Blackbox Multiclass Fairness

Preston Putzel, and Scott Lee

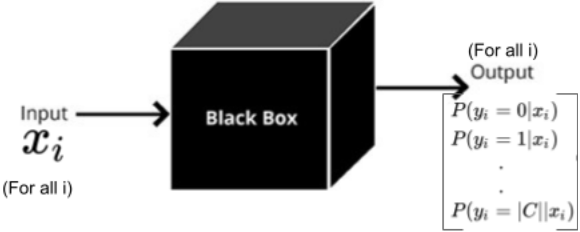
University of California, Irvine, and CDC

March 2022

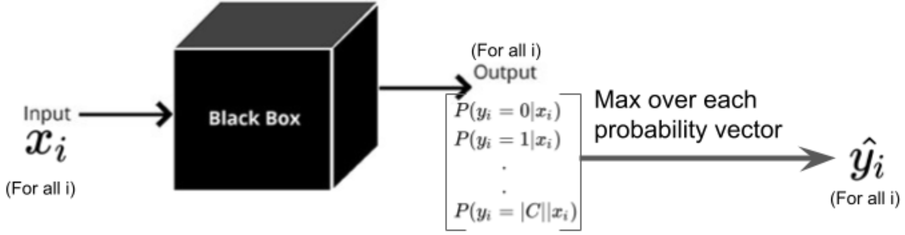
Motivating ML Use Case

- Data collection itself **is not** necessarily an ‘objective process’ [1]
- Ex: COMPASS criminal recidivism prediction. [2]

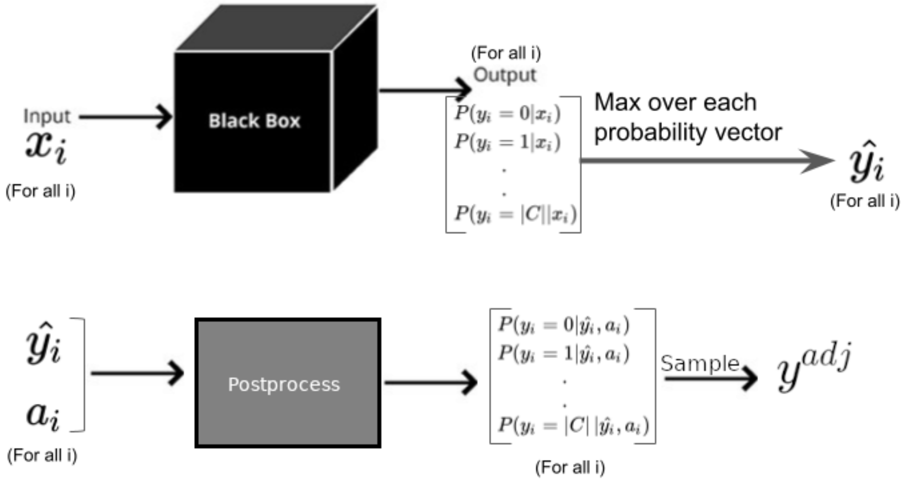
Our Approach: An Extension of Hardt 2016



Our Approach: An Extension of Hardt 2016



Our Approach: An Extension of Hardt 2016



The Group Conditional Confusion Matrix

A=0	Y		
Y^{adj}	$Pr(Y^{adj} = 0 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 0 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 0 \mid Y = 2, A = 0)$
	$Pr(Y^{adj} = 1 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 1 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 1 \mid Y = 2, A = 0)$
	$Pr(Y^{adj} = 2 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 2 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 2 \mid Y = 2, A = 0)$

The Group Conditional Confusion Matrix

A=0	Y		
Y^{adj}	$Pr(Y^{adj} = 0 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 0 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 0 \mid Y = 2, A = 0)$
	$Pr(Y^{adj} = 1 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 1 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 1 \mid Y = 2, A = 0)$
	$Pr(Y^{adj} = 2 \mid Y = 0, A = 0)$	$Pr(Y^{adj} = 2 \mid Y = 1, A = 0)$	$Pr(Y^{adj} = 2 \mid Y = 2, A = 0)$

