Fairness and discrimination in actuarial pricing

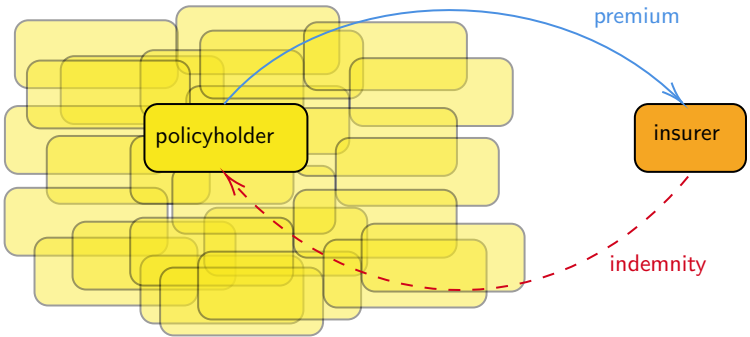
Arthur Charpentier[1]

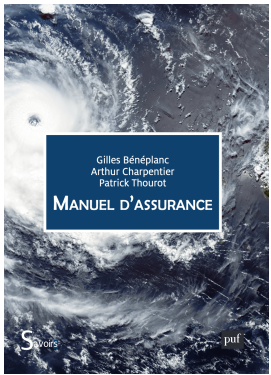
[1] Université du Québec à Montréal

Data Science for Actuaries – Institut des Actuaires

Preliminaries and Motivation

- Insurance is the contribution of the many to the misfortune of the few





There is no "*unique price*" for an individual... actuaries will use $\mathbb{E}_{\widehat{\mathbb{P}}_n}[Y|\boldsymbol{X}]$ as a proxy for $\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{\Omega}]$ to compute a "*fair actuarial premium*".

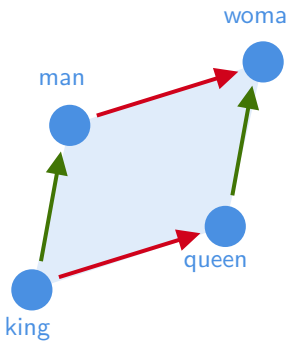
Preliminaries and Motivation

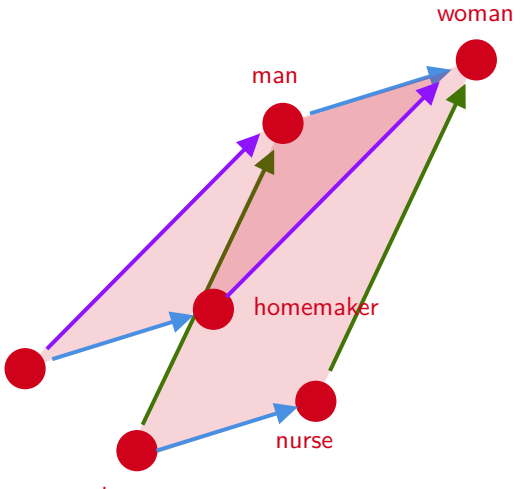
- "*Technology is neither good nor bad; nor is it neutral* " , Kranzberg (1986)
- "*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for* ", Kearns and Roth (2019)
- "*at the core of insurance business lies discrimination between risky and non-risky insureds*", Avraham (2017)
- Accuracy : $\pi(\boldsymbol{x}) = \mathbb{E}_{\widehat{\mathbb{P}}_n}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\widehat{\mathbb{P}}_n$ historical probability) (is)
- Fairness : $\pi^\star(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\mathbb{P}^\star$ targeted probability) (ought, Hume (1739))

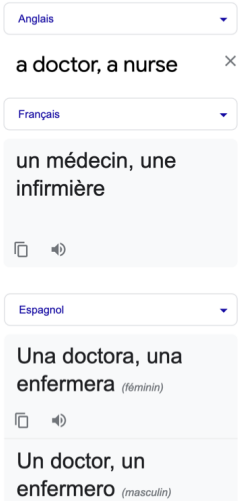
- Charpentier (2022a) Assurance: biais, discrimination et équité
- Charpentier (2022b) Insurance: biases, discrimination and fairness

Motivation

- Accuracy : $\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\mathbb{P}$ historical probability) (is)
- Fairness : $\pi^\star(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X} = \boldsymbol{x}]$ ($\mathbb{P}^\star$ targeted probability) (ought, Hume (1739))







Discrimination and Protected Attributes

California

Allowed (with applicable limitations): driving experience, marital status, address/zip code

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, residential status, insurance history

Notes & Clarifications: California's insurance commissioner banned gender as of January 2019. Occupation and education are permitted for use in group plans (i.e. for alumni associations and other membership programs).

