Fairness

Christos Dimitrakakis

October 2, 2019

Fairness

What is it?

Fairness

What is it?

- Meritocracy.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.
- Non-discrimination.

Meritocracy

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with 4+ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?
- Admit randomly?

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

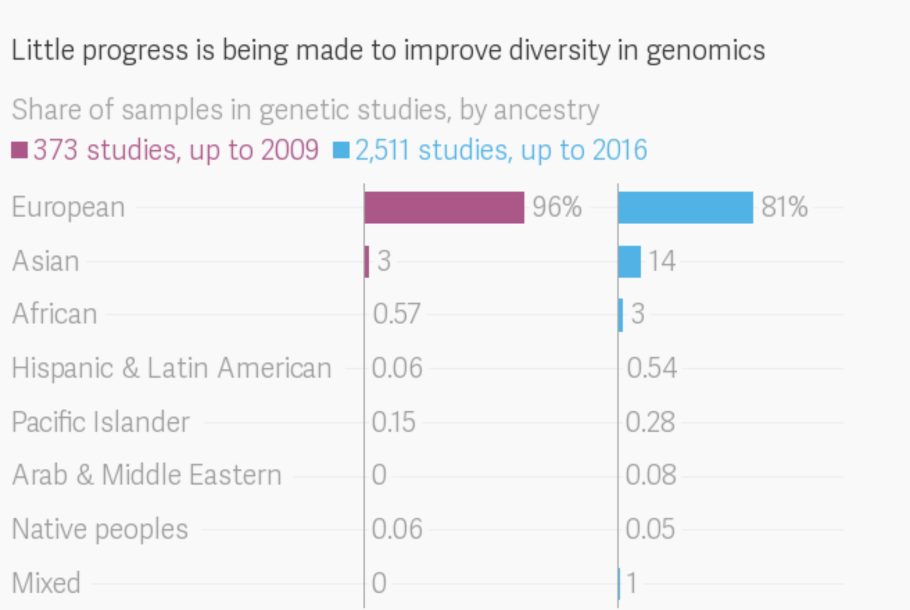
- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

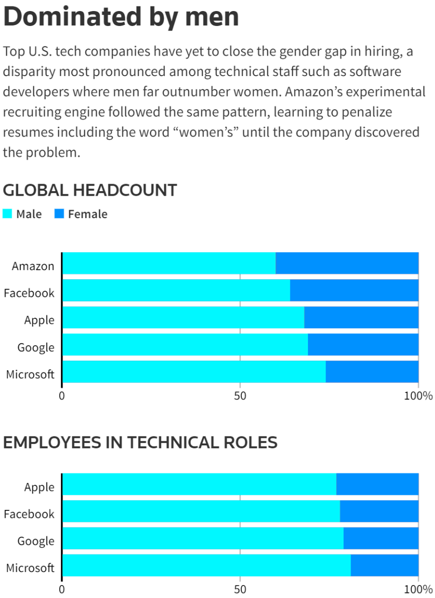
Solutions

- Admit everybody?
- Admit randomly?
- Use prediction of individual academic performance?

Proportional representation



Hiring decisions



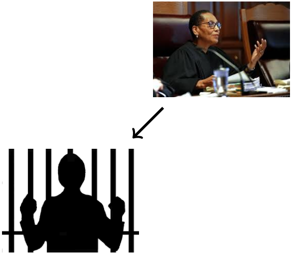
Fairness and information

Example 3 (College admissions data)

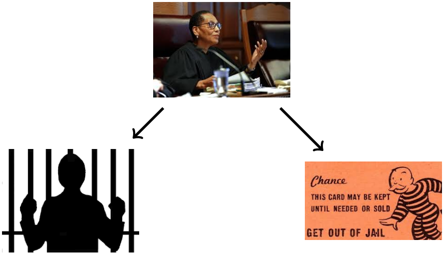
School	Male	Female
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%
Average	*45%*	*38%*