A=1	Y		
Y^{adj}	$Pr(Y^{adj} = 0 \mid Y = 0, A = 1)$	$Pr(Y^{adj} = 0 \mid Y = 1, A = 1)$	$Pr(Y^{adj} = 0 \mid Y = 2, A = 1)$
	$Pr(Y^{adj} = 1 \mid Y = 0, A = 1)$	$Pr(Y^{adj} = 1 \mid Y = 1, A = 1)$	$Pr(Y^{adj} = 1 \mid Y = 2, A = 1)$
	$Pr(Y^{adj} = 2 \mid Y = 0, A = 1)$	$Pr(Y^{adj} = 2 \mid Y = 1, A = 1)$	$Pr(Y^{adj} = 2 \mid Y = 2, A = 1)$

Types of Multiclass Fairness

- Term by Term Equality of Odds
- Classwise Equality of Odds
  - Diagonals and False Detection Rates: $Pr(Y^{adj} = i | Y \neq i, A = a)$
- Multiclass Equality of Opportunity
- Demographic Parity: $Pr(Y^{adj} = i | A = a)$

A=0	Y		
Y^{adj}	$Pr(Y^{adj} = 0 \| Y = 0, A = 0)$	$Pr(Y^{adj} = 0 \| Y = 1, A = 0)$	$Pr(Y^{adj} = 0 \| Y = 2, A = 0)$
	$Pr(Y^{adj} = 1 \| Y = 0, A = 0)$	$Pr(Y^{adj} = 1 \| Y = 1, A = 0)$	$Pr(Y^{adj} = 1 \| Y = 2, A = 0)$
	$Pr(Y^{adj} = 2 \| Y = 0, A = 0)$	$Pr(Y^{adj} = 2 \| Y = 1, A = 0)$	$Pr(Y^{adj} = 2 \| Y = 2, A = 0)$

The Linear Program

- All of the previous equalities under mild assumptions can be written as linear constraints on $Pr(Y^{adj}|Y, A)$
- We minimize a weighted sum of mismatch errors between Y^{adj} and Y:

$$\sum_{a \in A} \sum_{i=1}^{|C|} \sum_{j \neq i} Pr(Y^{\text{adj}} = i, Y = j, A = a) l(i, j, a)$$

The Linear Program

- All of the previous equalities under mild assumptions can be written as linear constraints on $Pr(Y^{adj}|Y, A)$
- We minimize a weighted sum of mismatch errors between Y^{adj} and Y:

$$\sum_{a \in A} \sum_{i=1}^{|C|} \sum_{j \neq i} \underbrace{Pr(Y^{\text{adj}} = i, Y = j, A = a)}_{\text{Joint Pr of error}} \underbrace{l(i, j, a)}_{\text{Weights}}$$

Synthetic Results

Experiments with $|A| = 3$

Hyperparameter	**Level**	**Change in Acc**	**Change in TDR**
Group Balance	No Minority One Slight Minority One Strong Minority Two Slight Minorities Two Strong Minorities		
Class Balance	Balanced One Rare Two Rare		
Pred Bias	Low One Low Two Medium One Medium Two High One High Two	-	

Synthetic Results

Experiments with $|A| = 3$

Hyperparameter	**Level**	**Change in Acc**	**Change in TDR**
Group Balance	No Minority	-	-
	One Slight Minority	-0.03	-0.02
	One Strong Minority	-0.04	-0.01
	Two Slight Minorities	-0.05	-0.02
	Two Strong Minorities	-0.07	-0.01
Class Balance	Balanced		
	One Rare		
	Two Rare		
Pred Bias	Low One		
	Low Two		
	Medium One	-	
	Medium Two		
	High One		
	High Two		

