Definitions of Fairness, Inherent Tradeoffs + Impossibilities

Today

1. Definitions of fairness
   - statistical parity
   - predictive rate parity
   - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness + Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Running Example

COMPAS : risk assessment program

Propublica concluded that COMPAS is biased:

The likelihood of blacks predicted to readivate given that they did Not is > likelihood for whites

	Humans		COMPAS	
	Black %	White %	Black %	White %
Accuracy*	68.2	67.6	64.9	65.7
False Positives	37.1	27.2	40.4	25.4
False Negatives	29.2	40.3	30.9	47.9

COMPAS DATA

- Recidivism rate for blacks 51%
- Recidivism rate for whites 39%

IS COMPAS BIASED?

Today

1. Definitions of fairness
   - statistical parity
   - predictive rate parity
   - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness & Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Definitions/Notation

$x \in \mathcal{U}$ feature vector [Typically $\mathcal{U} = \mathbb{R}^d$ or discretized version]

$y \in \{0,1\}$ actual value (we are trying to predict)

Underlying distribution is pair of r.v.'s (X, Y)

Classifier: maps x to $\hat{y} = f(x)$.

COMPAS EXAMPLE:

x : feature vector of offender

$y=1$: offender did recidivate, $y=0$ did not

$\hat{y}$: prediction for x

Probability Space, RV's

$\mathcal{U}$ $(= \mathbb{R}^d)$: set of all possible feature vectors (finite descriptions of a person — d attributes)

$p(u)$: probability (over all individuals in population) of having feature vector u

$X = p(\mathcal{U})$: random variable

(X, Y) : random variable

Confusion Matrix

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	TNR $\Pr[\hat{y}=0 \mid y=0]$	FPR $\Pr[\hat{y}=1 \mid y=0]$
$y=1$	FNR $\Pr[\hat{y}=0 \mid y=1]$	TPR $\Pr[\hat{y}=1 \mid y=1]$

TNP : true Negative rate

FNR : false Negative rate

FPR : False positive rate

TPR : true positive rate

Definitions/Notation

$x \in \mathbb{R}^d$ feature vector (may include A)

$y \in \{0,1\}$ actual value (we are trying to predict)

Underlying distribution is pair of r.v.'s (X, Y)

Classifier: maps x to $\hat{y} = f(x)$.

Sensitive variable: $A \in \{0,1\}$

Joint distribution $(X, Y, A, \hat{Y})$

Example: COMPAS

X: vector about offender

Y: whether offender will recidivate $(Y=1)$

A: black $(A=1)$ or white $(A=0)$

$\hat{y} = f(x)$: predicted value of y

Confusion Matrix A=1

	$\hat{y}=0$	$\hat{y}=1$
$y=0$		40.4
$y=1$	30.9	

Confusion Matrix A=0

	$\hat{y}=0$	$\hat{y}=1$
$y=0$		25.4
$y=1$	47.9	

	Humans		COMPAS	
	Black %	White %	Black %	White %
Accuracy*	68.2	67.6	64.9	65.7
False Positives	37.1	27.2	40.4	25.4
False Negatives	29.2	40.3	30.9	47.9

Definitions / Notation

$R = f(X)$

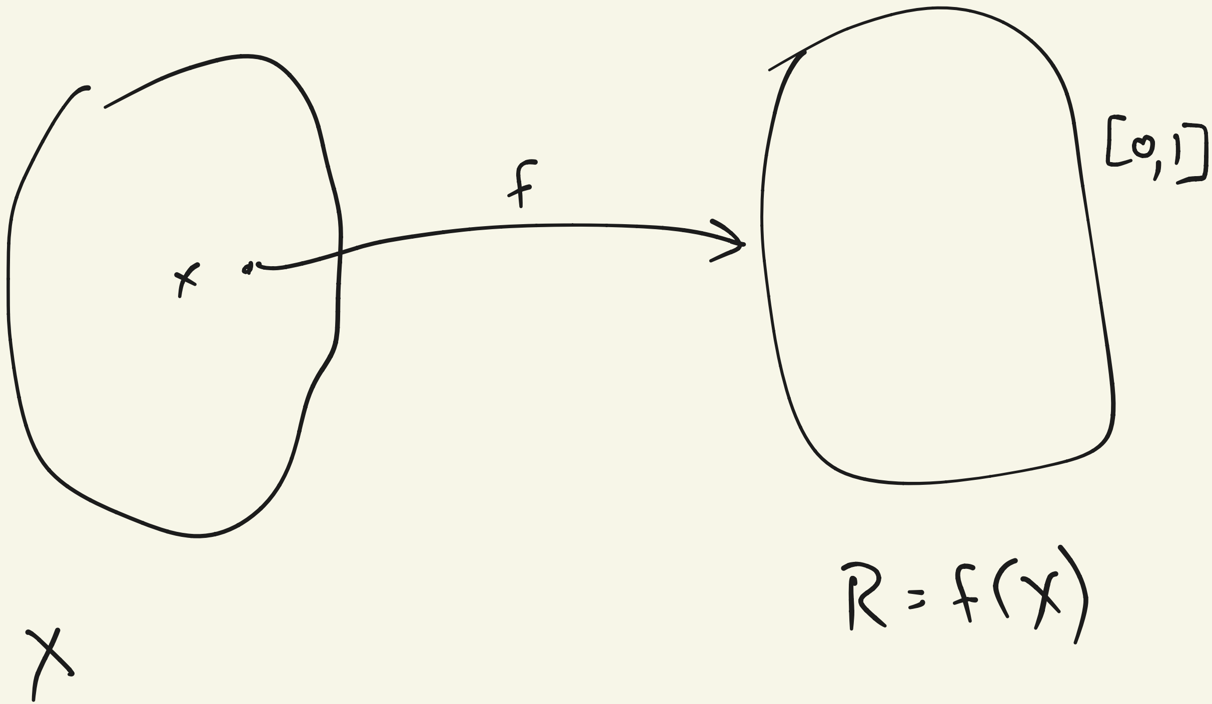
Classification often solved by first solving a regression problem to summarize data by a score, $f(x) \in \mathbb{R}$ (we assume $\in [0,1]$).

Natural score function: $f(x) = \mathbb{E}[Y \mid x]$

Score to $\hat{y}$: Pick threshold t

$\hat{Y} = 1$ iff $f(x) \geq t$

R generalizes $\hat{Y}$ – so from now on can think of $\hat{Y}$ as special case of R



x
f
$[0,1]$
X
$R = f(X)$

Definitions / Notation

1. Statistical Parity / group Parity / Independence

$R \perp A \quad (\hat{Y} \perp A) \qquad \Pr[R \mid A] = \Pr[R]$

2. Predictive Rate Parity / Sufficiency

$Y \perp A \mid R \qquad Y \perp A \mid \hat{Y} \qquad \Pr[Y \mid A, R] = \Pr[Y \mid R]$

3. Equalized odds / Separation

$R \perp A \mid Y \qquad \hat{Y} \perp A \mid Y \qquad \Pr[R \mid A, Y] = \Pr[R \mid Y]$

Definitions / Notation

1. Statistical Parity / group Parity

$$\forall b \in \{0,1\} \quad \Pr[\hat{y}=b \mid A=0] = \Pr[\hat{y}=b \mid A=1]$$

2. Predictive Rate Parity

$$(*) \quad \left\{ \forall b, b' \in \{0,1\} \quad \Pr[y=b \mid \hat{y}=b', A=0] = \Pr[y=b \mid \hat{y}=b', A=1] \right.$$

$\Updownarrow$ Equivalent to

$$(**) \quad \begin{cases} \Pr[y=1 \mid \hat{y}=1, A=0] = \Pr[y=1 \mid \hat{y}=1, A=1] \text{ and} \\ \Pr[y=0 \mid \hat{y}=0, A=0] = \Pr[y=0 \mid \hat{y}=0, A=1] \end{cases}$$

Definitions / Notation

1. Statistical Parity / group Parity

$$\forall b \in \{0,1\} \quad \Pr[\hat{y}=b \mid A=0] = \Pr[\hat{y}=b \mid A=1]$$

2. Predictive Rate Parity

$$\forall b, b' \in \{0,1\} \quad \Pr[y=b \mid \hat{y}=b', A=0] = \Pr[y=b \mid \hat{y}=b', A=1]$$

3. Equalized Odds

$$\Pr[\hat{y}=0 \mid y=1, A=0] = \Pr[\hat{y}=0 \mid y=1, A=0]$$

$$\Pr[\hat{y}=1 \mid y=0, A=0] = \Pr[\hat{y}=1 \mid y=0, A=0]$$

Definitions / Notation

1. Statistical Parity / group Parity

$$\forall b \in \{0,1\} \quad Pr[\hat{y}=b \mid A=0] = Pr(\hat{y}=b \mid A=1]$$

2. Predictive Rate Parity (PRP)

$$\forall b, b' \in \{0,1\} \quad Pr[y=b \mid \hat{y}=b', A=0] = Pr[y=b \mid \hat{y}=b', A=1]$$

