Fairness and Discrimination with Multiple Attributes

Agathe Fernandes Machado, Suzie Grondin, François Hu, Philipp Ratz & Arthur Charpentier

Insurance Data Science 2024, Stockholm
# Motivation

<table>
<thead>
<tr>
<th></th>
<th>CA</th>
<th>HI</th>
<th>GA</th>
<th>NC</th>
<th>NY</th>
<th>MA</th>
<th>PA</th>
<th>FL</th>
<th>TX</th>
<th>AL</th>
<th>ON</th>
<th>NB</th>
<th>NL</th>
<th>QC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Driving experience</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Credit history</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Occupation</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Employment status</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Housing situation</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Address/ZIP code</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
<tr>
<td><strong>Insurance history</strong></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
<td><em>x</em></td>
</tr>
</tbody>
</table>


pip install equipy
Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- $\mathbf{X} \in \mathcal{X}$: ‘non-sensitive’ features,
- $\mathbf{A} = (A_1, \cdots A_r) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_r$: $r$ sensitive features,
- $\hat{Y}$: prediction of response variable (continuous or score for a binary classifier)
- $m$: predictive model for $(\mathbf{x}, \mathbf{a})$, and $m^*$ the optimal Bayes estimator,
- $\nu_f$: distribution of $m(\mathbf{X}, \mathbf{A})$
  with cumulative distribution function $F_m$ and quantile function $Q_m$,
- $\nu_f|_{a_i}$: conditional distribution of $m(\mathbf{X}, \mathbf{A})|_{A_i = a_i}$ with $F_m|_{a_i}$ and $Q_m|_{a_i}$,
- $\mathcal{R}(m) = \mathbb{E}[(Y - m(\mathbf{X}, \mathbf{A}))^2]$: (theoretical ) quadratic risk.

Given a sample $\{(y_i, \mathbf{x}_i, \mathbf{a}_i)\}$, the empirical risk is $\hat{\mathcal{R}}_n(m) = \sum_{i=1}^n (y_i - m(\mathbf{x}_i, \mathbf{a}_i))^2$

Optimal Bayes estimator $m^* = \operatorname*{argmin}_{m} \{\mathcal{R}(m)\}$, is

$$m^*(\mathbf{x}, \mathbf{a}) = \mathbb{E}[Y|_{(\mathbf{X} = \mathbf{x}), (\mathbf{A} = \mathbf{a})}]$$
Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- Theoretical quadratic risk is $\mathcal{R}(m) = \mathbb{E}[(Y - m(X, A))^2]$
- Define Wasserstein (2) distance $\mathcal{W}_2^2(\nu_0, \nu_1) = \min_{Z_0 \sim \nu_0, Z_1 \sim \nu_1} \{\mathbb{E}[(Z_1 - Z_0)^2]\}$

if $Z_0$ and $Z_1$ are univariate (+ technical conditions),

$$\mathcal{W}_2^2(\nu_0, \nu_1) = \int_0^1 (Q_1(u) - Q_0(u))^2 du = \mathbb{E}[(Q_1 \circ F_0(Z_0) - Z_0)^2], \ Z_0 \sim \nu_0$$

$z \mapsto Q_1 \circ F_0(z)$ is the “optimal Monge” transport, Santambrogio (2015)
Notations, risk $\mathcal{R}$ and unfairness $\mathcal{U}$

- $m$ is strongly fair regarding a single sensitive attribute (SSA) $A_i$, according to demographic parity – $m(X, A) \perp \perp A_i$ – if and only if:

$$\mathcal{U}_i(m) = \max_{a_i \in A_i} \text{distance}(\nu_m, \nu_m|_{a_i}) = 0 \text{ unfairness w.r.t. } A_i$$

- $m$ is strongly fair regarding multiple sensitive attribute (MSA), if and only if:

$$\mathcal{U}(m) = \mathcal{U}_1(m) + \cdots + \mathcal{U}_r(m) = 0 \text{ unfairness w.r.t. } A$$

- $m$ is $\epsilon$-approximately fair regarding $A_i$ if and only if $\mathcal{U}_i(m) \leq \epsilon \cdot \mathcal{U}_i(m^*)$ (where $m^*$ is the optimal Bayes estimator)
The Price to Pay for Fairness

- Let $\mathcal{M}$ denote a (general) class of models,
- Let $\mathcal{M}_{F,i}$ denote the subset of fair models with respect to $A_i$,
  \[ \mathcal{M}_{F,i} = \{ m \in \mathcal{M} : U_i(m) = 0 \} \]
- Fairness is achieved by projection onto the fair subspace
  \[ \hat{m} \in \arg\min_{m \in \mathcal{M}} \{ \hat{R}_n(m) \} \text{ and } \hat{m}_{F,i} \in \arg\min_{m \in \mathcal{M}_{F,i}} \{ \hat{R}_n(m) \} \]
- The price of fairness of a model class $\mathcal{M}$
  \[ \mathcal{E}_{F,i}(\mathcal{M}) = \min_{m \in \mathcal{M}_{F,i}} \{ R(m) \} - \min_{m \in \mathcal{M}} \{ R(m) \} \geq 0. \]
- Chzhen et al. (2020) proved that (+mild technical assumptions)
  \[ \mathcal{E}_{F,i}(\mathcal{M}) = \min_{\mu \in \mathcal{M}} \left\{ \sum_{a_i \in A_i} \mathbb{P}(A_i = a_i) \cdot W_2^2(\nu_{m^*|a_i}, \nu_{\mu|a_i}) \right\} \]
The Price to Pay for Fairness

Given $K$ distributions $(\nu_1, \ldots, \nu_K)$, and weights $(w_1, \ldots, w_K) \in \mathbb{R}_+^K$, the $\mathcal{W}_2$-Barycenter is the minimizer:

$$\text{Bar}\{(w_k, \nu_k)_{k=1}^K\} = \arg\min_{\nu} \left\{ \sum_{k=1}^K w_k \cdot \mathcal{W}_2^2(\nu_k, \nu) \right\}.$$ 

SSA ($r = 1$) Chzhen et al. (2020)

$\nu_{m_{B_i}}$ corresponding to $\min_m \sum_{a_i \in A_i} \mathbb{P}(A_i = a_i) \cdot \mathcal{W}_2^2(\nu_{m*|a_i}, \nu_m)$

$$m_{B_i}(x, a_i) = \left( \sum_{a'{\in}A_i} \mathbb{P}(A_i = a') \cdot Q_{m*|a'} \right) \circ F_{m*|a_i}(m^*(x, a_i)), \ \forall (x, a_i) \in \mathcal{X} \times A_i.$$ 

→ EquiPy: implemented in the function `FairWasserstein` of the fairness module.
The Case of Multiple Attributes

- Consider a machine Learning model $m$, score predictions and two sensitive attributes, ethnic origin $A_1$ (White/Black) and gender $A_2$ (Male/Female).
- Consider densities of $\nu_{m|A_1=0}$, $\nu_{m|A_1=1}$ (left) and $\nu_{m|A_2=0}$, $\nu_{m|A_2=1}$ (right)
The Case of Multiple Attributes

- Consider a machine Learning model $m$, score predictions and two sensitive attributes, ethnic origin $A_1$ (White/Black) and gender $A_2$ (Male/Female).
- Consider densities of $\nu_{m|A_1=0}$, $\nu_{m|A_1=1}$ (left) and $\nu_{m|A_2=0}$, $\nu_{m|A_2=1}$ (right)
- Plot densities of barycenters, $\nu_{m_{B_1}}$ and $\nu_{m_{B_2}}$
The Case of Multiple Attributes

- **Intersectional Fairness**, MSA $\rightarrow$ Single sensitive attribute (SSA), by intersection, ethnic origin $A_1$, gender $A_2$

  $\mathbf{a} \in \mathcal{A} = A_1 \times A_2 = \{\text{white, black}\} \times \{\text{male, female}\}$

  Here $\mathcal{A}$ corresponds to $4 = 2 \times 2$ states,

  $\mathcal{A} = \{(\text{white, male}), (\text{white, female}), (\text{black, male}), (\text{black, female})\}$

- **Sequential Fairness**, MSA, in Hu et al. (2024)
The Case of Multiple Attributes

- MSA \((r \geq 1)\) Hu et al. (2024)

\[
m_B(x, a) := m_{B_1} \circ m_{B_2} \circ \cdots \circ m_{B_r}(x, a)
\]

where

\[
m_{B_i} \circ m_{B_j}(x, a) = \left( \sum_{a' \in A_i} \mathbb{P}(A_i = a') Q_{m_{B_j}|a_i'} \right) \circ F_{m_{B_j}|a_i}(m_{B_j}(x, a))
\]

∀ \((x, a) \in X \times A_{1:r}\), with the \(i\)-th component of \(a\) denoted \(a_i\).

- Hu et al. (2024) proved the associativity of Wasserstein barycenters (fairness mitigation remains unaffected by the order of \(A_{1:r}\)).

→ EquiPy: implemented in the function MultiWasserstein of fairness module.
The Case of Multiple Attributes

- Given $\nu_{m_{B_1}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on $A_1$ (no impact, already fair)
  - the barycenter $\nu_{m_{B_2}}$ conditional on $A_2$

- On the right, distribution of $\nu_{m_{B_2} \circ m_{B_1}}$
The Case of Multiple Attributes

- Given $\nu_{m_{B_2}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on $A_1$
  - the barycenter $\nu_{m_{B_2}}$ conditional on $A_2$ (no impact, already fair)

- On the left, distribution of $\nu_{m_{B_1} \circ m_{B_2}}$
The Case of Multiple Attributes

- The order of this sequential approach leads different interpretations,
  - left hand part, $A_2$ then $A_1$
  - right hand part, $A_1$ then $A_2$
Life insurance dataset

- Public SEER database: [https://seer.cancer.gov](https://seer.cancer.gov),
- Prediction of one-year mortality of US individuals with melanoma skin cancer, → Use the methodology presented in Sauce et al. (2023), we convert the dataset into survival data, by accounting for exposure over a given time interval.
- Sample size $n = 547,878$ from 2004 to 2018,
- Explanatory variables: 16 features describing patient characteristics (age, gender male/female, ethnic origin) and cancer attributes (tumor size, extent).
  → MSA framework: use of the function MultiWasserstein.
Model fitting

1. Split the data into train and test sets,
2. Fit Logistic Regression \( m \) (\( m \) can be any ML model, model-agnosticity)
3. Apply \( m \) on the test set to obtain \( \hat{y}_{\text{test}} \).

We consider different model fitting scenarios, in which we include or exclude sensitive attributes as explanatory variables:

<table>
<thead>
<tr>
<th>Ethnic origin</th>
<th>Gender</th>
<th>AUC</th>
<th>Unfairness</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>0.8652</td>
<td>0.2179</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>0.8672</td>
<td>0.2668</td>
</tr>
</tbody>
</table>
Transforming predictions

1. Split the test data into calibration and test sets,
2. Specify an order to sequentially correct: $A_1$ corresponds to ethnic origin and $A_2$ corresponds to gender,
3. Fit and transform your test predictions using MultiWasserstein from fairness module.
Visualizations

Unfairness and metric calculations with graphs module:

- `fair_waterfall_plot`: sequential gain in fairness for the specified order $A_1$ then $A_2$,
- `fair_multiple_arrow_plot`: fairness-performance relationship for all potential pathways.

![Sequential fairness: Origin then Gender](image1)

![Sequential fairness operation](image2)
When correcting biases related to gender, we reduce fairness regarding origin:

<table>
<thead>
<tr>
<th>Fairness step</th>
<th>Unfairness in origin</th>
<th>Unfairness in gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base model</td>
<td>0.2371</td>
<td>0.0297</td>
</tr>
<tr>
<td>Origin</td>
<td>0.0345</td>
<td>0.0309</td>
</tr>
<tr>
<td>Origin &amp; Gender</td>
<td>0.0469</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

We can prioritize fairness across attributes by specifying $\epsilon = [0, 0.5]$ corresponding to exact fairness in $A_1$ and 0.5-approximate fairness in $A_2$.

$$m_B = 0.5 \cdot (m_{B_2} \circ m_{B_1}) + 0.5 \cdot m_{B_1}$$