Bail decisions

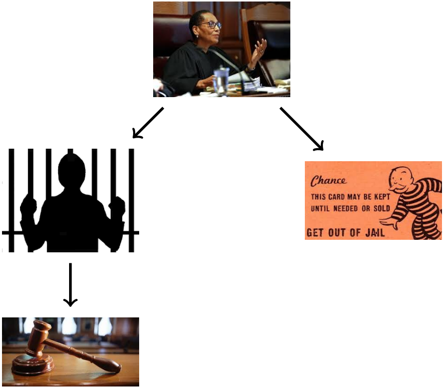
Bail decisions



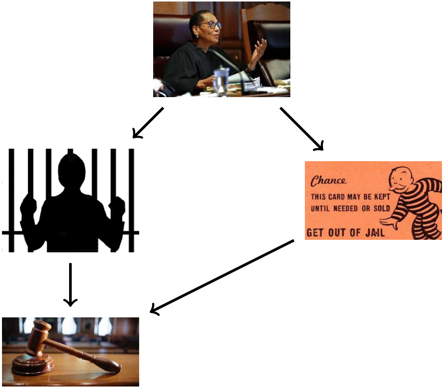
Bail decisions



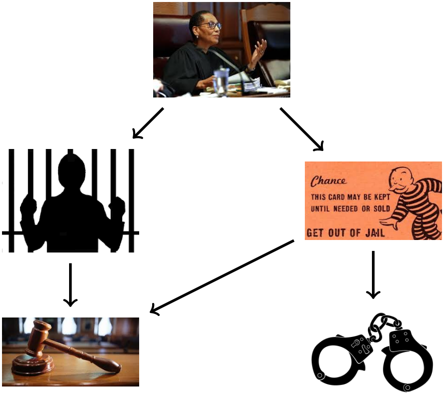
Bail decisions



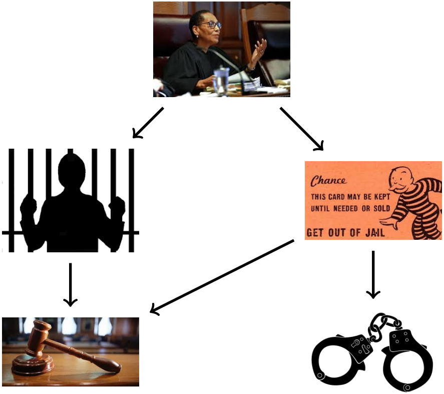
Bail decisions

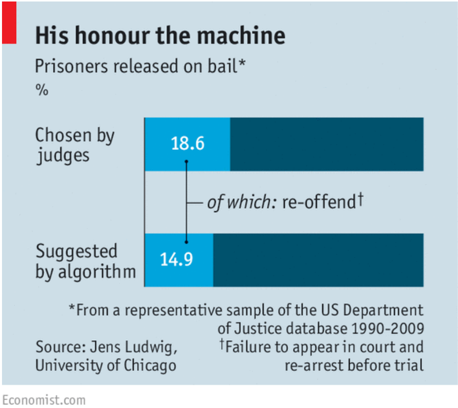


Bail decisions



Bail decisions





Whites get lower scores than blacks[1]

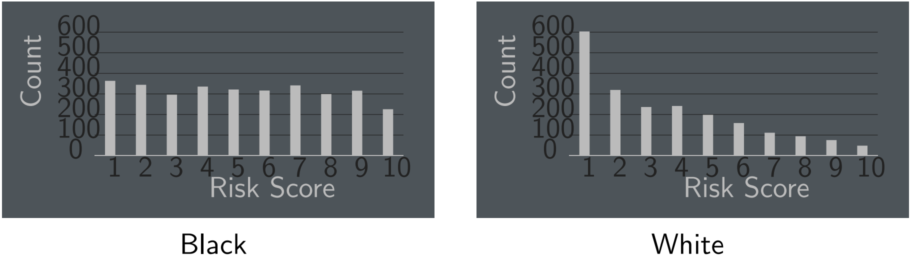


Figure: Apparent bias in risk scores towards black versus white defendants.

[1] Pro-publica, 2016

But scores equally accurately predict recidivsm[2]

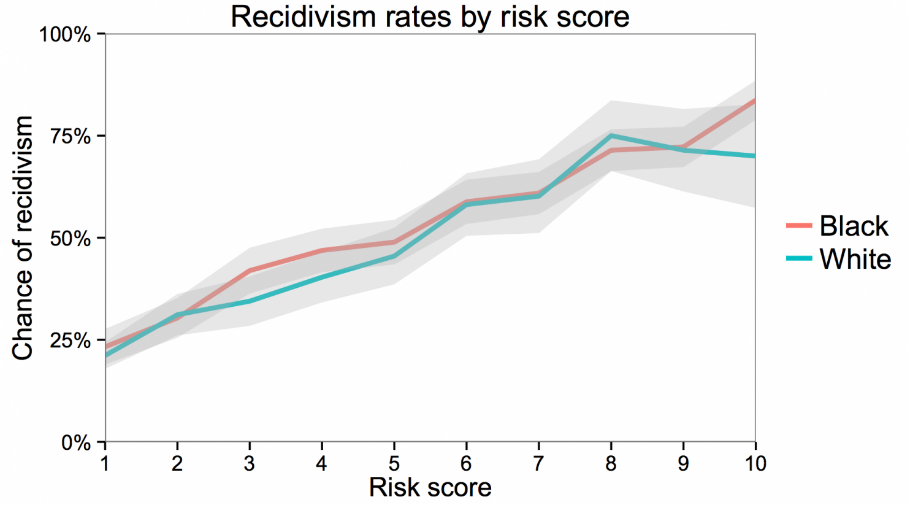


Figure: Recidivism rates by risk score.

[2] Washington Post, 2016

But non-offending blacks get higher scores

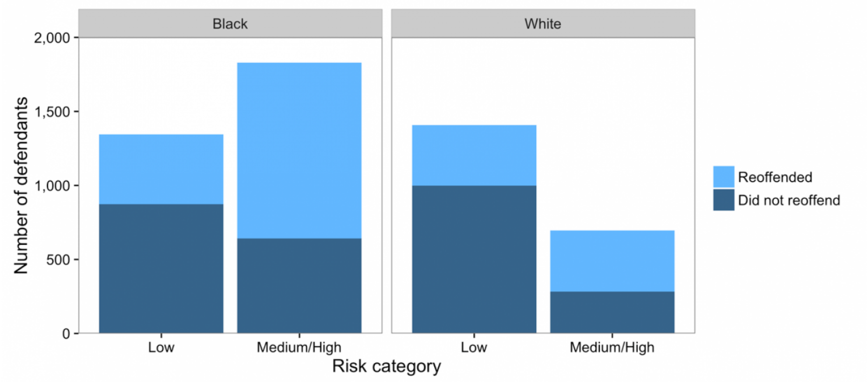


Figure: Score breakdown based on recidivism rates.

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

Graphical models

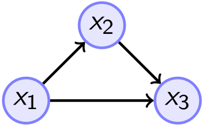


Figure: Graphical model (directed acyclic graph) for three variables.