3. Equalized Odds

FNR
$$Pr[\hat{y}=0 \mid y=1, A=0] = Pr[\hat{y}=0 \mid y=1, A=0]$$

FPR
$$Pr[\hat{y}=1 \mid y=0, A=0] = Pr[\hat{y}=1 \mid y=0, A=0]$$

Today

1. Definitions of fairness
   - statistical parity
   - predictive rate parity
   - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness & Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Impossibility Theorem

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

Impossibility Theorem (Indep vs sufficiency)

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

1 - 2 Statistical parity & predictive rate parity are mutually exclusive unless $A \perp Y$

$$A \perp \hat{Y} \text{ and } A \perp Y \mid \hat{Y} \Rightarrow A \perp Y$$

Impossibility Theorem

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

① - ② Statistical parity & predictive rate parity are mutually exclusive unless $A \perp Y$

$$A \perp \hat{y} \text{ and } A \perp y \mid \hat{y} \Rightarrow A \perp (y, \hat{y}) \Rightarrow A \perp Y$$

<u>Pf</u> $A \perp y \mid \hat{y}$: $\Pr[A \mid y, \hat{y}] = \Pr[A \mid \hat{y}]$

$A \perp \hat{y}$: $\Pr[A \mid \hat{y}] = \Pr[A]$

So $$\Pr(A \mid y] = \sum_b \Pr[\hat{y} = b] \Pr[A \mid y, \hat{y} = b] = \sum_b \Pr(\hat{y} = b) \Pr[A \mid \hat{y} = b]$$

$$= \Pr(A)$$

Impossibility Theorem (Separation vs Sufficiency)

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

② - ③ Predictive Rate parity and Equalized odds are mutually exclusive unless $A \perp Y$

$$A \perp \hat{Y} \mid Y \quad \text{and} \quad A \perp Y \mid \hat{Y} \Rightarrow A \perp (\hat{Y}, Y) \Rightarrow A \perp Y$$

$$\Pr(A \mid \hat{Y}, Y) = \Pr(A \mid \hat{Y}) \quad \text{and}$$

$$\Pr(A \mid \hat{Y}, Y) = \Pr(A \mid Y)$$

$$\Pr(A \mid \hat{y}=1) = \Pr(y=0)\Pr(A \mid \hat{y}=1, y=0) + \Pr(y=1)\Pr(A \mid \hat{y}=1, y=1)$$

$$= \Pr(y=0)\Pr(A \mid y=0) \Pr(y=1)\Pr(A \mid y=1)$$

$$= \Pr(A)$$

$$\Pr(A \mid \hat{y}=1, y=0) = \Pr(A \mid y=0)$$

$$\Pr(A \mid \hat{y}=1, y=1) = \Pr(A \mid y=1)$$

Impossibility Theorem (Indep vs Separation)

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

① - ③ Statistical Parity & Equalized Odds (* For binary Y*)

are mutually exclusive unless $A \perp Y$

or $\hat{Y} \perp Y$

$$A \perp \hat{Y} \text{ and } A \perp \hat{Y} \mid Y \Rightarrow A \perp Y \text{ or } \hat{Y} \perp Y$$

Impossibility Theorem (Indep vs Separation)

any 2 of the 3 definitions of fairness are mutually exclusive (except in degenerate cases)

① - ③ Statistical Parity & Equalized Odds (* For binary Y *)

are mutually exclusive unless $A \perp Y$ or $\hat{Y} \perp Y$

$$A \perp \hat{Y} \text{ and } A \perp \hat{Y} \mid Y \Rightarrow A \perp Y \text{ or } \hat{Y} \perp Y$$

$$\Pr[\hat{Y} = b] = \Pr[\hat{Y} = b \mid A = a] = \sum_{y} \Pr[\hat{Y} = b \mid A = a, Y = y] \Pr[Y = y \mid A = a]$$

$$= \sum_{y} \Pr[\hat{Y} = b \mid Y = y] \Pr[Y = y \mid A = a]$$

$$\Pr[\hat{Y} = b] = \sum_{y} \Pr[\hat{Y} = b \mid Y = y] \Pr[Y = y]$$

So

$$\sum_y \underbrace{\Pr[\hat{Y}=b \mid Y=y]}_{b_y} \underbrace{\Pr[Y=y]}_{p} = \sum_y \underbrace{\Pr[\hat{Y}=b \mid Y=y]}_{b_y} \underbrace{\Pr[Y=y \mid A=a]}_{p_a}$$

$$p b_0 + (1-p) b_1 = p_a b_0 + (1-p_a) b_1$$

$$p(b_0 - b_1) + b_1 = p_a (b_0 - b_1) + b_1$$

$$p(b_0 - b_1) = p_a (b_0 - b_1)$$

so either $\underbrace{b_0 = b_1}_{\hat{Y} \perp Y}$ or $\underbrace{p = p_a}_{Y \perp A}$

BACK TO COMPAS

Probublica says:

Blacks face higher false positive rates

so violates equalized odds

Northpointe's defense:

Scores satisfy predictive rate parity

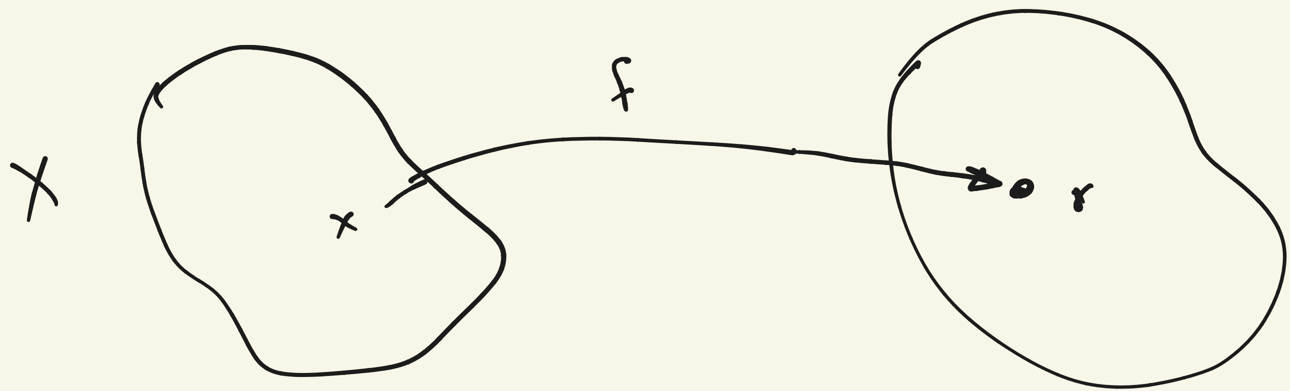
and we can't have both.

CALIBRATION ≈ PREDICTIVE RATE PARITY

Defn. A score R is calibrated if $\forall r$

$$\Pr[Y=1 \mid R=r] = r$$

* Note the natural score function $R(x) = \mathbb{E}[Y=1 \mid x]$ is calibrated



$R = f(X)$

CALIBRATION ≈ PREDICTIVE RATE PARITY

Defn. A score R is calibrated by group if $\forall r$

$$\Pr[Y=1 \mid R=r, A] = r$$

* Note the natural score function $R(x) = \mathbb{E}[Y=1 \mid x, A]$ is calibrated by group

CALIBRATION $\approx$ PREDICTIVE RATE PARITY

Defn. A score R is calibrated by group if $\forall r$

$$\Pr[Y=1 \mid R=r, A] = r$$

Lemma

1. R calibrated by group $\Longrightarrow$ R satisfies predictive rate parity (PRP)

2. R satisfies predictive rate parity $\Longrightarrow$ $\exists$ score function ℓ s.t. $\ell(R)$ satisfies calibration by group

① Calibrated by group $\Longrightarrow$ PRP $\qquad Y \perp R \mid A$

Assume $\forall r, a$

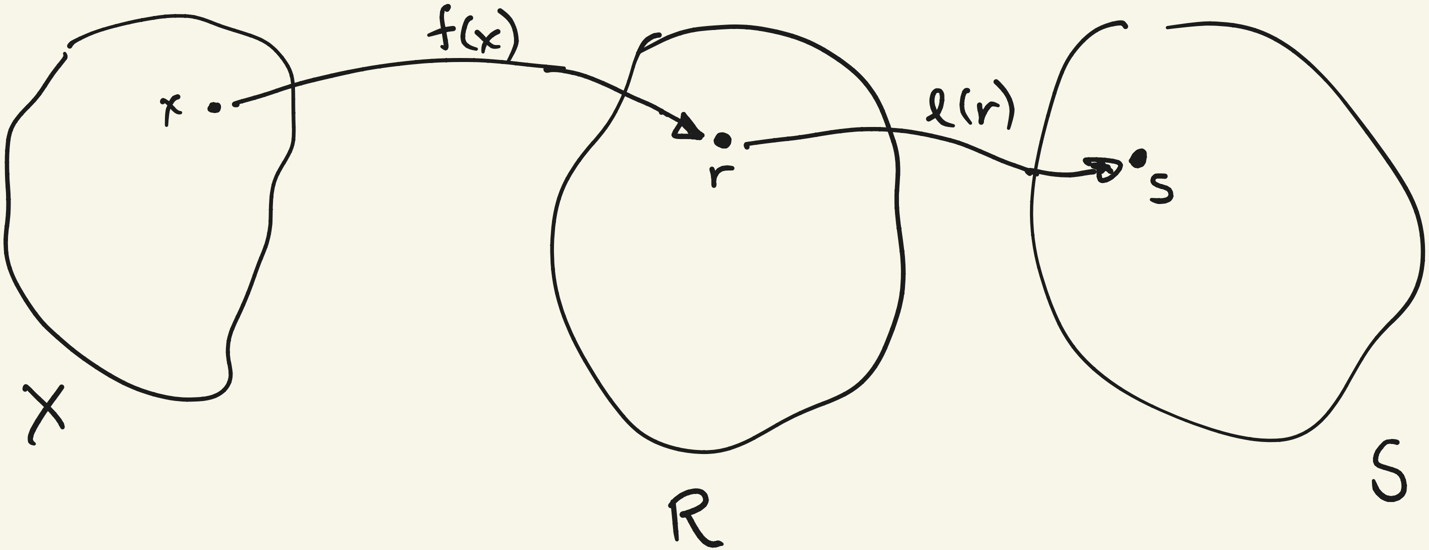
$$\Pr[Y=1 \mid R=r, A=a] = r$$

Then

$$\Pr[Y=1 \mid R, A] = \Pr[Y=1 \mid R] \quad \checkmark$$

(2) R satisfies PRP $\Rightarrow \exists\, \ell$ s.t. $\ell(R)$ satisfies calibration by group

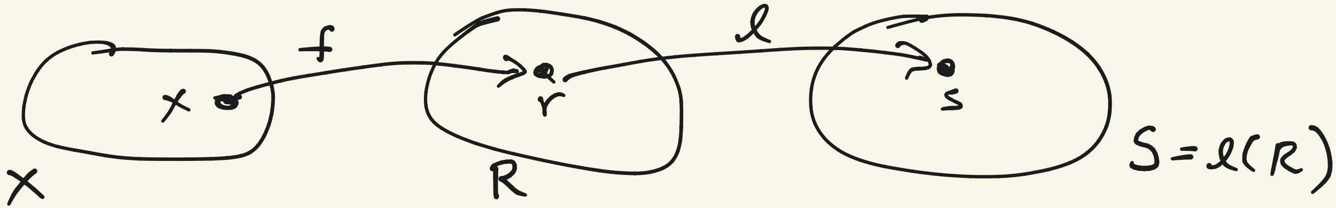
Define $\ell(r) = \Pr[Y=1 \mid R=r, A=b]$



Show $\Pr[Y=1 \mid S, A] = \Pr[Y=1 \mid S]$

(2) R satisfies PRP $\Rightarrow \exists \ell$ s.t. $\ell(R)$ satisfies calibration by group

Define $\ell(r) = \Pr[Y=1 \mid R=r, A=b]$



$$\Pr[Y=1 \mid S=s, A=a]$$
$$= \Pr[Y=1 \mid r \in \ell^{-1}(s), A=a]$$
$$= \Pr[Y=1 \mid r \in \ell^{-1}(s)]$$
$$= \Pr[Y=1 \mid S=s]$$

$\therefore \ell(R)$ is calibrated by group.

(Semi-) Intuitive Proof (calibration vs equalized odds)

N_a = # people in group $A=a$

N_a^+ = # people in group $A=a$, with $Y=1$

N_a^- = " " " " " " $Y=0$

R_a = total score (sum of scores) for people in group $A=a$

R_a^+ = " " " " " $A=a$, and $Y=1$

R_a^- = " " " " " $A=a$, and $Y=0$

$\overline{R}_a^+$ = avg score of people in $A=a$, $Y=1$ = R_a^+ / N_a^+

$\overline{R}_a^-$ = R_a^- / N_a^-

$$R_a = N_a^+ \overline{R}_a^+ + N_a^- \overline{R}_a^-$$

Calibration : $R_a = N_a^+$

all scores

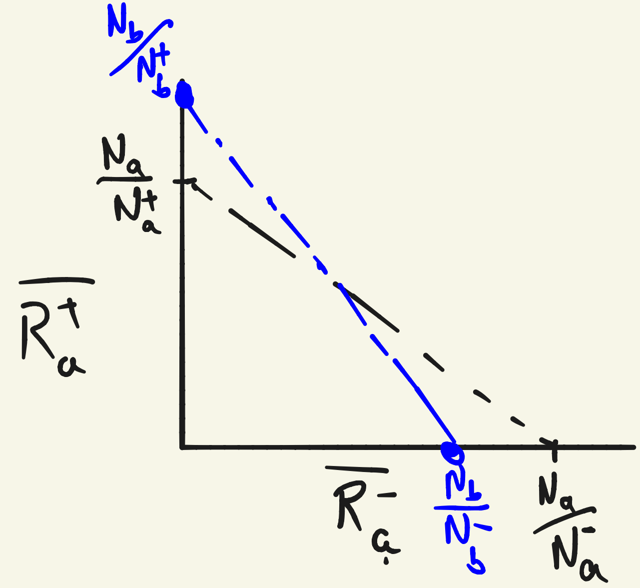
total score for bin r = (# people in bin r) r

$$= (N_{a,r})\, r = N^+_{a,r}$$

so

$$N_a^+ = \sum_r N^+_{a,r} = R_a$$

so $R_a = N_a^+ \underbrace{\overline{R_a^+}}_{y} + N_a^- \underbrace{\overline{R_a^-}}_{x} = N_a^+$



Today

1. Definitions of fairness
    - statistical parity
    - predictive rate parity
    - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness + Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Tradeoff between Fairness + Accuracy

<u>Example</u> Suppose $Y=1$ iff $A=1$

Then accuracy obviously at odds with fairness

"Inherent Tradeoffs in Learning Fair Representations"
[Zhao, Gordon]

- quantitative tradeoffs between statistical parity + accuracy via distance between distributions $\mathcal{D}_{A=0}(Y)$ and $\mathcal{D}_{A=1}(Y)$
- I.e., $\hat{Y}$ satisfies stat. parity,
  error $\geq d_{TV}(\mathcal{D}_0(Y), \mathcal{D}_1(Y))$

Today

1. Definitions of fairness
   - statistical parity
   - predictive rate parity
   - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness & Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Simplicity/Fairness Tradeoffs

(Kleinberg, Mullainathan, 2019)

Setup:

- Set of applicants, can accept an r fraction
- $x \in R^K$, $(K+1)^{st}$ dimension x_{K+1} is membership in A
- $s(x)$ gives a score to x
- s is simple if it doesnt depend on A
- top r percent (based on score) are admitted

Simplicity / Fairness Tradeoffs

Conditions:

① <u>DISADVANTAGE condition on s</u>

Let $\mu(x, A=b)$ be fraction of population with value (x, b)

For all x such that $s(x) > s(x')$

$$\frac{\mu(x, A=0)}{\mu(x, A=1)} > \frac{\mu(x', A=0)}{\mu(x', A=1)}$$

Simplicity/Fairness Tradeoffs

Conditions:

① DISADVANTAGE condition on s

Let $\mu(x, A=b)$ be fraction of population with value (x,b)