Georgia

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, and employment status

Notes & Clarifications: none

Hawaii

Allowed (with applicable limitations): address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Illinois

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): none

Notes & Clarifications: none

Massachusetts

Allowed (with applicable limitations): years of driving experience, address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Michigan

Allowed (with applicable limitations): gender (group-rated policies), age, years of driving experience, credit history, education, occupation, employment status, marital status (group-rated policies), residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): gender (non-group policies), marital status (non-group policies)

Notes & Clarifications: Gender and marital status are permitted only in rate-making for group plans (i.e. for alumni associations and other membership programs). **UPDATE:** Michigan lawmakers approved a major insurance reform bill in May 2019 that will ban insurers in the state from using gender, marital status, address/zipcode, residential status, education and occupation in rate setting. The ban will be enforced starting in July 2020. Insurers will be permitted to use "territory" as approved by the state regulators instead of zip code.

New York

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, employment status

Notes & Clarifications: none

via The Zebra (2022)

Discrimination and Protected Attributes

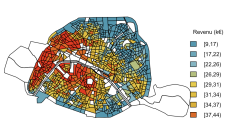
	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	X	X	●	X	●	X	X	●	●	●	●	X	X	●
Age	X	X	●	X*	●	X	●	●	●	●*	●	X	X	●
Driving experience	●	X	●	●	●	●	●	●	●	●	●	●	●	●
Credit history	X	X	●	●	●	X	●*	●	●	X*	X	●*	X	●
Education	X	X	X	X	X	X	●	●	●	●	●	●	●	●
Occupation	X	X	X	●	X	X	●	●	●	●	●	●	●	●
Employment status	X	X	X	●	X	X	●	●	●	●	●	●	●	●
Marital status	●	X	●	●	●	X	●	●	●	●	●	●	●	●
Housing situation	X	X	●	●	●	X	●	●	●	X	X	●	●	●
Address/ZIP code	●	●	●	●	●	●	●	●	●	X	X	●	●	●
Insurance history	●	●	●	●	●	●	●	●	●	●	●	●	●	●

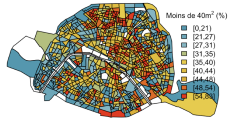
CA: Californie, HI: Hawaii, GA: Georgia, NC: Caroline du nord, NY: New York, MA: Massachusetts, PA: Pennsylvanie, FL: Floride, TX: Texas, AL: Alberta, ON: Ontario, NB: Nouveau-Brunswick, NL: Terre-Neuve-et-Labrador, QC: Québec

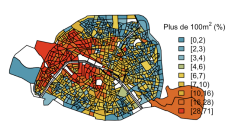
Protected Attributes ?

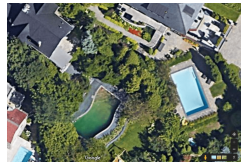
- location (policyholder home adress)

  Jean et al. (2016), Seresinhe et al. (2017), Gebru et al. (2017), Law et al. (2019), Ilic et al. (2019), Kita and Kidziński (2019), see also redlining

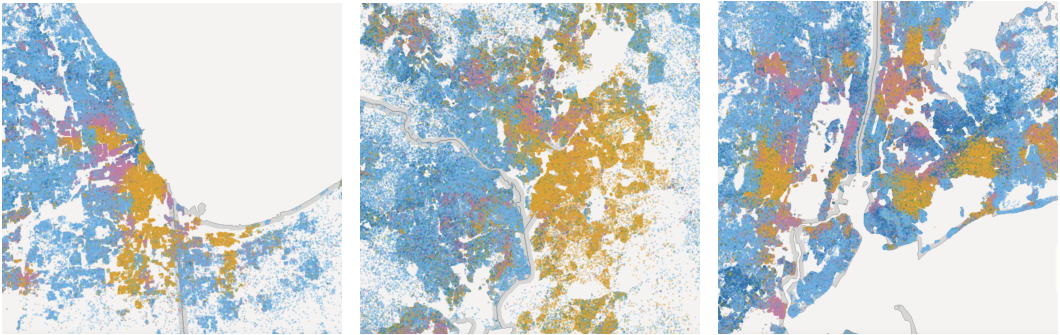








Protected Attributes ?



Protected Attributes ?

female, age: 38
female (0.997)
age: 34
joy (74%)

female, age: 23
female (0.989)
age: 20
joy (85%)

male, age: 37
male (0.967)
age: 27
joy (81%)

male, age: 53
male (0.985)
age: 38
joy (73%)

male, age: 37
male (0.996)
age: 38
joy (56%)

Faces generated by Karras et al. (2020).

Algorithms are from https://gender.toolpie.com/, https://picpurify.com/
https://cloud.google.com/vision/, https://howolddoyoulook.com/ and https://www.facialage.com/.

Protected Attributes ?

male, age: 24
male (0.944)
age: 26
joy (70%)

male, age: 33
male (0.981)
age: 32
joy (81%)

male, age: 34
female (0.905)
age: 34
joy (82%)

male, age: 48
male (0.989)
age: 48
joy (83%)

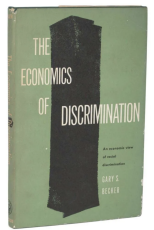
male, age: 43
male (0.984)
age: 38
joy (78%)

Faces generated by Karras et al. (2020).

Algorithms are from https://gender.toolpie.com/, https://picpurify.com/
https://cloud.google.com/vision/, https://howolddoyoulook.com/ and https://www.facialage.com/.

Protected Attributes ?

Bohren et al. (2019) on statistical discrimination, or efficient discrimination, as in Becker (1957) (inspired by Edgeworth (1922) up to Phelps (1972))



Becker (2005) says "*if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism*",

"*racial profiling*" is "*effective*", even though "*such profiling is 'unfair' to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest* (...) *That could be made up in part by compensating groups who are forced to go through more careful airport screening through putting them in shorter security lines, or in other ways. Similarly, innocent shoppers who are stopped and searched could be compensated for their embarrassment and time*"

Defining Group Fairness when y is binary

$$\begin{cases} \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ p \in \{0,1\} : \text{protected variable or sensitive attribute} \\ y \in \{0,1\} : \text{variable of interest (binary)} \\ s = m(\boldsymbol{x}, p) : \text{score} \\ \widehat{y} = \mathbf{1}(s > \text{threshold}) : \text{prediction} \end{cases}$$

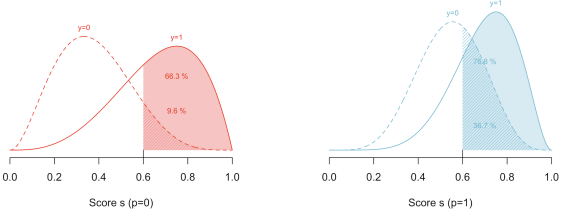
Fairness Through Unawareness, Kusner et al. (2017)
Protected attribute p is not explicitly used in decision function $\widehat{y}$.

Defining Group Fairness when y is binary

Demographic Parity, (Corbett-Davies et al. (2017), Agarwal (2021))
Decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp P$, i.e.

$$\mathbb{P}[\widehat{Y} = y|P = 0] = \mathbb{P}[\widehat{Y} = y|P = 1], \ \forall y \text{ or } \mathbb{E}[\widehat{Y}|P = 0] = \mathbb{E}[\widehat{Y}|P = 1]$$

We can compare $s(\boldsymbol{X})$ conditional on Y, but also on P



Defining Group Fairness when y is binary

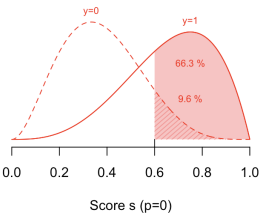
Equal Opportunity, Hardt et al. (2016)

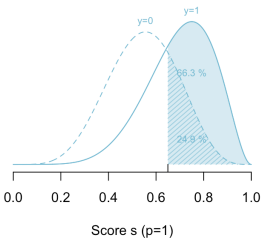
True positive parity

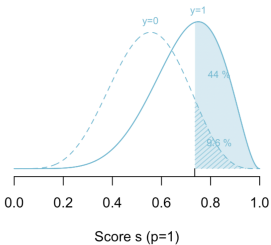
$$\mathbb{P}[\widehat{Y} = 1|P = 0, Y = 1] = \mathbb{P}[\widehat{Y} = 1|P = 1, Y = 1]$$

or false positive parity

$$\mathbb{P}[\widehat{Y} = 1|P = 0, Y = 0] = \mathbb{P}[\widehat{Y} = 1|P = 1, Y = 0]$$







Defining Group Fairness when y is binary

statistical parity	Dwork et al. (2012)	$\mathbb{P}[\widehat{Y} = 1 \vert P = p] = \text{cst}, \ \forall p$	independence
conditional statistical parity	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y} = 1 \vert P = p, X = x] = \text{cst}_x, \ \forall p, y$	$\widehat{Y} \perp\!\!\!\perp P$
equalized odds	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y} = 1 \vert P = p, Y = y] = \text{cst}_y, \ \forall p, y$	separation
equalized opportunity	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y} = 1 \vert P = p, Y = 1] = \text{cst}, \ \forall p$	
predictive equality	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y} = 1 \vert P = p, Y = 0] = \text{cst}, \ \forall p$	$\widehat{Y} \perp\!\!\!\perp P \mid Y$
balance (positive)	Kleinberg et al. (2017)	$\mathbb{E}[S \vert P = p, Y = 1] = \text{cst}, \ \forall p$	$S \perp\!\!\!\perp P \mid Y$
balance (negative)	Kleinberg et al. (2017)	$\mathbb{E}[S \vert P = p, Y = 0] = \text{cst}, \ \forall p$	
conditional accuracy equality	Berk et al. (2017)	$\mathbb{P}[Y = y \vert P = p, \widehat{Y} = y] = \text{cst}_y, \ \forall p, y$	sufficiency
predictive parity	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert P = p, \widehat{Y} = 1] = \text{cst}, \ \forall p$	
calibration	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert P = p, S = s] = \text{cst}_s, \ \forall p, s$	$Y \perp\!\!\!\perp P \mid \widehat{Y}$
well-calibration	Chouldechova (2017)	$\mathbb{P}[Y = 1 \vert P = p, S = s] = s, \ \forall p, s$	
accuracy equality	Berk et al. (2017)	$\mathbb{P}[\widehat{Y} = Y \vert P = p] = \text{cst}, \ \forall p$	
treatment equality	Berk et al. (2017)	$\frac{\text{FN}_p}{\text{FP}_p} = \text{cst}_p, \ \forall p$	

Defining Group Fairness when y is binary

Calibration, Krüger and Ziegel (2021) "the forecast X of Y is an auto-calibrated forecast of Y if $\mathbb{E}(Y|X) = X$ almost surely ", or $\mathbb{E}(Y|\widehat{Y} = y) = y, \ \forall y$

"*Suppose that a forecaster sequentially assigns probabilities to events. He is* **well calibrated** *if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent.* ", **?**,

(Well) Calibration, Chouldechova (2017) We have calibration parity if

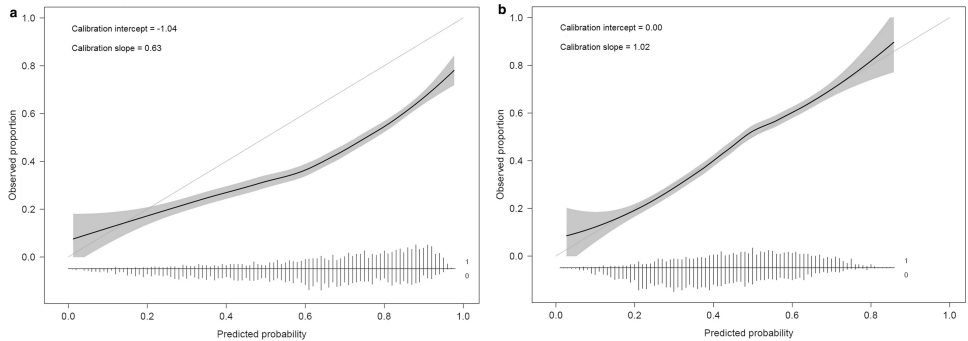
$$\mathbb{P}[Y = 1|\widehat{Y} = y, P = 0] = \mathbb{P}[Y = 1|\widehat{Y} = y, P = 1], \ \forall y \in (0, 1).$$

We have an fairness of good calibration if

$$\mathbb{P}[Y = 1|\widehat{Y} = y, P = 0] = \mathbb{P}[Y = 1|\widehat{Y} = y, P = 1] = y, \ \forall y \in (0, 1).$$

Defining Group Fairness when y is binary

See Denuit et al. (2021) for more details



Source: Van Calster et al. (2019) Calibration: the Achilles heel of predictive analytics

Dependence measures and discrimination mitigation

Group fairness is characterized be independence or conditional independence properties. Given two random variables U and V,

$$C(U,V) = \begin{cases} \text{corr}[U,V] \text{ Pearson's correlation} \\ \text{corr}[F_U(U), F_V(V)] \text{ Spearman's rank correlation} \\ \tau[U,V] \text{ Kendall's tau} \end{cases}$$

that can be extended to conditional measures, as Lawrance (1976), since

$$\text{corr}[U,V] = \mathbb{E}\big[UV\big] \text{ when } \begin{cases} \mathbb{E}[U] = \mathbb{E}[V] = 0 \\ \mathbb{E}[U^2] = \mathbb{E}[V^2] = 1 \end{cases}$$

$$\begin{cases} \textbf{Demographic Parity}: \ \widehat{Y} \perp\!\!\!\perp P & \implies C(\widehat{Y}, P) = 0 \\ \textbf{Equalized Odds}: \ \widehat{Y} \perp\!\!\!\perp P | Y & \implies C(\widehat{Y}, P | Y = y) = 0, \ \forall y \end{cases}$$

Dependence measures and discrimination mitigation

Hirschfeld (1935), Gebelein (1941) and Rényi (1959)

$$HGR(U,V) = \max\left\{\text{corr}[f(U), g(V)]\right\} = \max_{f\in\mathcal{S}_U, g\in\mathcal{S}_V}\left\{\mathbb{E}[f(U)g(V)]\right\}$$

where $\mathcal{S}_U = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}[f(U)] = 0 \text{ and } \mathbb{E}[f(U)^2] = 1\}$ and similarly $\mathcal{S}_V$.
One can also consider a conditional version,

$$HGR(U,V|Z) = \max_{f\in\mathcal{S}_{U|Z}, g\in\mathcal{S}_{V|Z}}\Big\{\mathbb{E}[f(U)g(V)|Z]\Big\}$$

where $\mathcal{S}_{U|Z} = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}[f(U)|Z] = 0 \text{ and } \mathbb{E}[f(U)^2|Z] = 1\}$.

This measure can be used to characterize independence,

$$U \perp\!\!\!\perp V \iff HGR(U,V) = 0,$$

and if (U,V) is a Gaussian vector, $HGR(U,V) = |\text{corr}(U,V)|$.

Dependence measures and discrimination mitigation

Thus, this measure can be used to characterize fairness,

$$\begin{cases} \textbf{Demographic Parity}: \ \widehat{Y} \perp\!\!\!\perp P & \iff HGR(\widehat{Y}, P) = 0 \\ \textbf{Equalized Odds}: \ \widehat{Y} \perp\!\!\!\perp P | Y & \iff HGR(\widehat{Y}, P | Y = y) = 0, \ \forall y \end{cases}$$

HGR can be difficult to estimate, but one can use some Neural Networks

$$HGR_{NN}(U, V) = \max_{\omega_f, \omega_g} \left\{ \mathbb{E}[f_{\omega_f}(U) g_{\omega_g}(V)] \right\}$$

See also Breiman and Friedman (1985) on the estimation of this maximal correlation, in the context of regression

Dependence measures and discrimination mitigation

More generally, $\boldsymbol{V}$ can be a vector on $\mathcal{V} \subset \mathbb{R}^k$, then

$$HGR(U, \boldsymbol{V}) = \max_{h:\mathcal{V}\to\mathbb{R}} \left\{ HGR[U, h(\boldsymbol{V})] \right\} = \max_{f\in\mathcal{S}_U, g\in\mathcal{S}_V} \left\{ \mathbb{E}[f(U)g(\boldsymbol{V})] \right\}$$

where $\mathcal{S}_V = \{g : \mathcal{V} \to \mathbb{R} : \mathbb{E}[g(\boldsymbol{V})] = 0 \text{ and } \mathbb{E}[g(\boldsymbol{V})^2] = 1\}$. A conditional version exists, and one can estimate that measure using a neural network,

$$HGR_{NN}(U, \boldsymbol{V}) = \max_{\boldsymbol{\omega}_f, \boldsymbol{\omega}_g} \left\{ \mathbb{E}[f_{\boldsymbol{\omega}_f}(U) g_{\boldsymbol{\omega}_g}(\boldsymbol{V})] \right\}$$

$$\begin{cases} \textbf{Demographic Parity}: \ \widehat{Y} \perp\!\!\!\perp \boldsymbol{P} & \iff HGR(\widehat{Y}, \boldsymbol{P}) = 0 \\ \textbf{Equalized Odds}: \ \widehat{Y} \perp\!\!\!\perp \boldsymbol{P} | Y & \iff HGR(\widehat{Y}, \boldsymbol{P} | Y = y) = 0, \ \forall y \end{cases}$$

In-processing mitigation and adversarial approach

In a classical ML or econometric pricing model, solve

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \mathcal{L}\left(\widehat{y}, y\right) \right\}, \text{ where } \mathcal{L}\left(\widehat{y}, y\right) = \sum_{i=1}^{n} \ell\left(\widehat{y}_i, y_i\right) \text{ and } \widehat{y} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

either related to some loss, or some log-likelihood,
To avoid over-fit: penalize complexity (penalty $\mathcal{P}$)

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \mathcal{L}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right) + \lambda \mathcal{P}(h_{\boldsymbol{\theta}}) \right\}$$

In-processing mitigation and adversarial approach

Inspired by Goodfellow et al. (2018) (but also Bechavod and Ligett (2017) or Cho et al. (2020)), to avoid un-fairness: penalize according to $HGR(\widehat{y}, p)$ (for demographic parity),

$$\underset{\boldsymbol{\theta},\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}{\operatorname{argmin}} \left\{ \mathcal{L}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\big) + \lambda \text{HGR}_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}(h_{\boldsymbol{\theta}}(\boldsymbol{x}), p) \right\}$$

i.e.

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \max_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g} \left\{ \mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) + \lambda \mathbb{E}_{(\boldsymbol{X},S)\sim\mathcal{D}}(\widehat{f}_{\boldsymbol{\omega}_f}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_g}(P) \right\} \right\}$$

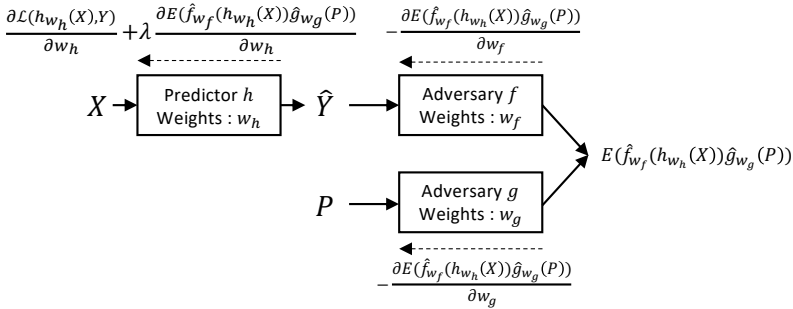
or $HGR(\widehat{y}, p|y)$ (for equalized odds), i.e. when $y \in \{0, 1\}$

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Bigg\{ \max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\boldsymbol{\omega}_{f1},\boldsymbol{\omega}_{g1}} \Big\{ \mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) &+ \lambda_0 \mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_0}(\widehat{f}_{\boldsymbol{\omega}_{f_0}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_0}}(P)) \\ &+ \lambda_1 \mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_1}(\widehat{f}_{\boldsymbol{\omega}_{f_1}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_1}}(P)) \Big\} \Bigg\} \end{aligned}$$

or, more generally when $y \in \Omega_Y$ (e.g. $\{0, 1, 2, 3+\}$), if $k = \#\Omega_y$

In-processing mitigation and adversarial approach

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{ \max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\cdots,\boldsymbol{\omega}_{fk},\boldsymbol{\omega}_{gk}} \left\{ \mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) + \sum_{y\in\Omega_y} \lambda_y \mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_y}(\widehat{f}_{\boldsymbol{\omega}_{f_y}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_y}}(P)) \right\} \right\}$$



Post-processing mitigation

Consider k models, $m_1, \cdots, m_k$.

- classification problem, $m_j(\boldsymbol{x}) \approx \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$
- regression problem, $m_j(\boldsymbol{x}) \approx \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$

With a bagging approach, $M(\boldsymbol{x}) = \sum_{j=1}^{k} \omega_j m_j(\boldsymbol{x}) = \boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x})$.

Following Friedler et al. (2019), we can solve

$$\min_{\boldsymbol{\omega} \in \mathcal{S}_k} \left\{ \left| \frac{1}{n_1} \sum_{i:P_i=1} \boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i) - \frac{1}{n_0} \sum_{i:P_i=0} \boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i) \right|^\alpha + \frac{\lambda}{n} \sum_i \ell\big(\boldsymbol{\omega}^\top \boldsymbol{m}(\boldsymbol{x}_i), y_i\big) \right\}$$

for some $\alpha > 0$, where $\mathcal{S}_k = \big\{ \boldsymbol{w} \in \mathbb{R}_+^k : \boldsymbol{w}^\top \mathbf{1} = 1 \big\}$.

Post-processing mitigation

As shown in Fermanian and Guegan (2021), if $\alpha = 2$ and $\ell = \ell_2$, then

$$\boldsymbol{\omega}_{dp}^{\star} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^{\top}\Sigma^{-1}\mathbf{1}} \text{ where } \Sigma = \boldsymbol{A}\boldsymbol{A}^{\top} + \lambda \boldsymbol{B},$$

where

$$\boldsymbol{A} = \frac{1}{n_1} \sum_{i:P_i=1} \boldsymbol{m}(\boldsymbol{x}_i) - \frac{1}{n_0} \sum_{i:P_i=0} \boldsymbol{m}(\boldsymbol{x}_i)$$

and

$$\boldsymbol{B} = \frac{1}{n} \sum_{i=1}^{n} \big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i \mathbf{1}\big)\big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i \mathbf{1}\big)^{\top}.$$

Post-processing mitigation

For equalized odds, the optimization problem is

$$\min_{\boldsymbol{\omega}\in\mathcal{S}_{\|}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left|\widehat{e}^{(1)}(y_i) - \widehat{e}^{(0)}(y_i)\right|^{\alpha} + \frac{\lambda}{n}\sum_{i} \ell\big(\boldsymbol{\omega}^{\top}\boldsymbol{m}(\boldsymbol{x}_i), y_i\big) \right\}$$

where $\widehat{e}^{(s)}(y_i)$ is an estimation of $\mathbb{E}[\widehat{Y}|Y = y_i, S = s]$. For instance, consider some standard nonparametric estimators, kernel based,

$$\widehat{e}^{(s)}(y) \sum_{i:s_i=s} \boldsymbol{\omega}^{\top}\boldsymbol{m}(\boldsymbol{x}_i)K_h(y_i - y) = \boldsymbol{\omega}^{\top}\boldsymbol{v}_h^{(s)}(y)$$

Post-processing mitigation

As shown in Fermanian and Guegan (2021), if $\alpha = 2$ and with $\ell = \ell_2$,

$$\boldsymbol{\omega}_{eo}^{\star} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^{\top}\Sigma^{-1}\mathbf{1}} \text{ where } \Sigma = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^{\top} + \lambda\boldsymbol{B},$$

where

$$\boldsymbol{\gamma}_i = \boldsymbol{v}_h^{(0)}(y_i) - \boldsymbol{v}_h(y_i)$$

and

$$\boldsymbol{B} = \frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i\mathbf{1}\big)\big(\boldsymbol{m}(\boldsymbol{x}_i) - y_i\mathbf{1}\big)^{\top}.$$

From correlation to causality

- "*classifying projection methods as using demographic/actuarial models or non-demographic/causal models*"

  Keilman (2003) and Hudson (2007)

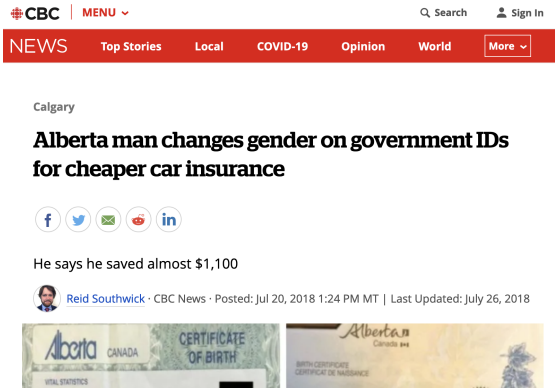
- "*Article 5(2) allowed Member States to Permit proportionate differences in individuals premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data.*"

  Thiery and Van Schoubroeck (2006) and Schmeiser et al. (2014)

- "*Two judges on the Supreme Court dissented in the Zurich case. In their view, an insurer must not only prove a statistical correlation between a particular group and higher risk, but a causal connection*"

  Gomery et al. (2011)

From correlation to causality


CBC | MENU Search Sign In

NEWS Top Stories Local COVID-19 Opinion World More

Calgary

Alberta man changes gender on government IDs for cheaper car insurance

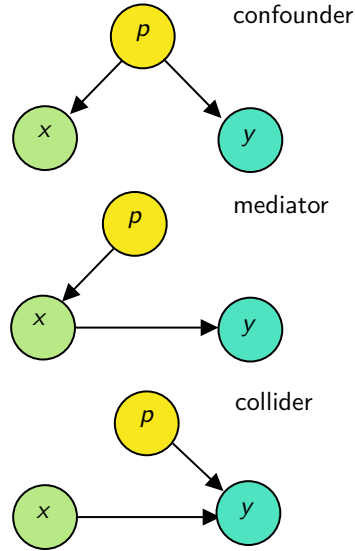
He says he saved almost $1,100

Reid Southwick · CBC News · Posted: Jul 20, 2018 1:24 PM MT | Last Updated: July 26, 2018

Alberta CANADA CERTIFICATE OF BIRTH

Alberta Canada

- DAGs are important
- Looking for a counterfactual



From correlation to causality

Consider some distances D on $\{0,1\} \times \{0,1\}$ or $[0,1] \times [0,1]$, and d on $\mathbb{R}^p \times \mathbb{R}^p$,

Lipschitz property, Duivesteijn and Feelders (2008)

$$D(\widehat{y}_i, \widehat{y}_j) \text{ or } D(s_i, s_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_j), \ \forall i, j = 1, \cdots, n.$$

Conterfactual fairness, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

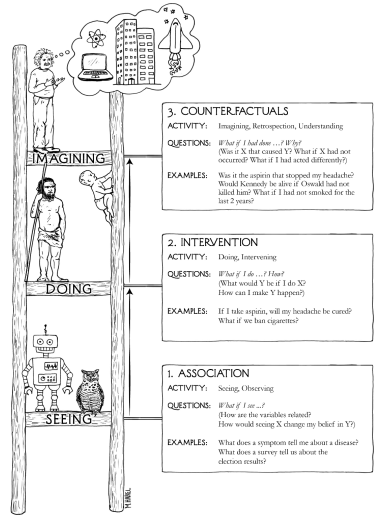
$$\mathbb{P}\big[Y^{\star}_{P \leftarrow p} = y | \boldsymbol{X} = \boldsymbol{x}\big] = \mathbb{P}\big[Y^{\star}_{P \leftarrow p'} = y | \boldsymbol{X} = \boldsymbol{x}\big], \ \forall p', \boldsymbol{x}, y.$$

From correlation to causality

- counterfactuals
  (*what if I had done...?*)
- intervention
- association
  (*what if I see...?*)

what would have happened if this person had had treatment 1 instead of treatment 0 ?

(picture Pearl & Mackenzie (2018))



From correlation to causality

Causal inference litterature,

- t some binary treatment ($t \in \{0, 1\}$)
- $\boldsymbol{x}$ some covariates
- y denote the observed outcome, $y^{\star}_{i,T\leftarrow 1}$ and $y^{\star}_{i,T\leftarrow 0}$ the potential outcomes

	treatement	outcome			age	gender	height	weight
	t_i	y_i	$y^{\star}_{i,T\leftarrow 1}$	$y^{\star}_{i,T\leftarrow o}$	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$	$x_{4,i}$
1	1	121	121	?	37	F	160	56
2	0	109	?	109	28	F	156	54
3	1	162	162	?	53	M	190	87

There will be a significant impact of treatment t on y if $y^{\star}_{T\leftarrow 0} \neq y^{\star}_{T\leftarrow 1}$ (see Rubin (1974), Hernán and Robins (2010) or Imai (2018)).
The causal effect for individual i is $\tau_i = y^{\star}_{i,T\leftarrow 1} - y^{\star}_{i,T\leftarrow 0}$

From correlation to causality

One can define the sample average treatment effect (SATE)

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n} y_{i,T\leftarrow 1}^{\star} - y_{i,T\leftarrow 0}^{\star}$$

the average treatment effect (ATE)

$$\tau = \text{ATE} = \mathbb{E}\big[Y_{i,T\leftarrow 1}^{\star} - Y_{i,T\leftarrow 0}^{\star}\big]$$

and, for possibly heterogeneous effects, conditional average treatment effect (CATE)

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y_{i,T\leftarrow 1}^{\star} - Y_{i,T\leftarrow 0}^{\star}\big|\boldsymbol{X} = \boldsymbol{x}\big]$$

See Charpentier et al. (2023) for some extention with optimal transport.

References

,

Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Barry, L. and Charpentier, A. (2022). The Fairness of Machine Learning in Insurance: New Rags for an Old Man? . *ArXiv*.

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv*, 1707.00044.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Becker, G. S. (2005). Is ethnic and other profiling discrimination? *The Becker-Posner Blog*, 01-23-2005.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.

References

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Charpentier, A. (2022a). *Assurance: biais, discrimination et équité*. Institut Louis Bachelier.

Charpentier, A. (2022b). *Insurance: biases, discrimination and fairness*. Institut Louis Bachelier.

Charpentier, A. (2023). Quantifying fairness and discrimination in predictive models. In Kreinovich, V., SriboonchiNa, S., and Yamaka, W., editors, *Machine Learning for Econometrics and Related Topics*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Cho, J., Hwang, G., and Suh, C. (2020). A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2521–2526. IEEE.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

References

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Edgeworth, F. Y. (1922). Equal pay to men and women for equal work. *The Economic Journal*, 32(128):431–457.

Fermanian, J.-D. and Guegan, D. (2021). Fair learning with bagging. *SSRN*, 3969362.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

References

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

Gomery, S., Renault, O., and John, N. (2011). Gender neutral. *Canadian Underwriter*.

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520524.

Hudson, R. (2007). Mortality projections and unisex pricing of annuities in the uk. *Journal of Financial Regulation and Compliance*.

Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

Ilic, L., Sawada, M., and Zarzelli, A. (2019). Deep mapping gentrification in a large canadian city using deep learning and google street view. *PloS one*, 14(3):e0212814.

References

Imai, K. (2018). *Quantitative social science: an introduction*. Princeton University Press.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Keilman, N. (2003). Types of models for projecting mortality. *Perspectives on mortality forecasting*, 1:19–27.

Kita, K. and Kidziński, Ł. (2019). Google street view image of a house predicts car accident risk of its resident. *arXiv*, 1904.05270.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

References

Krüger, F. and Ziegel, J. F. (2021). Generic conditions for forecast dominance. *Journal of Business & Economic Statistics*, 39(4):972–983.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Law, S., Paige, B., and Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology*, 10(5).

Lawrance, A. (1976). On conditional and partial correlation. *The American Statistician*, 30(3):146–149.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schmeiser, H., Störmer, T., and Wagner, J. (2014). Unisex insurance pricing: consumers perception and market implications. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 39(2):322–350.

References

Seresinhe, C. I., Preis, T., and Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7):170170.

The Zebra (2022). Car insurance rating factors by state. *https://www.thezebra.com/*.

Thiery, Y. and Van Schoubroeck, C. (2006). Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.