Joint probability

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then $\boldsymbol{x} : \Omega \to X$, $X = \prod_i X_i$ and:

$$\mathbb{P}(\boldsymbol{x} \in A) = P(\{\omega \in \Omega \mid \boldsymbol{x}(\omega) \in A\}).$$

Factorisation

$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{x}_B \mid \boldsymbol{x}_C)\,\mathbb{P}(\boldsymbol{x}_C), \qquad B, C \subset [n]$$

Graphical models

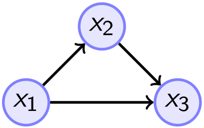


Figure: Graphical model (directed acyclic graph) for three variables.

Joint probability

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then $\boldsymbol{x} : \Omega \to X$, $X = \prod_i X_i$ and:

$$\mathbb{P}(\boldsymbol{x} \in A) = P(\{\omega \in \Omega \mid \boldsymbol{x}(\omega) \in A\}).$$

Factorisation

So we can write any joint distribution as

$$\mathbb{P}(x_1) \, \mathbb{P}(x_2 \mid x_1) \, \mathbb{P}(x_3 \mid x_1, x_2) \cdots \mathbb{P}(x_n \mid x_1, \ldots, x_{n-1}).$$

Directed graphical models

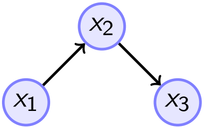


Figure: Graphical model for the factorisation $\mathbb{P}(x_3 \mid x_2)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_1)$.

Conditional independence

We say x_i is conditionally independent of $\boldsymbol{x}_B$ given $\boldsymbol{x}_D$ and write $x_i \mid \boldsymbol{x}_D \perp\!\!\!\perp \boldsymbol{x}_B$ iff

$$\mathbb{P}(x_i, \boldsymbol{x}_B \mid \boldsymbol{x}_D) = \mathbb{P}(x_i \mid \boldsymbol{x}_D)\,\mathbb{P}(\boldsymbol{x}_B \mid \boldsymbol{x}_D).$$

Example 4 (Smoking and lung cancer)

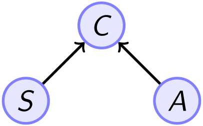


Figure: Smoking and lung cancer graphical model, where S: Smoking, C: cancer, A: asbestos exposure.

Explaining away

Even though S, A are independent, they become dependent once you know C.

Example 5 (Time of arrival at work)

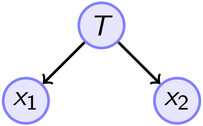


Figure: Time of arrival at work graphical model where T is a traffic jam and x_1 is the time John arrives at the office and x_2 is the time Jane arrives at the office.

Conditional independence

Even though x_1, x_2 are correlated, they become independent once you know T.

Example 6 (Treatment effects)

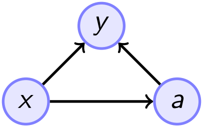


Figure: Kidney treatment model, where x: severity, y: result, a: treatment applied

	Treatment A	Treatment B
Small stones	87	270
Large stones	263	80

Severity	Treatment A	Treatment B
Small stones)	93%	87%
Large stones	73%	69%
Average	78%	83%

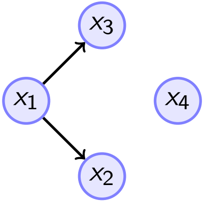
Example 7 (School admission)

Figure: School admission graphical model, where z: gender, s: school applied to, a: whether you were admitted.

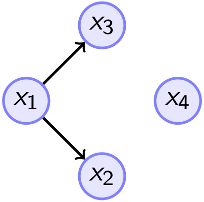
School	Male	Female
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%
Average	*45%*	*38%*

Exercise 1

Factorise the following graphical model.



Exercise 1

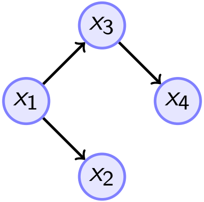


Factorise the following graphical model.

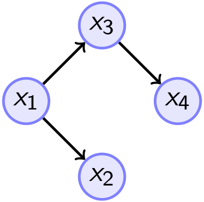
$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(x_1)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_3 \mid x_1)\,\mathbb{P}(x_4)$$

Exercise 2

Factorise the following graphical model.



Exercise 2



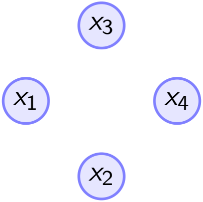
Factorise the following graphical model.

$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(x_1)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_3 \mid x_1)\,\mathbb{P}(x_4 \mid x_3)$$

Exercise 3

What dependencies does the following factorisation imply?

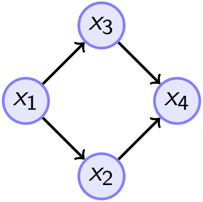
$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(x_1)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_3 \mid x_1)\,\mathbb{P}(x_4 \mid x_2, x_3)$$



Exercise 3

What dependencies does the following factorisation imply?

$$\mathbb{P}(\boldsymbol{x}) = \mathbb{P}(x_1)\,\mathbb{P}(x_2 \mid x_1)\,\mathbb{P}(x_3 \mid x_1)\,\mathbb{P}(x_4 \mid x_2, x_3)$$



Deciding conditional independence

There is an algorithm for deciding conditional independence of any two variables in a graphical model.

Inference and prediction in graphical models

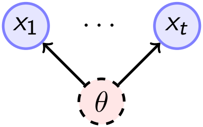


Figure: Inference and prediction in a graphical model.

Inference of latent variables

$$\mathbb{P}(\theta \mid x_1, \ldots, x_t)$$

- Model parameters.
- System states.

Inference and prediction in graphical models

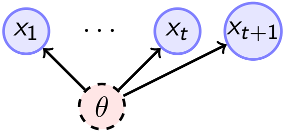


Figure: Inference and prediction in a graphical model.