Synthetic Results

Experiments with $|A| = 3$

Hyperparameter	**Level**	**Change in Acc**	**Change in TDR**
Group Balance	No Minority	-	-
	One Slight Minority	-0.03	-0.02
	One Strong Minority	-0.04	-0.01
	Two Slight Minorities	-0.05	-0.02
	Two Strong Minorities	-0.07	-0.01
Class Balance	Balanced	-	-
	One Rare	0.02	-0.04
	Two Rare	0.07	-0.18
Pred Bias	Low One		
	Low Two		
	Medium One	-	
	Medium Two		
	High One		
	High Two		

Synthetic Results

Experiments with $|A| = 3$

Hyperparameter	**Level**	**Change in Acc**	**Change in TDR**
Group Balance	No Minority	-	-
	One Slight Minority	-0.03	-0.02
	One Strong Minority	-0.04	-0.01
	Two Slight Minorities	-0.05	-0.02
	Two Strong Minorities	-0.07	-0.01
Class Balance	Balanced	-	-
	One Rare	0.02	-0.04
	Two Rare	0.07	-0.18
Pred Bias	Low One	-	-
	Low Two	0.00	-0.00
	Medium One	-0.06	-0.06
	Medium Two	-0.04	-0.06
	High One	-0.18	-0.16
	High Two	-0.15	-0.13

Real-World Data Results

In-Sample Results

Dataset (N)	# Terms in $\mathbf{P^a}$	% change Old Acc $\rightarrow$ New Acc	% change Pre $\rightarrow$ Post-Adj Disparity
Bar (N=22406)	18	-1% (88 % $\rightarrow$ 88%)	-100% (0.11 $\rightarrow$ 0.00)
Parkinsons (N=5875)	18	-2% (93% $\rightarrow$ 91%)	-100% (0.04 $\rightarrow$ 0.00)
Cannabis (N=1885)	18	-4% (74% $\rightarrow$ 71%)	-100% (0.07 $\rightarrow$ 0.00)
Obesity (N=1490)	50	-7% (78% $\rightarrow$ 73%)	-100% (0.05 $\rightarrow$ 0.00)

Real-World Data Results

In-Sample Results

Dataset (N)	# Terms in **P**[a]	% change Old Acc → New Acc	% change Pre → Post-Adj Disparity
Bar (N=22406)	18	-1% (88 % → 88%)	-100% (0.11 → 0.00)
Parkinsons (N=5875)	18	-2% (93% → 91%)	-100% (0.04 → 0.00)
Cannabis (N=1885)	18	-4% (74% → 71%)	-100% (0.07 → 0.00)
Obesity (N=1490)	50	-7% (78% → 73%)	-100% (0.05 → 0.00)

Out of Sample Results

Dataset (N)	# Terms in **P**[a]	% change Old Acc → New Acc	% change Pre → Post-Adj Disparity
Bar (N=22406)	18	-6% (88 % → 83%)	-95% (0.11 → 0.01)
Parkinsons (N=5875)	18	-12% (93% → 82%)	33% (0.04 → 0.05)
Cannabis (N=1885)	18	-18% (74% → 61%)	124% (0.07 → 0.16)
Obesity (N=1490)	50	-47% (78% → 41%)	45% (0.05 → 0.07)

Key Takeaways

- ML methods can propagate dataset bias
- Blackbox post-processing addreses fairness by updating model outputs for fairness
- Our linear-program based approach works well when given enough training data to reliably estimate probabilities empirically, but fails on out-of-sample data otherwise.

Acknowledgments

- Thanks to the support of my PhD advisor Padhraic Smyth, and ORISE supervisor Chad Heilig.
- This project was supported in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and CDC.
- The views presented are the author's own and do not necessarily represent an official position of the Centers for Disease Control Prevention.

Bibliography

[1] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 54(6), 2021
[2] Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COM- PAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc.