UCLA

STAR AI RESEARCH LAB UCLA

Group Fairness by Probabilistic Modeling with Latent Fair Decisions

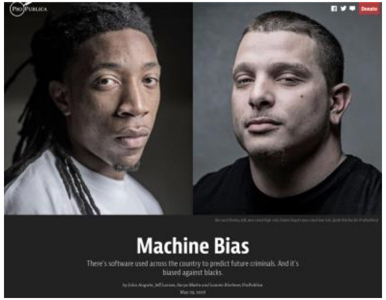
YooJung Choi, Meihua Dang, Guy Van den Broeck

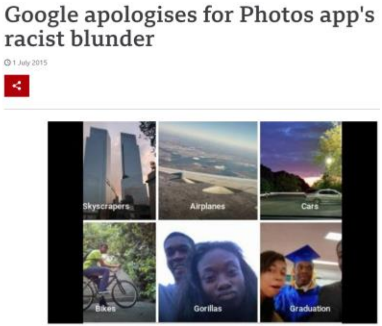
AAAI 2021

Why algorithmic fairness

AI systems are increasingly being adopted in areas with personal and societal impact.

Societal bias may be perpetuated and amplified by AI/ML models





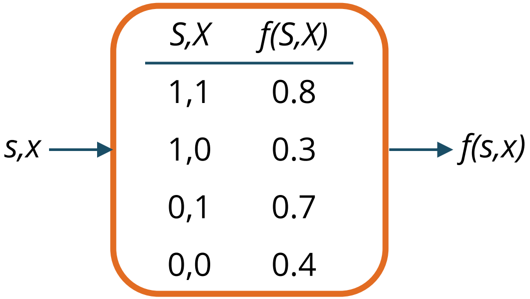
Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.



Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.

$s,x \rightarrow$

S,X	$f(S,X)$
1,1	0.8
1,0	0.3
0,1	0.7
0,0	0.4

$\rightarrow f(s,x)$

S,X	$P_{data}(X\|S)$
1,1	0.7
1,0	0.3
0,1	0.4
0,0	0.6

Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.

s,x → $f(s,x)$

S,X	$f(S,X)$
1,1	0.8
1,0	0.3
0,1	0.7
0,0	0.4

S,X	$P_{data}(X\|S)$
1,1	0.7
1,0	0.3
0,1	0.4
0,0	0.6

$$\mathbb{E}_{P_{data}}[f|S=1] - \mathbb{E}_{P_{data}}[f|S=0] = 0.13$$

f does not satisfy demographic parity!

Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.

s,x → *f(s,x)*

S,X	*f(S,X)*
1,1	0.8
1,0	0.3
0,1	0.7
0,0	0.4

S,X	*Q(X|S)*
1,1	0.5
1,0	0.5
0,1	0.5
0,0	0.5

Motivation

Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.

s,x →

S,X	$f(S,X)$
1,1	0.8
1,0	0.3
0,1	0.7
0,0	0.4

→ $f(s,x)$

S,X	$Q(X \mid S)$
1,1	0.5
1,0	0.5
0,1	0.5
0,0	0.5

f is considered fair with respect to *Q*

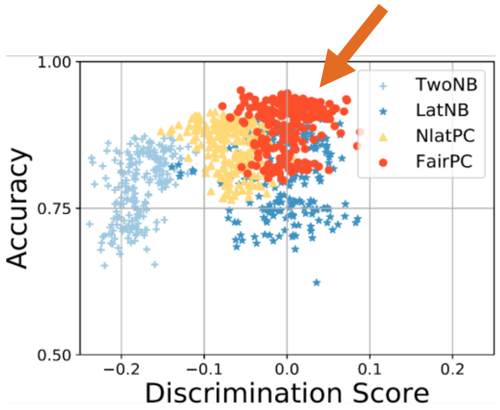
Motivation

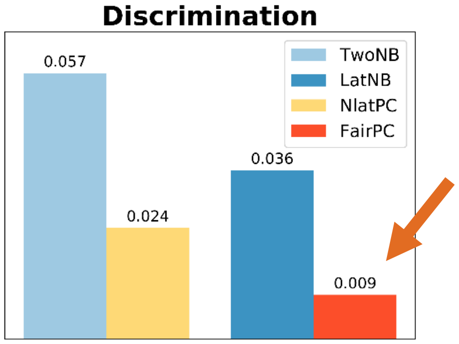
Challenge #1: When learning classifiers, the labels may have historical bias or be proxies to the true target variable.

