SCIENCES SORBONNE UNIVERSITÉ

eXplainable Artificial Intelligence

Cours 12 : Fairness

presented by Jean-Noël Vittaut

This lecture is a presentation of some of V. Grari's work :
Adversarial mitigation to reduce unwanted biases in machine learning, PhD thesis, Sorbonne Université, 2022.

- GDPR (General Data Protection Regulation)
- Strictly regulates the collection and use of sensitive personal data. With the aim of obtaining non-discriminatory algorithms
- Article 9(1) : "Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.".

1 Group fairness
- Overview
- Problem statement
- Measuring fairness
- Ensuring group fairness

2 Individual fairness
- Overview
- Problem statement
- Measuring fairness

3 Conclusion

4 TME

1 Group fairness
- Overview
- Problem statement
- Measuring fairness
- Ensuring group fairness

2 Individual fairness
- Overview
- Problem statement
- Measuring fairness

3 Conclusion

4 TME

Groups can be subject to unfair decisions

- men vs. women ; Caucasian vs. non-Caucasians ; ...

Measures

- statistical independance : *Demographic parity*
- statistical separation criterion : *Equalized odds* ; *Residuals parity*

Mitigation algorithms

- pre-processing : suppressing sensitive attributes ; changing class labels ; resampling
- post-processing : modifying the output of the classifier
- in-processing : penalty in the cost function ; adversarial methods

Mitigation algorithms

- Prediction retreatment
- Fair representation

Classification or regression problem

- predict Y from $X \in \mathbb{R}^d$
- S : sensitive attributes

Demographic parity (Dwork et al., 2012)

- Output prediction $\hat{Y}$ from features X is independent of the sensitive attribute S
- $\mathbb{P}\{\hat{Y}|S\} = \mathbb{P}\{\hat{Y}\}$

Example : patient tumors and gender

- Classification model (0 = no tumor, 1 = tumor)
- Sensitive feature : gender
- Model predicts more tumors on male than female

the positive rate would be equal for males and females
$\rightarrow$ increase the predictive error by detecting fewer tumors for males and more for females

Equalized odds

Equalized odds (Hardt et al., 2016)

- output prediction $\hat{Y}$ from features X is independent of the sensitive attribute S given the outcome true value Y
- $\mathbb{P}\{\hat{Y}|S, Y\} = \mathbb{P}\{\hat{Y}|Y\}$
- $\mathbb{P}\{\hat{Y}|S, Y = 1\} = \mathbb{P}\{\hat{Y}|Y = 1\}$: *equalized opportunity*

Example : patient tumors and gender

- Classification model (0 = no tumor, 1 = tumor)
- Sensitive feature : gender
- Model predicts more tumors on male than female

false-positive and false-negative rates will be the same for males and females
$\rightarrow$ more appropriate for this medical application

Equalized residuals

Equalized residuals (Grari et al., 2020)

- Residuals $\hat{Y} - Y$ from features X is independent of the sensitive attribute S
- $\mathbb{P}\{\hat{Y} - Y | S\} = \mathbb{P}\{\hat{Y} - Y\}$

Example : car insurance pricing and age

- Regression model (real cost)
- Sensitive feature : age (younger vs older)

Demographic parity : older people pay more than their real cost

Equalized residuals : residuals between the predictions and the real claim cost are preserved, independently from the sensitive variable age

→ What is more appropriate ?

How to mathematically quantify the previous fairness objectives ?

Binary setting

- predict $Y \in \{0, 1\}$
- $S \in \{0, 1\}$: sensitive attribute

Demographic Parity

- Each demographic group has the same chance for a positive outcome
- A classifier is considered fair according to the demographic parity principle if $\mathbb{P}\{\hat{Y} = 1 | S = 0\} = \mathbb{P}\{\hat{Y} = 1 | S = 1\}$

p-rule

- $\min \left\{ \frac{\mathbb{P}\{\hat{Y}=1|S=1\}}{\mathbb{P}\{\hat{Y}=1|S=0\}}, \frac{\mathbb{P}\{\hat{Y}=1|S=0\}}{\mathbb{P}\{\hat{Y}=1|S=1\}} \right\}$
- 100% rule : totally fair
- 0% rule : totally unfair

Disparate Impact (DI) assessment (Feldman et al., 2015)

- Absolute difference of outcome distributions for subpopulations with different sensitive attribute values
- $|\mathbb{P}\{\hat{Y} = 1 | S = 1\} - \mathbb{P}\{\hat{Y} = 1 | S = 0\}|$
- The smaller the difference, the fairer the model

Equalized Odds

- $\mathbb{P}\{\hat{Y}|S, Y\} = \mathbb{P}\{\hat{Y}|Y\}$

Disparate Mistreatment (DM) (Zafar et al., 2017)

- Absolute difference between the false positive rate (FPR) and the false-negative rate (FNR) for both demographics
- $D_{FPR} = |\mathbb{P}\{\hat{Y} = 1|Y = 0, S = 1\} - \mathbb{P}\{\hat{Y} = 1|Y = 0, S = 0\}|$
- $D_{FNR} = |\mathbb{P}\{\hat{Y} = 0|Y = 1, S = 1\} - \mathbb{P}\{\hat{Y} = 0|Y = 1, S = 0\}|$
- chances of being correctly (or incorrectly) classi- fied as positive should be the same across groups.
- The closer the values of D_{FPR} and D_{FNR} to 0, the lower the degree of disparate mistreatment of the classifier

Continuous setting

- predict $Y \in \mathbb{R}$
- $S \in \mathbb{R}$: sensitive attribute

Measuring dependence

- assessing fairness $\rightarrow$ measuring statistical dependance
- Pearson's correlation
- Kendall's tau
- Spearman's rank correlation
- ...
- $\rightarrow$ only capture a limited class of association patterns.

Dependence via Information Theory

Mutual Information

- $I(U,V) = \int_{\mathbb{R}} \int_{\mathbb{R}} P_{UV}(u,v) * \log(Q(u,v)) dudv$
- with $Q(u,v) = \frac{P_{UV}(u,v)}{\sqrt{P_U(u)}\sqrt{P_V(v)}}$

χ^2 divergence

- $\chi^2(P_{UV}, P_U.P_V) = \int_{\mathbb{R}} \int_{\mathbb{R}} Q(u,v) dudv - 1$

Drawbacks

- difficult to measure, interpret, compute
- usual method : estimate the density function via KDE (Kernel Density Estimation)

Dual representation (Broniatowski and Leorato, 2006)

- The χ^2 divergence admits the following representation :
  $\chi^2(P, Q) = \sup_f \mathbb{E}(f(P)) - \mathbb{E}(f(Q) + \frac{1}{4}f^2(Q))$

Algorithm 1 χ^2 Neural Estimation

Input: Distributions $P_{U,V}$ and P_V, Neural Network f_θ, Batchsize b, Learning rate α
repeat
Draw b samples from the joint distribution:
$(u_1, v_1), ..., (u_b, v_b) \sim P_{UV}$
Draw b samples from the V marginal distribution:
$\bar{v}_1, ..., \bar{v}_b \sim P_V$
Evaluate the lower bound:
$J(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_\theta(u_i, v_i) - \frac{1}{b}\sum_{i=1}^{b}(f_\theta(u_i, \bar{v}_i) + \frac{1}{4}f_\theta(u_i, \bar{v}_i)^2)$
Update the network parameters by gradient ascent:
$\theta \leftarrow \theta + \alpha \nabla J(\theta)$
until convergence

Measure of concordance (Hirschfeld, 1935) and (Gebelein, 1941)

- $HGR(U, V) = \max_{f,g} \{r(f(U), g(V))\}$
  where r is Pearson's correlation

Alternative expression

- $HGR(U, V) = \max_{f \in \mathcal{S}(U), g \in \mathcal{S}(V)} \{\mathbb{E}(f(U)g(V))\}$
  where $\mathcal{S}(X) = \left\{f : \mathbb{E}(f(X)) = 0 \text{ and } \mathbb{E}(f^2(X)) = 1\right\}$ (standardized transformations)

HGR coefficient is equal to 0 if, and only if, the two random variables are independent.

Algorithm 2 HGR Estimation by Neural Network

Input: Distributions $P_{U,V}$, Neural Networks f_{ω_f} and g_{ω_g}, Batchsize b, Learning rates α_f, α_g

repeat

Draw b samples from the joint distribution:

$(u_1, v_1), ..., (u_b, v_b) \sim P_{UV}$

Calculate the average and variance of the transformation predictions:

$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\omega_f}(u_i)\ ;\ \sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{\omega_f}(u_i) - m_f)^2$

$m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{\omega_g}(v_i)\ ;\ \sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{\omega_g}(v_i) - m_g)^2$

Standardize w.r.t. the minibatch:

$\forall i : \hat{f}_{\omega_f}(u_i) \leftarrow \frac{f_{\omega_f}(u_i) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}; \hat{g}_{\omega_g}(v_i) \leftarrow \frac{g_{\omega_g}(v_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$

Maximize the following objective function J by gradient ascent:

$J(\omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{\omega_f}(u_i) * \hat{g}_{\omega_g}(v_i)$

$\omega_f \leftarrow \omega_f + \alpha_f \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_f};\ \omega_g \leftarrow \omega_g + \alpha_g \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_g}$

until convergence

Demographic parity

- A machine learning algorithm achieves Demographic Parity if the associated prediction $\hat{Y}$ and the sensitive attribute S satisfies : $HGR(\hat{Y}, S) = 0$

FairQuant metric (Grari et al., 2020)

- Metric based on discretization of the sensitive attribute
- Splits the set samples X in K quantiles with regards to the sensitive attribute
- We define K as the number of quantiles, m_k as the mean of the predictions $h(X_k)$ in the k-th quantile set X_k, and m its mean on the full sample X
- $FairQuant = \frac{1}{K} \sum_{k=1}^{K} |m_k - m|$

Equalized residuals

- A machine learning algorithm achieves equalized residuals if the associated residuals $\hat{Y} - Y$ and the sensitive attribute S satisfies : $HGR(\hat{Y} - Y, S) = 0$

FairQuant metric (Grari et al., 2020)

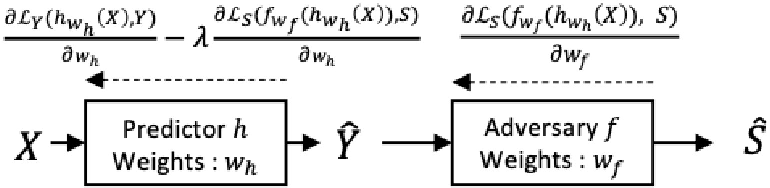
- FairQuant on the mean of the residuals

Achieving the demographic parity

- $\arg\min_h \{\mathcal{L}(h(X), Y) + \lambda p(h(X), S)\}$
- p : penalization term which evaluates the correlation loss between two variables

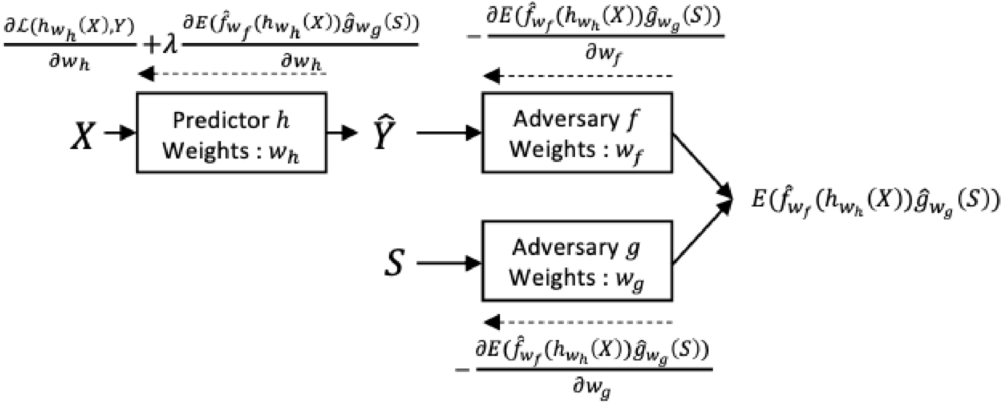
Adversarial simple architecture

- $\arg\min_h \left\{\max_f \left\{\mathcal{L}_{pred}(h(X), Y) + \lambda \mathcal{L}_{adv}(f(h(X)), S)\right\}\right\}$



Adversarial HGR architecture (Grari et al., 2020)

- arg $\min_h \left\{ \max_{f,g} \left\{ \mathcal{L}_{pred}(h(X), Y) + \lambda \mathbb{E}_{XS}(\hat{f}(h(X))\hat{g}(S) \right\} \right\}$



Fair algorithm (Grari et al., 2020)

Algorithm 3 Fair Rényi Algorithm for Demographic Parity

Input: Training set $\mathcal{T}$, Loss function $\mathcal{L}$, Batchsize b,
Neural Networks h_{ω_h}, f_{ω_f} and g_{ω_g},
Learning rates α_f, α_g and α_h, Fairness control λ

Repeat

Draw b samples $(x_1, s_1, y_1), ..., (x_b, s_b, y_b)$ from $\mathcal{T}$

Calculate the mean and variance of the transformations:

$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{\omega_f}(h_{\omega_h}(x_i))$; $m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{\omega_g}(s_i)$

$\sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{\omega_f}(h_{\omega_h}(x_i)) - m_f)^2$

$\sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{\omega_g}(s_i) - m_g)^2$

Standardize the transformations:

$\forall i : \hat{f}_{\omega_f}(h_{\omega_h}(x_i)) \leftarrow \frac{f_{\omega_f}(h_{\omega_h}(x_i)) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}$

$\forall i : \hat{g}_{\omega_g}(s_i) \leftarrow \frac{g_{\omega_g}(s_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$

Compute the objectives:

$J(\omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{\omega_f}(h_{\omega_h}(x_i)) * \hat{g}_{\omega_g}(s_i)$

$L(\omega_h, \omega_f, \omega_g) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{\omega_h}(x_i), y_i) + \lambda J(\omega_f, \omega_g)$

Update the adversary by gradient ascent:

$\omega_f \leftarrow \omega_f + \alpha_f \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_f}$; $\omega_g \leftarrow \omega_g + \alpha_g \frac{\partial J(\omega_f, \omega_g)}{\partial \omega_g}$

Update the predictor model h_{ω_h} by gradient descent:

$\omega_h \leftarrow \omega_h - \alpha_h (\frac{\partial L(\omega_h, \omega_f, \omega_g)}{\partial \omega_h})$

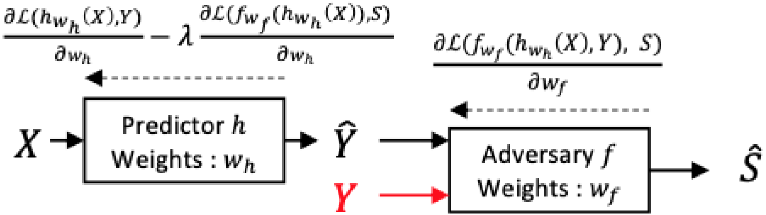
Improving equalized odds

- $\arg\min_h \{\mathcal{L}(h(X), Y) + \lambda p(h(X), S, \color{red}{Y})\}$
- the penalization term evaluates the correlation loss between the output prediction and the sensitive attribute given the expected outcome

SCIENCES SORBONNE UNIVERSITÉ

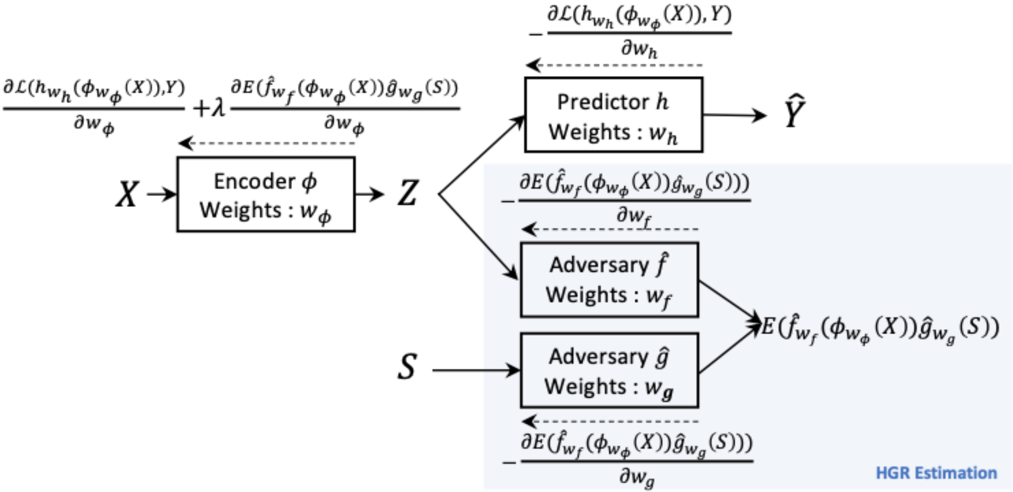
Adversarial simple architecture

- $\arg\min_h \{\max_f \{\mathcal{L}_{pred}(h(X), Y) + \lambda \mathcal{L}_{adv}(f(h(X), Y), S)\}\}$



Rényi adversarial architecture extension (Grari et al., 2020)

- $\min_{h,\phi}\left\{\max_{f,g}\left\{\mathcal{L}_{pred}(h(X),Y)+\lambda\mathbb{E}_{XS}(\hat{f}(\phi(X))\hat{g}(S)\right\}\right\}$



SCIENCES SORBONNE UNIVERSITÉ

Algorithm 4 Rényi Fair Representation

Input: Training set $\mathcal{T}$, Loss function $\mathcal{L}$, Batchsize b, Epochs for HGR n_{HGR}
Neural Networks ϕ_{w_ϕ},h_{w_h}, f_{w_f} and g_{w_g},
Learning rates α_f, α_g, α_ϕ and α_h. Fairness control λ

Repeat

Draw b samples $(x_1, s_1, y_1), ..., (x_b, s_b, y_b)$ from $\mathcal{T}$

Compute the predictor objective:

$L_Y(w_h, \phi_{w_\phi}) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{w_h}(\phi_{w_\phi}(x_i)), y_i)$

Update the predictor model h_{w_h} by gradient descent:

$w_h \leftarrow w_h - \alpha_h(\frac{\partial L_Y}{\partial w_h})$

Repeat n_{HGR} times

Calculate the mean and variance of the transformations:

$m_f \leftarrow \frac{1}{b}\sum_{i=1}^{b} f_{w_f}(\phi_{w_\phi}(x_i))$; $m_g \leftarrow \frac{1}{b}\sum_{i=1}^{b} g_{w_g}(s_i)$

$\sigma_f^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(f_{w_f}(\phi_{w_\phi}(x_i)) - m_f)^2$

$\sigma_g^2 \leftarrow \frac{1}{b}\sum_{i=1}^{b}(g_{w_g}(s_i) - m_g)^2$

Standardize the transformations:

$\forall i : \hat{f}_{w_f}(\phi_{w_\phi}(x_i)) \leftarrow \frac{f_{w_f}(\phi_{w_\phi}(x_i)) - m_f}{\sqrt{\sigma_f^2 + \epsilon}}$

$\forall i : \hat{g}_{w_g}(s_i) \leftarrow \frac{g_{w_g}(s_i) - m_g}{\sqrt{\sigma_g^2 + \epsilon}}$

Compute the objectives:

$J(w_f, w_g, w_\phi) = \frac{1}{b}\sum_{i=1}^{b} \hat{f}_{w_f}(\phi_{w_\phi}(x_i)) * \hat{g}_{w_g}(s_i)$

$L_E(w_h, w_\phi, w_f, w_g) = \frac{1}{b}\sum_{i=1}^{b} \mathcal{L}(h_{w_h}(\phi_{w_\phi}(x_i)), y_i) + \lambda J(w_f, w_g, w_\phi)$

Update the adversary by gradient ascent:

$w_f \leftarrow w_f + \alpha_f \frac{\partial J}{\partial w_f}$; $w_g \leftarrow w_g + \alpha_g \frac{\partial J}{\partial w_g}$

Update the encoder model ϕ_{w_ϕ} by gradient descent:

$w_\phi \leftarrow w_\phi - \alpha_\phi(\frac{\partial L_E}{\partial w_\phi})$

1 Group fairness
- Overview
- Problem statement
- Measuring fairness
- Ensuring group fairness

2 Individual fairness
- Overview
- Problem statement
- Measuring fairness

3 Conclusion

4 TME

Individual fairness

- *"similar people should be treated similarly"* $\rightarrow$ existence of similarity measure
  - finding individuals with the most disparate treatment
  - oracle indentifying fairness violations
- Counterfactual fairness
  - produce similar outcomes for every alternate version of any individual

Classification or regression problem

- X : attributes
- S : sensitive attributes
- predict $Y \in \mathcal{Y}$ from $X \in \mathcal{X}$
- predictor : $y = h(x)$
- $d(x, x')$: distance metric between individuals x and x'

Fairness Through Awareness (FTA) (Dwork et al., 2012)

- A predictor h achieves *Individual Fairness* with respect to a distance metric d on the input space $\mathcal{X}$ if h is K-lipschitz for a certain K.
- $\forall x, x' \in \mathcal{X}, |h(x) - h(x')| \leqslant Kd(x, x')$

$\rightarrow$ **The metric $d(\cdot, \cdot)$ must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling.**

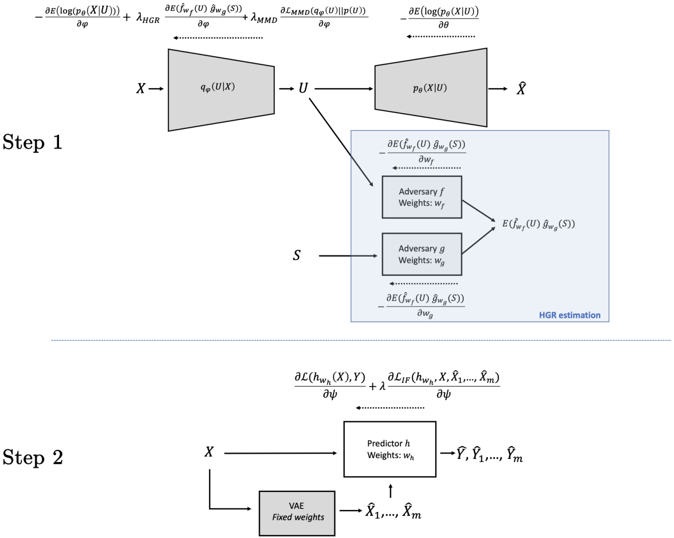
Example : promotion denied

- Classification model (promoted ; not promoted)
- Sensitive feature : privileged group

Demographic parity : not promoted for belonging to a privileged group

FTA : merit within the group is taken into account

Fairness Through Awareness (FTA)



Counterfactual Individual Fairness (Kusner et al., 2017)

- A predictor h achieves *Counterfactual Fairness* if under any context $X = x$ and $S = s$, the outcome probability is the same for individual $x_{S \leftarrow s}$ and counterfactual $x_{S \leftarrow s'}$ for all $s' \neq s$.
- $\forall y \in \mathcal{Y}, \forall s' \neq s, \mathbb{P}\{\hat{Y}_{S \leftarrow s} = y | X = x, S = s\} = \mathbb{P}\{\hat{Y}_{S \leftarrow s'} = y | X = x, S = s\}$

Example : accident rate

- Classification model (number of accidents)
- Sensitive feature S : race
- Unobserved variable U : aggressive driving
  - causes drivers to be more likely have an accident
  - causes individuals to prefer red cars (in X)
- Individuals belonging to a certain race A are more likely to drive red cars
- These individuals are no more likely to be aggressive or to get in accidents than anyone else

Omitting S may introduce unfairness

Total Causal Effect

- $TCE = \mathbb{P}\{Y_{S \leftarrow s'}\} - \mathbb{P}\{Y_{S \leftarrow s}\}$

Total Predictions Effect

- $TPE = \mathbb{P}\{h(X_{S \leftarrow s'})\} - \mathbb{P}\{h(X_{S \leftarrow s})\}$

1 Group fairness
- Overview
- Problem statement
- Measuring fairness
- Ensuring group fairness

2 Individual fairness
- Overview
- Problem statement
- Measuring fairness

3 Conclusion

4 TME

Conclusion

- Mathematically modeling fairness
- Group fairness vs. individual fairness
- Quantify different fairness objectives
- Bias mitigation algorithms on output predictions / latent representation

1 Group fairness
- Overview
- Problem statement
- Measuring fairness
- Ensuring group fairness

2 Individual fairness
- Overview
- Problem statement
- Measuring fairness

3 Conclusion

4 TME

Datasets

- The **US Census demographic** data set is an extraction of the 2015 American Community Survey 5-year estimates. It contains 37 information features about 74,000 American census tracts. Predict the percentage of children below the poverty line. Consider gender as a sensitive attribute encoded as the percentage of the women in the census tract.
- The **Crime data** set is obtained from the UCI Machine Learning Repository (Dua and Graff, 2017). This data set includes a total of 128 attributes for 1,994 instances from communities in the US. Predict the number of violent crimes per population for US communities. Use the race information with the ratio of an ethnic group per population as sensitive attribute.
- The **COMPAS data** set (Angwin et al., 2016) contains 13 attributes of about 7,000 convicted criminals with class labels that state whether or not the individual reoffended within 2 years of their most recent crime. Use age as sensitive attribute.

Work to do

- Choose one data set
- Learn a prediction model
- Mesure the fairness of the model using one appropriate metric