Prediction

$$\mathbb{P}(x_{t+1} \mid x_1, \ldots, x_t) = \int_{\Theta} P(x_{t+1} \mid x_1, \ldots, x_t) \, \mathrm{d}P(\theta \mid x_1, \ldots, x_t)$$

Predictions are testable.

Coin tossing, revisited

Example 8

The Beta-Bernoulli prior

Figure: Graphical model for a Beta-Bernoulli prior

$$\theta \sim \mathcal{B}eta(\xi_1, \xi_2), \quad \text{i.e. } \xi \text{ are Beta distribution parameters} \tag{3.1}$$

$$x \mid \theta \sim \mathcal{B}ernoulli(\theta), \quad \text{i.e. } P_\theta(x) \text{ is a Bernoulli} \tag{3.2}$$

Example 9

The n-meteorologists problem (again)

- Meteorological models $\mathcal{M} = \{\mu_1, \ldots, \mu_n\}$
- Rain predictions at time t: $p_{t,\mu} \triangleq P_\mu(x_t = \mathrm{rain})$.
- Prior probability $\xi(\mu) = 1/n$ for each model.
- Decision a, resulting in utility $U(a, x_{t+1})$

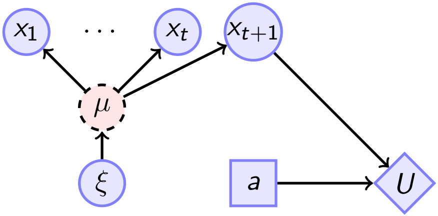


Figure: Inference, prediction and decisions in a graphical model.

Measuring independence

Theorem 10

If $x_i \mid \boldsymbol{x}_D \perp\!\!\!\perp \boldsymbol{x}_B$ *then*

$$\mathbb{P}(x_i \mid \boldsymbol{x}_B, \boldsymbol{x}_D) = \mathbb{P}(x_i \mid \boldsymbol{x}_D)$$

Example 11

$$\| \mathbb{P}(a \mid y, z) - \mathbb{P}(a \mid y) \|_1$$

which for discrete a, y, z is:

$$\max_{i,j} \| \mathbb{P}(a \mid y = i, z = j) - \mathbb{P}(a \mid y = i) \|_1 = \max_{i,j} \| \sum_k \mathbb{P}(a = k \mid y = i, z = j) -$$

Measuring independence

Theorem 10

If $x_i \mid \boldsymbol{x}_D \perp\!\!\!\perp \boldsymbol{x}_B$ *then*

$$\mathbb{P}(x_i \mid \boldsymbol{x}_B, \boldsymbol{x}_D) = \mathbb{P}(x_i \mid \boldsymbol{x}_D)$$

This implies

$$\mathbb{P}(x_i \mid \boldsymbol{x}_B = b, \boldsymbol{x}_D) = \mathbb{P}(x_i \mid \boldsymbol{x}_B = b', \boldsymbol{x}_D)$$

so we can measure independence by seeing how the distribution of x_i changes when we vary $\boldsymbol{x}_B$, keeping $\boldsymbol{x}_D$ fixed.

Example 11

$$\| \mathbb{P}(a \mid y, z) - \mathbb{P}(a \mid y) \|_1$$

which for discrete a, y, z is:

$$\max_{i,j} \| \mathbb{P}(a \mid y = i, z = j) - \mathbb{P}(a \mid y = i) \|_1 = \max_{i,j} \| \sum_k \mathbb{P}(a = k \mid y = i, z = j) -$$

Example 12

An alternative model for coin-tossing This is an elaboration of Example **??** for hypothesis testing.

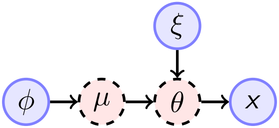


Figure: Graphical model for a hierarchical prior

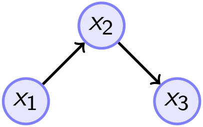
- μ_1: A Beta-Bernoulli model with $\mathcal{B}eta(\xi_1, \xi_2)$
- μ_0: The coin is fair.

$$\theta \mid \mu = \mu_0 \sim \mathcal{D}(0.5), \qquad \text{i.e. } \theta \text{ is always } 0.5 \tag{3.3}$$

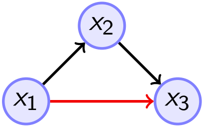
$$\theta \mid \mu = \mu_1 \sim \mathcal{B}eta(\xi_1, \xi_2), \qquad \text{i.e. } \theta \text{ has a Beta distribution} \tag{3.4}$$

$$x \mid \theta \sim \mathcal{B}ernoulli(\theta), \qquad \text{i.e. } P_\theta(x) \text{ is Bernoulli} \tag{3.5}$$

Bayesian testing of independence



(a) Θ_0 assumes independence



(b) Θ_1 does not assume independence

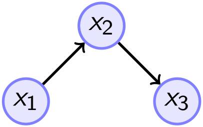
Example 13

Assume data $D = \{x_1^t, x_2^t, x_3^t \mid t = 1, \ldots, T\}$ with $x_i^t \in \{0, 1\}$.

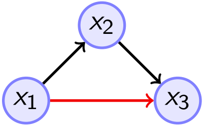
$$P_\theta(D) = \prod_t P_\theta(x_3^t \mid x_2^t) P_\theta(x_2^t \mid x_1^t) P_\theta(x_1^t), \qquad \theta \in \Theta_0 \tag{3.6}$$

$$P_\theta(D) = \prod_t P_\theta(x_3^t \mid x_2^t, x_1^t) P_\theta(x_2^t \mid x_1^t) P_\theta(x_1^t), \qquad \theta \in \Theta_1 \tag{3.7}$$

Bayesian testing of independence



(a) Θ_0 assumes independence



(b) Θ_1 does not assume independence

Example 13

$$\theta_1 \triangleq P_\theta(x_1^t = 1) \qquad (\mu_0, \mu_1)$$

$$\theta_{2|1}^i \triangleq P_\theta(x_2^t = 1 \mid x_1^t = i) \qquad (\mu_0, \mu_1)$$

$$\theta_{3|2}^j \triangleq P_\theta(x_3^t = 1 \mid x_2^t = j) \qquad (\mu_0)$$

$$\theta_{3|2,1}^{i,j} \triangleq P_\theta(x_3^t = 1 \mid x_2^t = j, x_1^t = i) \qquad (\mu_1)$$

Figure: Hierarchical model.

$$\mu_i \sim \phi \tag{3.6}$$

$$\theta \mid \mu = \mu_i \sim \xi_i \tag{3.7}$$

Marginal likelihood

$$\mathbb{P}_\phi(D) = \phi(\mu_0)\,\mathbb{P}_{\mu_0}(D) + \phi(\mu_1)\,\mathbb{P}_{\mu_1}(D) \tag{3.8}$$

$$\mathbb{P}_{\mu_i}(D) = \int_{\Theta_i} P_\theta(D)\,\mathrm{d}\xi_i(\theta). \tag{3.9}$$

$$\mu \longrightarrow \theta \longrightarrow D$$

Figure: Hierarchical model.

Marginal likelihood

$$\mathbb{P}_\phi(D) = \phi(\mu_0)\,\mathbb{P}_{\mu_0}(D) + \phi(\mu_1)\,\mathbb{P}_{\mu_1}(D) \tag{3.6}$$

$$\mathbb{P}_{\mu_i}(D) = \int_{\Theta_i} P_\theta(D)\,\mathrm{d}\xi_i(\theta). \tag{3.7}$$

Model posterior

$$\phi(\mu \mid D) = \frac{\mathbb{P}_\mu(D)\phi(\mu)}{\sum_i \mathbb{P}_{\mu_i}(D)\phi(\mu_i)} \tag{3.8}$$

Calculating the marginal likelihood

Monte-Carlo approximation

$$\int_{\Theta} P_\theta(D) \, \mathrm{d}\xi(\theta) \approx \sum_{n=1}^{N} P_{\theta_n}(D) + O(1/\sqrt{N}), \qquad \theta_n \sim \xi \tag{3.9}$$

Importance sampling

$$\int_{\Theta} P_\theta(D) \, \mathrm{d}\xi(\theta) \tag{3.10}$$

Calculating the marginal likelihood

Monte-Carlo approximation

$$\int_\Theta P_\theta(D)\,\mathrm{d}\xi(\theta) \approx \sum_{n=1}^N P_{\theta_n}(D) + O(1/\sqrt{N}), \qquad \theta_n \sim \xi \tag{3.9}$$

Importance sampling

$$\int_\Theta P_\theta(D)\,\mathrm{d}\xi(\theta) = \int_\Theta P_\theta(D)\frac{\mathrm{d}\psi(\theta)}{\mathrm{d}\psi(\theta)}\,\mathrm{d}\xi(\theta) \tag{3.10}$$

Calculating the marginal likelihood

Monte-Carlo approximation

$$\int_\Theta P_\theta(D)\,\mathrm{d}\xi(\theta) \approx \sum_{n=1}^N P_{\theta_n}(D) + O(1/\sqrt{N}), \qquad \theta_n \sim \xi \tag{3.9}$$

Importance sampling

$$\int_\Theta P_\theta(D)\,\mathrm{d}\xi(\theta) = \int_\Theta P_\theta(D) \frac{\mathrm{d}\xi(\theta)}{\mathrm{d}\psi(\theta)}\,\mathrm{d}\psi(\theta) \tag{3.10}$$

Calculating the marginal likelihood

Monte-Carlo approximation

$$\int_{\Theta} P_\theta(D) \, \mathrm{d}\xi(\theta) \approx \sum_{n=1}^{N} P_{\theta_n}(D) + O(1/\sqrt{N}), \qquad \theta_n \sim \xi \tag{3.9}$$

Importance sampling

$$\int_{\Theta} P_\theta(D) \, \mathrm{d}\xi(\theta) \approx \sum_{n=1}^{N} P_\theta(D) \frac{\mathrm{d}\xi(\theta_n)}{\mathrm{d}\psi(\theta_n)}, \qquad \theta_n \sim \psi \tag{3.10}$$

Sequential updating of the marginal likelihood

$$\mathbb{P}_\xi(D) \tag{3.14}$$

Example 14 (Beta-Bernoulli)

$$\mathbb{P}_\xi(x_t = 1 \mid x_1, \ldots, x_{t-1}) = \frac{\alpha_t}{\alpha_t + \beta_t},$$

with $\alpha_t = \alpha_0 + \sum_{n=1}^{t-1} x_n, \quad \beta_t = \beta_0 + \sum_{n=1}^{t-1}(1 - x_n)$

Sequential updating of the marginal likelihood

$$\mathbb{P}_\xi(D) = \mathbb{P}_\xi(x_1, \ldots, x_T) \tag{3.14}$$

Example 14 (Beta-Bernoulli)

$$\mathbb{P}_\xi(x_t = 1 \mid x_1, \ldots, x_{t-1}) = \frac{\alpha_t}{\alpha_t + \beta_t},$$

with $\alpha_t = \alpha_0 + \sum_{n=1}^{t-1} x_n, \quad \beta_t = \beta_0 + \sum_{n=1}^{t-1} (1 - x_n)$

Sequential updating of the marginal likelihood

$$\mathbb{P}_\xi(D) = \mathbb{P}_\xi(x_1, \ldots, x_T) \tag{3.11}$$
$$= \mathbb{P}_\xi(x_2, \ldots, x_T \mid x_1)\, \mathbb{P}_\xi(x_1)$$

$$\tag{3.14}$$

Example 14 (Beta-Bernoulli)

$$\mathbb{P}_\xi(x_t = 1 \mid x_1, \ldots, x_{t-1}) = \frac{\alpha_t}{\alpha_t + \beta_t},$$

with $\alpha_t = \alpha_0 + \sum_{n=1}^{t-1} x_n, \quad \beta_t = \beta_0 + \sum_{n=1}^{t-1}(1 - x_n)$

Sequential updating of the marginal likelihood

$$\mathbb{P}_\xi(D) = \mathbb{P}_\xi(x_1, \ldots, x_T) \tag{3.11}$$

$$= \mathbb{P}_\xi(x_2, \ldots, x_T \mid x_1)\, \mathbb{P}_\xi(x_1) \tag{3.12}$$

$$= \prod_{t=1}^{T} \mathbb{P}_\xi(x_t \mid x_1, \ldots, x_{t-1})$$

$$\tag{3.14}$$

Example 14 (Beta-Bernoulli)

$$\mathbb{P}_\xi(x_t = 1 \mid x_1, \ldots, x_{t-1}) = \frac{\alpha_t}{\alpha_t + \beta_t},$$

with $\alpha_t = \alpha_0 + \sum_{n=1}^{t-1} x_n, \quad \beta_t = \beta_0 + \sum_{n=1}^{t-1} (1 - x_n)$

Sequential updating of the marginal likelihood

$$
\begin{aligned}
\mathbb{P}_\xi(D) &= \mathbb{P}_\xi(x_1, \ldots, x_T) && (3.11)\\
&= \mathbb{P}_\xi(x_2, \ldots, x_T \mid x_1)\, \mathbb{P}_\xi(x_1) && (3.12)\\
&= \prod_{t=1}^T \mathbb{P}_\xi(x_t \mid x_1, \ldots, x_{t-1}) && (3.13)\\
&= \prod_{t=1}^T \int_\Theta P_{\theta_n}(x_t)\, \mathrm{d}\underbrace{\xi(\theta \mid x_1, \ldots, x_{t-1})}_{\text{posterior at time } t} && (3.14)
\end{aligned}
$$

Example 14 (Beta-Bernoulli)

$$
\mathbb{P}_\xi(x_t = 1 \mid x_1, \ldots, x_{t-1}) = \frac{\alpha_t}{\alpha_t + \beta_t},
$$

with $\alpha_t = \alpha_0 + \sum_{n=1}^{t-1} x_n, \quad \beta_t = \beta_0 + \sum_{n=1}^{t-1}(1 - x_n)$

Further reading

Python sources

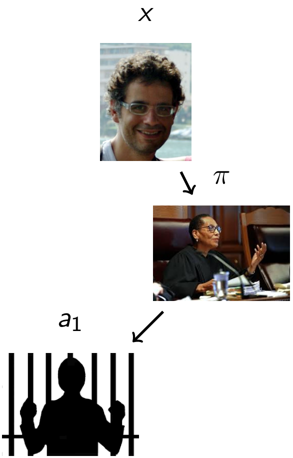
- A simple python measure of conditional independence
  `src/fairness/ci_test.py`
- A simple test for discrete Bayesian network
  `src/fairness/DirichletTest.py`
- Using the PyMC package
  `https://docs.pymc.io/notebooks/Bayes_factor.html`

Bail decisions, revisited

x

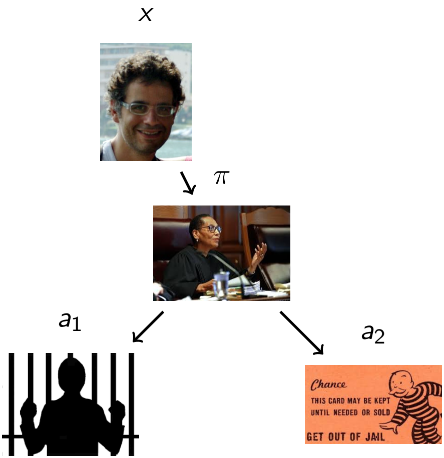
$\searrow \pi$

Bail decisions, revisited



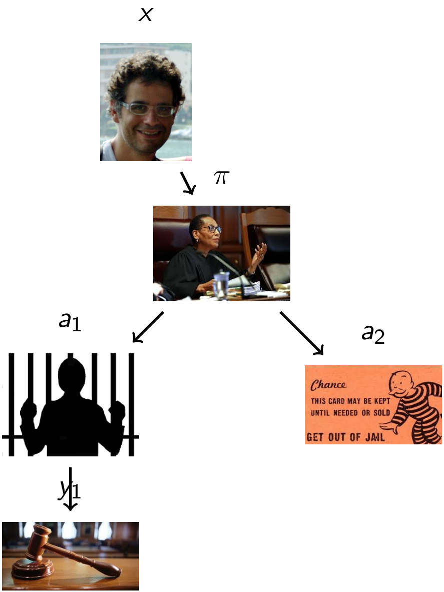
$$\pi(a \mid x) \qquad \text{(policy)}$$

Bail decisions, revisited



$\pi(a \mid x)$ (policy)

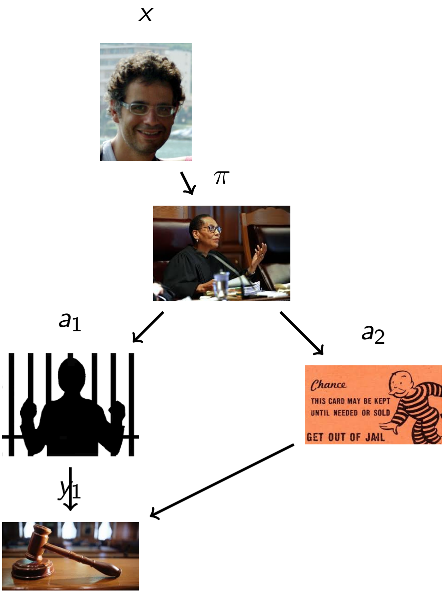
Bail decisions, revisited



$$\pi(a \mid x) \qquad \text{(policy)}$$

$$\mathbb{P}(y \mid a, x) \qquad \text{(outcome)}$$

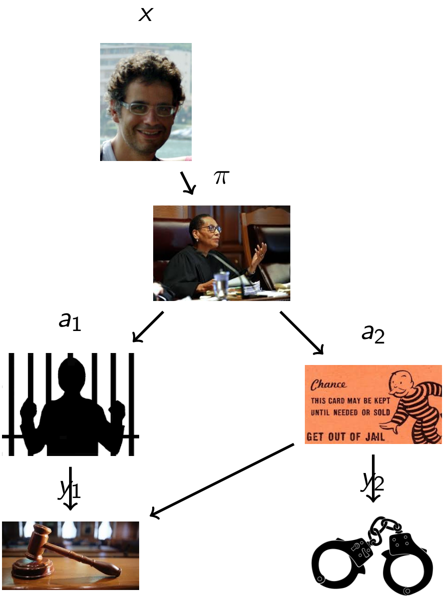
Bail decisions, revisited



$$\pi(a \mid x) \qquad \text{(policy)}$$

$$\mathbb{P}(y \mid a, x) \qquad \text{(outcome)}$$

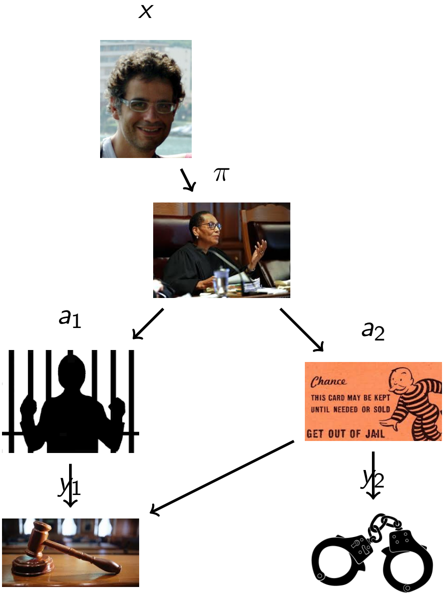
Bail decisions, revisited



$$\pi(a \mid x) \qquad \text{(policy)}$$

$$\mathbb{P}(y \mid a, x) \qquad \text{(outcome)}$$

Bail decisions, revisited

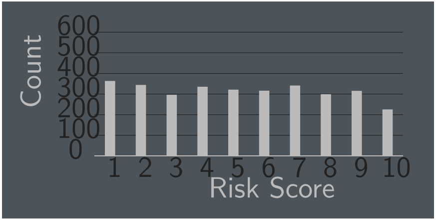


$\pi(a \mid x)$ (policy)

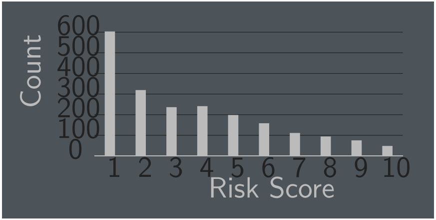
$\mathbb{P}(y \mid a, x)$ (outcome)

$U(a, y)$ (utility)

Independence



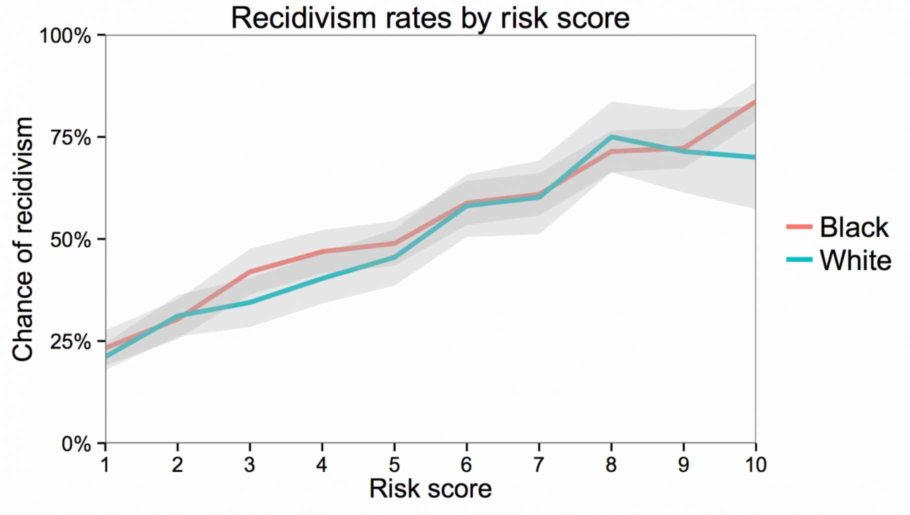
Black



White

Figure: Apparent bias in risk scores towards black versus white defendants.

