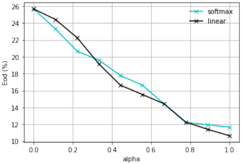
(c) CelebA, gender classification (age)



(d) COMPAS (sex)



(e) New Adult (gender)



(f) CelebA, attractiveness classification (gender)

Figure: Change of equalized odds as α varies.

Summary

Knowledge distillation for fairness without demographics

Effectiveness of label smoothing

Linear and softmax normalization

Connection between soft labelling and reweighing

Theoretical guarantee for fairness

PURDUE UNIVERSITY

Elmore Family School of Electrical and Computer Engineering

Thank you

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering