“Insurance is the contribution of the many to the misfortune of the few,” Bénéplanc et al. (2022). at the core of insurance business lies discrimination between risky and non-risky insureds, Avraham (2017), to ask for an actuarily fair premium.

By looking into earlier debates on discrimination, we show that some algorithmic biases are a renewed version of older ones, while others show a reversal of the previous order. Paradoxically, while the insurance practice has not deeply changed nor are most of these biases new, the machine learning era still deeply shakes the conception of insurance fairness.

Fairness of Machine Learning in Insurance
New Rags for an Old Man?

Arthur Charpentier1,2, Laurence Barry3
1Université du Québec à Montréal (charpentier.arthur@uqam.ca) 2Chaire PARIX

INTRODUCTION

Classically, $y$ is the variable of interest, e.g. accident occurrence (0, 1), frequency ($) or annual cost, before time insurance recovery ($R_k$, etc), $x$ some covariates and $p$ is a protected (sensitive) attribute (in $0, 1$).

Dwork et al. (2012), Hardt et al. (2016), Corbett-Davies et al. (2017). Berk et al. (2017) introduced various concepts of fairness

• Demographic parity, $P(Y = 1|P = p) = c_{st}$, $\forall p$
• Equalized odds, $P(Y = 1|P = p, Y = y) = c_{st}$, $\forall p, y$
• Equalized opportunity, $P(Y = 1|P = p, Y = 1) = c_{st}$, $\forall p$
• Predictive equality, $P(Y = 1|P = p, Y = 0) = c_{st}$, $\forall p$
• Conditional accuracy equality $P(Y = y|P = p, Y = y) = c_{st}$, $\forall p, y$

Those notions are simply related to independence $\hat{Y} \perp P$, or conditional independence $\hat{Y} \perp P \mid Y$ or $Y \perp P \mid \hat{Y}$. Let $s$ denote the score function, $s(x,p) = P(Y = 1|x, p)$ so that $\hat{Y} = \mathbb{1}(s > \text{threshold})$.

Inspired by Goodfellow et al. (2018), to avoid un-fairness, penalize according to $HGR(y, p)$, e.g.

$$\min_{G, f} \mathbb{E}_{(x, y, p)} \left[ L(h(x), y) + \lambda \text{HGR}(y, p) \right]$$

where $HGR(U, V)$ is max $\{ \text{corr}(f(U), g(V)) \} = \max_{g(U) \in \mathcal{S}_U, g(V) \in \mathcal{S}_V} \mathbb{E}[f(U)g(V)]$ and where $\mathcal{S}_U = \{ : U \rightarrow \mathbb{R} : \mathbb{E}[U|Y = 0] = 0 \} \text{ and } \mathbb{E}[Y|U] = 1$ and similarly $\mathcal{S}_V$. But a conditional version, can also be considered

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

REFERENCES


To go further

The question we address is

• How to identify the source of the discrimination ?
• How to get a ratemaking formula that is “discrimination free” ?

with technical challenges, continuous protected attribute, Grari et al. (2022), non-observed protected attribute, Grati et al. (2021), Racicot et al. (2021) and connections with causal-related fairness, and prevention.

INDIVIDUAL FAIRNESS

• Lipschitz property, $D((\hat{y}_i, \hat{y}_j)) \leq d(x_i, x_j)$, $Y, \hat{y}$.
• Counterfactual fairness,

$$P \{ Y_i^{r_{counter}} = y \mid X = x \} = P \{ Y^{r_{counter}} = y \mid X = x \}, \forall y, \forall y.$$

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.

Black et al. (2020), Torous et al. (2021) and de Lara et al. (2021) suggested

• Transport-based fairness, $P_{(X^\theta \mid m, T^\ast)} = P_{(X^\theta \mid m, T^\ast)}$

\[
\begin{align*}
\{X^\theta_{\ast}(m, T^\ast) = \{x \in X' : m(x,0) > m(T^\ast)(x,1)\} \\
\{X^\theta_{\ast}(m, T^\ast) = \{x \in X' : m(x,0) < m(T^\ast)(x,1)\},
\end{align*}
\]

where $P_{(\theta)} = P(\theta = 0)$, $P_{(\theta)} = P(\theta = 1)$, and

$$T^\ast = \min_{T : T_{\theta \rightarrow \theta'} = \int_{R^d} \{x - T(x)\}^2 dP(x).$$