Challenge #2: Fairness guarantees hold only if the real-world distribution is captured.

Our contribution: *address both challenges using probabilistic modeling with latent fair decisions*

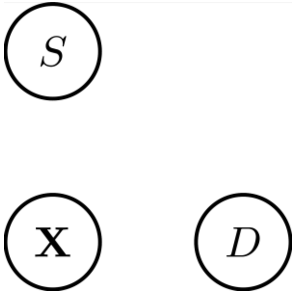
Spoiler alert





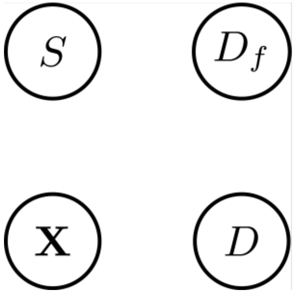
Results: *closely modeling the observed data distribution and bias mechanism leads to competitive classification accuracy and better fairness guarantees.*

Latent fair decisions



Sensitive attribute S, set of features $\boldsymbol{X}$, label D

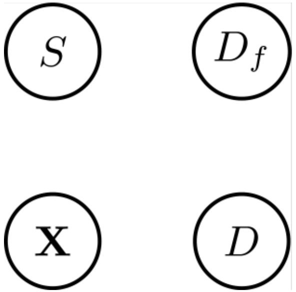
Latent fair decisions



Sensitive attribute S, set of features $\boldsymbol{X}$, label D

Latent variable D_f to represent the hidden, fair label.

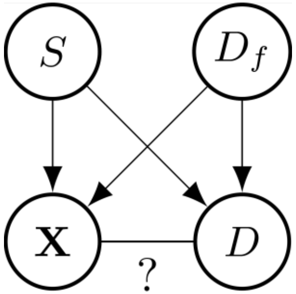
Latent fair decisions



Assumption #1: D_f satisfies demographic parity.

$$\mathbb{E}_P[f(\mathbf{X}, S) \mid S = 1] = \mathbb{E}_P[f(\mathbf{X}, S) \mid S = 0]$$

Latent fair decisions

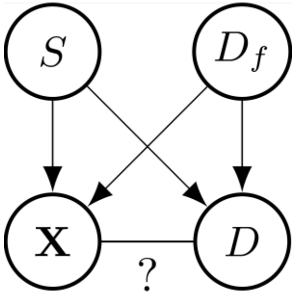


Assumption #1: D_f satisfies demographic parity.

$$\mathbb{E}_P[f(\mathbf{X}, S) \mid S = 1] = \mathbb{E}_P[f(\mathbf{X}, S) \mid S = 0]$$

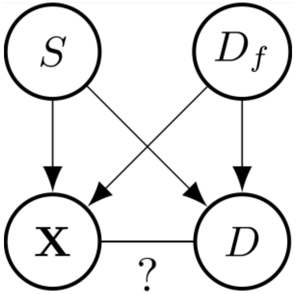
$\Rightarrow D_f \perp S$ for probabilistic classifier $f(\mathbf{X}, S) = P(D_f | \mathbf{X}, S)$

Latent fair decisions



Assumption #2: data provides information about D_f

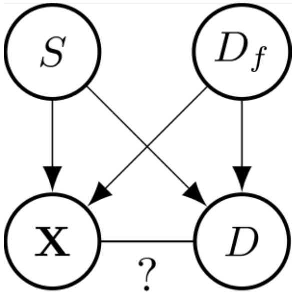
Latent fair decisions



S,X,D	*P(S,X,D)*
1,1,1	0.2
1,1,0	0.1
⋮	⋮
0,0,0	0.3

Assumption #2: data provides information about D_f

Latent fair decisions

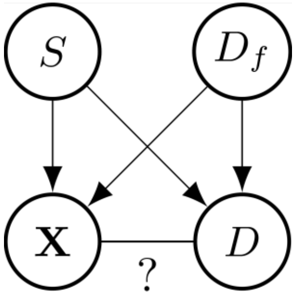


Assumption #2: data provides information about D_f

S,X,D	P(S,X,D)
1,1,1	0.2
1,1,0	0.1
⋮	⋮
0,0,0	0.3

S,X,D	$P(S,X,D,D_f{=}1)$	$P(S,X,D,D_f{=}0)$
1,1,1	0.15	0.05
1,1,0	0.05	0.05
⋮	⋮	⋮
0,0,0	0.1	0.2

Latent fair decisions



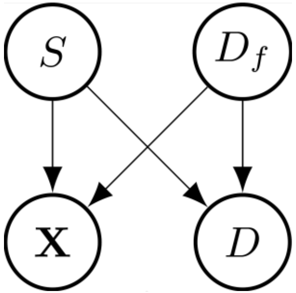
Assumption #2: data provides information about D_f