For all x such that $s(x) > s(x')$

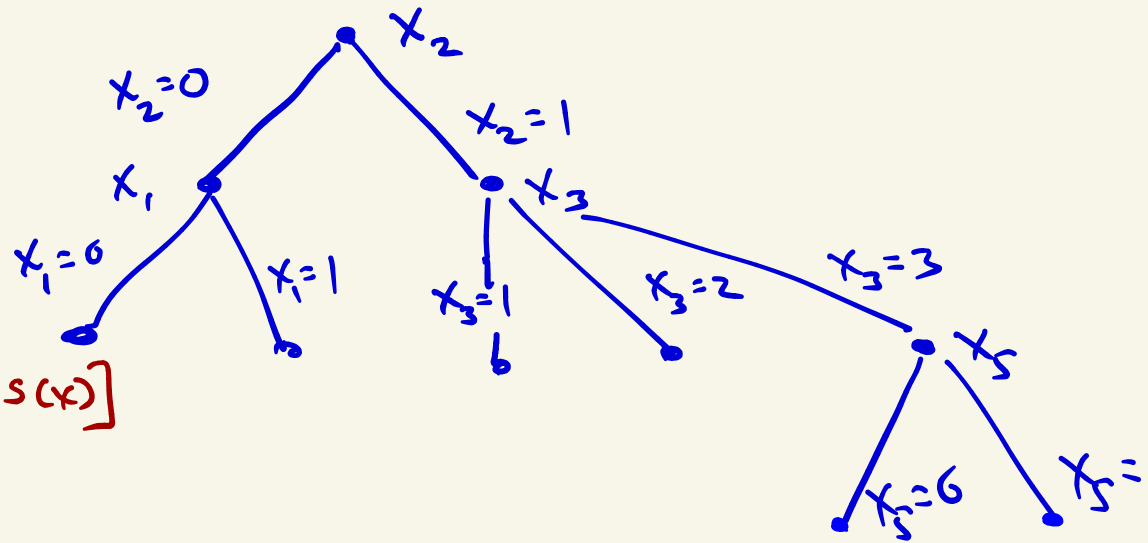
$$\frac{\mu(x, A=0)}{\mu(x, A=1)} > \frac{\mu(x', A=0)}{\mu(x', A=1)}$$

② GENERICITY condition on s for S, T subsets of applicants, $E[s(S)] \neq E[s(T)]$

ensures NO further simplification of score is possible

Simplicity/Fairness Tradeoffs

Simple s-approximators: Decision trees



$$\mathop{\mathbb{E}}_{\substack{x,\ x_2=0 \\ x_1=0}}[s(x)]$$

For a path p in decision tree with partial assignment ρ, label leaf of p with $\displaystyle\mathop{\mathbb{E}}_{x,\ x \text{ consistent with } \rho}[s(x)]$

Simplicity/Fairness Tradeoffs

Simple s-approximators: Decision trees

A decision tree f approximates s as follows:

order subcubes highest to lowest (by leaf value) and output individuals in this order until we reach rate r

Efficiency of f, $V_f(r)$: avg value of s for the admitted people

Equity of f, $W_f(r)$: fraction of admitted people who belong to $A=1$ (disadvantaged group)

Simplicity/Fairness Tradeoffs

Theorem 1 Let s satisfy disadvantage & genericity conditions. Then every simple s-approximator is strictly improvable:

For every nontrivial simple approximator g to s, there is a refinement h of g that is better:

$$\forall r \quad V_g(r) \leq V_h(r), \quad W_g(r) \leq W_h(r)$$

and

$$\exists r^* \text{ st } V_g(r^*) < V_h(r^*), \quad V_g(r^*) < V_h(r^*)$$

Simplicity/Fairness Tradeoffs

Theorem 2 Say that an s-approximator f (Not nec. simple) is "group-agnostic" if $f(x, A=0) = f(x, A=1)$ $\forall x$

Let f be a group agnostic approximator and let f' be the approximator to f obtained by splitting/refining each cell c_i of f according to group membership in A.

Then $V_{g'} > V_g$ but $W_{g'} < W_g$

i.e. if we try to approx s by a group agnostic g, this incentivizes a rule that depends on A where value improves at expense of equity

Simplicity/Fairness Tradeoffs

Simple s-approximators: Decision trees

A decision tree f approximates s as follows:
order subcubes highest to lowest (by leaf value)
and output individuals in this order
until we reach rate r

Efficiency of f, $V_f(r)$: avg value of s for the admitted people

Equity of f, $W_f(r)$: fraction of admitted people who belong to $A = 1$ (disadvantaged group)

Today

1. Definitions of fairness
   - statistical parity
   - predictive rate parity
   - equalized odds
2. Impossibility results: any 2 of the 3 fairness conditions cannot be achieved (except in degenerate situations)
3. Tradeoff between Fairness & Accuracy
4. Tradeoffs between simplicity and Fairness
5. Other

Fairness Under Composition (Dwork, Ilvento ITCS 2019)