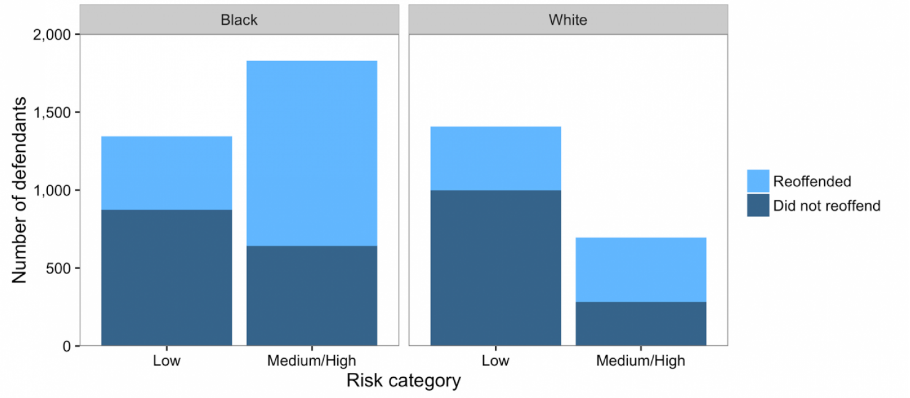
$$\mathbb{P}_\theta^\pi(a \mid z) = \mathbb{P}_\theta^\pi(a) \qquad \text{(non-discrimination)}$$



- y Result.
- a Assigned score.
- z Race.

$$\mathbb{P}^\pi(y \mid a, z) = \mathbb{P}^\pi(y \mid a) \qquad \text{(calibration)}$$

$$\mathbb{P}^\pi(a \mid y, z) = \mathbb{P}^\pi(a \mid y) \qquad \text{(balance)}$$



- y Result.
- a Assigned score.
- z Race.

$$\mathbb{P}^\pi(y \mid a, z) = \mathbb{P}^\pi(y \mid a) \qquad \text{(calibration)}$$

$$\mathbb{P}^\pi(a \mid y, z) = \mathbb{P}^\pi(a \mid y) \qquad \text{(balance)}$$

Meritocratic decision

$$a_t(\theta, x_t) \in \arg\max_a \mathbb{E}_\theta(U \mid a, x_t) = \int_{\mathcal{Y}} U(a_t, y) \, \mathbb{E}_\theta(U \mid a_t, x_t) \tag{4.1}$$

Smooth fairness

$$D[\pi(a \mid x), \pi(a \mid x')] \leq \rho(x, x'). \tag{4.2}$$

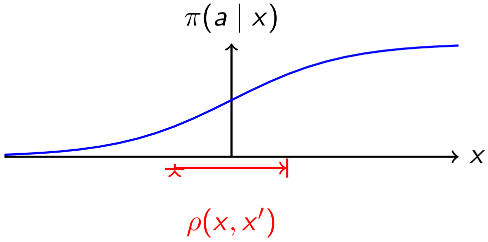


Figure: A Lipschitz function

The constrained maximisation problem

$$\max_{\pi} \left\{ U(\pi) \mid \rho(x, x') \leq \epsilon \right\} \tag{4.3}$$

The value of a policy

Fairness metrics: balance

$$F_{\text{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} |\, \mathbb{P}^{\pi}_{\theta}(a \mid y, z) - \mathbb{P}^{\pi}_{\theta}(a \mid y)|^2 \tag{4.4}$$

The value of a policy

Fairness metrics: balance

$$F_{\mathrm{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} |\, \mathbb{P}^{\pi}_{\theta}(a \mid y, z) - \mathbb{P}^{\pi}_{\theta}(a \mid y)|^2 \tag{4.4}$$

Utility: Classification accuracy

$$U(\theta, \pi) = \mathbb{P}^{\pi}_{\theta}(y_t = a_t)$$

The value of a policy

Fairness metrics: balance

$$F_{\mathrm{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} \left| \mathbb{P}^{\pi}_{\theta}(a \mid y, z) - \mathbb{P}^{\pi}_{\theta}(a \mid y) \right|^2 \tag{4.4}$$

Utility: Classification accuracy

$$U(\theta, \pi) = \mathbb{P}^{\pi}_{\theta}(y_t = a_t)$$

Use λ to trade-off utility and fairness

$$V(\lambda, \theta, \pi) = (1 - \lambda) \overbrace{U(\theta, \pi)}^{\text{utility}} - \lambda \underbrace{F(\theta, \pi)}_{\text{unfairness}} \tag{4.5}$$

Model uncertainty

θ is unknown

Theorem 15

A decision rule in the form of a lottery, i.e.

$$\pi(a \mid x) = p_a$$

can be the only way to satisfy balance for all possible θ.

Possible solutions

- Marginalize over θ ("expected" model)
- Use Bayesian reasoning

The value of a policy

Let λ represent the trade-off between utility and fairness.

$$V(\lambda, \theta, \pi) = \lambda \overbrace{U(\theta, \pi)}^{\text{utility}} - \underbrace{(1 - \lambda) F(\theta, \pi)}_{\text{fairness violation}} \tag{4.6}$$

The Bayesian decision problem

The Bayesian value of a policy

$$V(\lambda, \xi, \pi) = \int_{\Theta} V(\lambda, \theta, \pi) \, \mathrm{d}\xi(\theta). \tag{4.7}$$

Online resources

- COMPAS analysis by propublica
  `https://github.com/propublica/compas-analysis`
- Open policing database `https://openpolicing.stanford.edu/`

Learning outcomes

Understanding

- Graphical models.
- Conditional independence.
- Fairness as independence.
- Fairness as meritocracy.

Skills

- Be able to specify a graphical model capturing dependencies between variables.
- Be able to verify if a policy satisfies a fairness condition.

Reflection

- When looking at sensitive attributes, how easy is it to determine fairness?
- How should we balance the needs of individuals, the decision maker