S,X,D	P(S,X,D)
1,1,1	0.2
1,1,0	0.1
⋮	⋮
0,0,0	0.3

S,X,D	$P(S,X,D,D_f{=}1)$	$P(S,X,D,D_f{=}0)$
1,1,1	0.15	0.05
1,1,0	0.05	0.05
⋮	⋮	⋮
0,0,0	0.1	0.2

S,X,D	$P(S,X,D,D_f{=}1)$	$P(S,X,D,D_f{=}0)$
1,1,1	0.2	0
1,1,0	0.1	0
⋮	⋮	⋮
0,0,0	0.3	0

Latent fair decisions



Assumption #2: data provides information about D_f

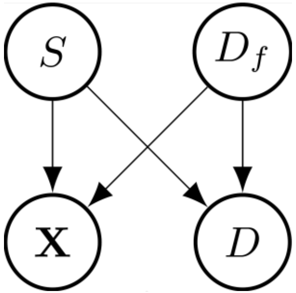
$\Rightarrow D \perp \boldsymbol{X} \mid D_f, S$ to model dependence to D_f

S,X,D	P(S,X,D)
1,1,1	0.2
1,1,0	0.1
⋮	⋮
0,0,0	0.3

S,X,D	$P(S,X,D,D_f\text{=}1)$	$P(S,X,D,D_f\text{=}0)$
1,1,1	0.15	0.05
1,1,0	0.05	0.05
⋮	⋮	⋮
0,0,0	0.1	0.2

S,X,D	$P(S,X,D,D_f\text{=}1)$	$P(S,X,D,D_f\text{=}0)$
1,1,1	0.2	0
1,1,0	0.1	0
⋮	⋮	⋮
0,0,0	0.3	0

Latent fair decisions



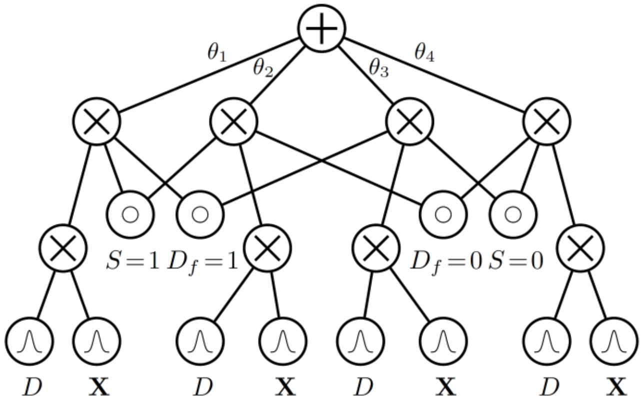
Learn the distribution that best fits the data while ensuring $D_f \perp S$ and $D \perp \boldsymbol{X} \mid D_f, S$.

Probabilistic circuits

Recursively define distributions using
sums, *products*, and *univariate distributions.*

$$\Pr_n(\mathbf{x}) = \begin{cases} f_n(\mathbf{x}) & \text{if } n \text{ is a leaf} \\ \prod_{c \in \mathsf{ch}(n)} \Pr_c(\mathbf{x}) & \text{if } n \text{ is a product} \\ \sum_{c \in \mathsf{ch}(n)} \theta_{n,c} \Pr_c(\mathbf{x}) & \text{if } n \text{ is a sum} \end{cases}$$

- Expressive: closely model the data
- Tractable: efficiently compute conditionals
- Structure encodes independencies



Learning fair probabilistic circuits

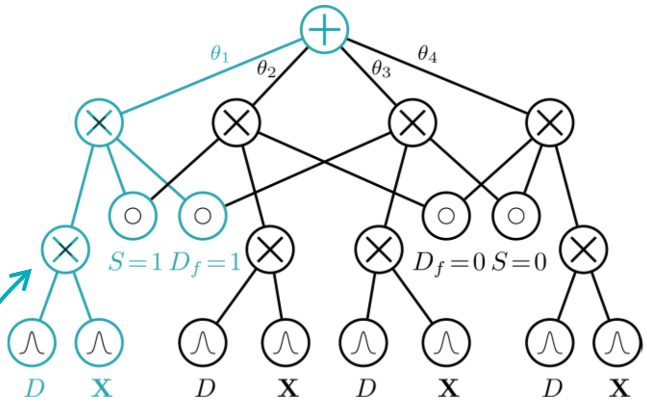
$P(D, \boldsymbol{X}, D_f = 1, S = 1)$

Parameters are conditional probabilities

$\theta_1 = P(D_f = 1, S = 1)$

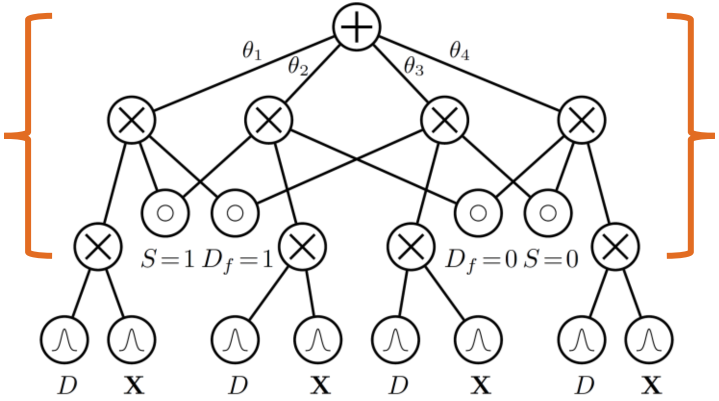
Structure encodes conditional independence

$P(D, \boldsymbol{X} \mid D_f, S) = P(D \mid D_f, S) \cdot P(\boldsymbol{X} \mid D_f, S)$



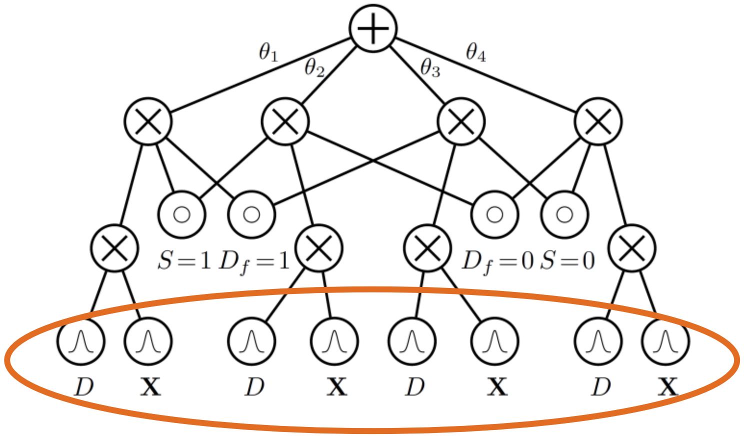
Learning fair probabilistic circuits

- Encode independence assumptions by fixing top-level structure and parameter tying.



Learning fair probabilistic circuits

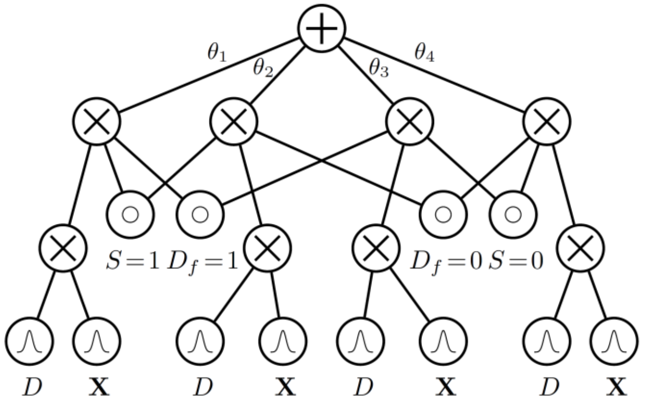
- Encode independence assumptions by fixing top-level structure and parameter tying.
- Learn the structure for *X* from data.



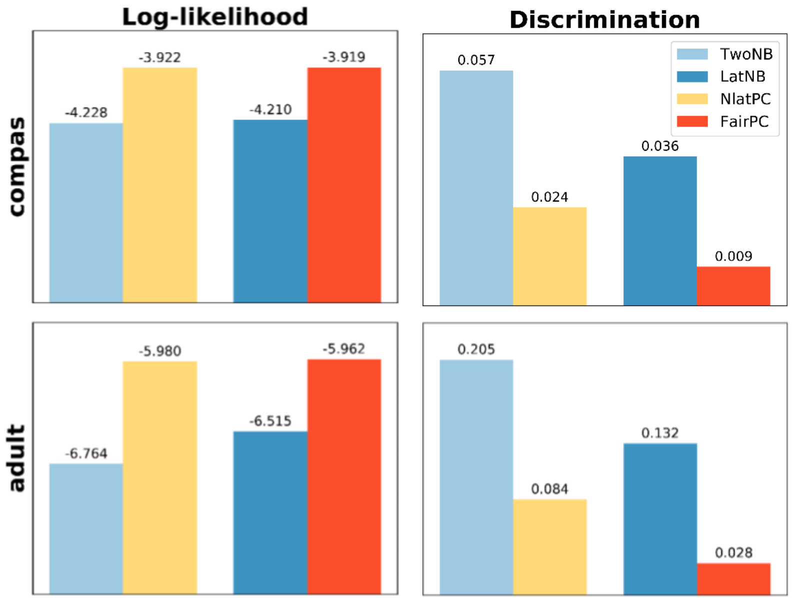
Learning fair probabilistic circuits

- Encode independence assumptions by fixing top-level structure and parameter tying.
- Learn the structure for *X* from data.
- Learn the parameters via EM:

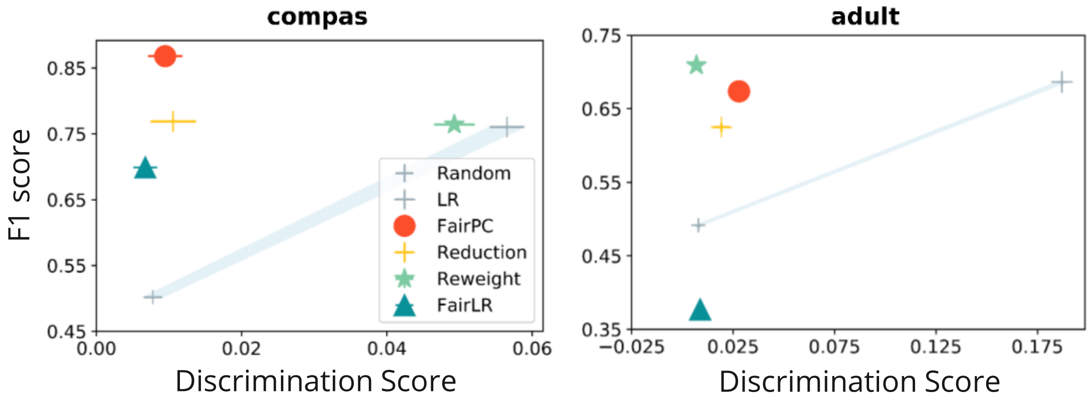
$$\theta_{n,c}^{(\text{new})} = \text{EF}_{\mathcal{D},\theta}(n,c) / \sum_{c \in \mathsf{ch}(n)} \text{EF}_{\mathcal{D},\theta}(n,c).$$



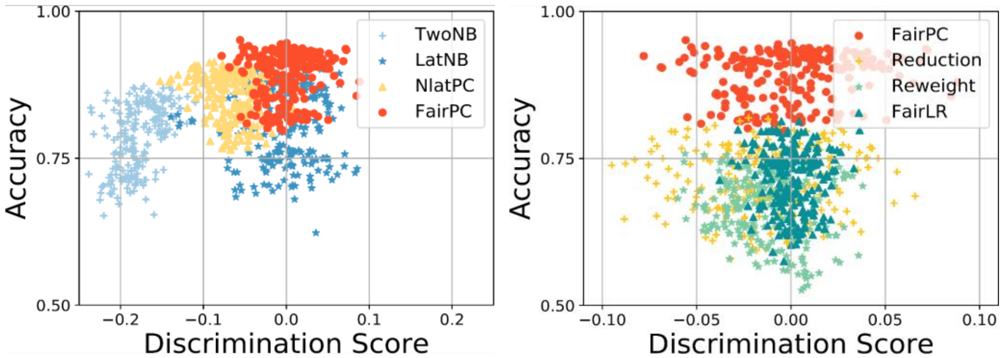
Experiments: modeling the data



Experiments: similarity to observed labels



Experiments: synthetic data



Conclusion

1. Latent variable approach can learn ***fair decisions*** while explaining the data with ***biased labels***.
2. Closely modeling the data leads to ***lower discrimination scores***.
3. Latent decision variables from FairPC retain ***high similarity*** *to observed labels.*