Fairness in Machine Learning algorithms

Jean-David Fermanian (Crest)
Telecom-ACPR

11 January 2021

Introductory example: scoring

Goal: to allocate (or not) a loan that is requested by a customer ... with the help of a scoring/ML predictive model.

- $Y = 1$ (resp. $Y = 0$) when default (resp. no default);
- a vector of individual features X;
- a sensitive feature S, as gender, race, color, religion, disability, etc.
- output of the model: $\hat{p}(x) \simeq \mathbb{P}(Y = 1|X = x)$, the default probability given $X = x$. When $\hat{p}(x) > c$, the loan is refused.

  More generally: a predictor $\hat{Y} \in \{0, 1\}$, a function of X.

Introductory example: scoring

To simplify, $S \in \{0, 1\}$.

Calibration of a “fine and smart” statistical/ML model:

1. by cutting-edge statistical techniques,
2. nice fit in sample on a large dataset,
3. good performances out-of-sample.

“Unfortunately”, we observe $\hat{p}(1, z) > \hat{p}(0, z)$ for most z values. Or even $\mathbb{E}_Z[\hat{p}(1, Z)] > \mathbb{E}_Z[\hat{p}(0, Z)]$ only.

Ex.: All other things being equal, being a black man increases the probability of being rejected.

Under an ethical (not statistical!) point-of-view, this may be not satisfying.

Q1.: Can we formalize the problem?

Q2.: Is it possible to correct it?

A predicted value $\hat{Y} = g(X) = g(S, Z)$.

Y and $\hat{Y}$ may be discrete in $\{0, 1, \ldots, p\}$, or continuous.

(a) *Demographic parity* : If $\hat{Y}$ is discrete,

$$\mathbb{P}(\hat{Y} = j | S = k) = \mathbb{P}(\hat{Y} = j),\ j = 0, \ldots, p - 1; k = 0, 1.$$

When $p = 2$, this is equivalent to

$$\mathbb{P}(\hat{Y} = 1 | S = 0) = \mathbb{P}(\hat{Y} = 1 | S = 1).$$

In the case of a continuous predicted variable $\hat{Y}$,

$$\mathbb{P}(\hat{Y} \leq y | S = 0) = \mathbb{P}(\hat{Y} \leq y | S = 1) = \mathbb{P}(\hat{Y} \leq y), \ \forall y.$$

This implies (but is not equivalent to)

$$\mathbb{E}(\hat{Y} | S = 0) = \mathbb{E}(\hat{Y} | S = 1) = \mathbb{E}(\hat{Y}).$$

Equalized Odds (EO)

Equalized odds means

$$\mathbb{P}(\hat{Y} = j | S = k, Y = l) = \mathbb{P}(\hat{Y} = j | Y = l), \; j, l = 0, \ldots, p-1; k = 0, 1.$$

When $p = 2$ (binary Y and $\hat{Y}$), this means

$$\mathbb{E}[\hat{Y} | S = 0, Y = l] = \mathbb{E}[\hat{Y} | S = 1, Y = l] = \mathbb{E}[\hat{Y} | Y = l], \; l = 0, 1.$$

Ex.: $S =$ the gender and $Y =$ the recidivism variable. EO means the predicted probability of recidivism of a person is the same given this person is a male or a female *and* he/she reoffends.

Equalized Odds (EO)

For continuous explained variables Y and $\hat{Y}$, EO means

$$\begin{aligned}\mathbb{P}(\hat{Y} \leq y | S = 0, Y = y') &= \mathbb{P}(\hat{Y} \leq y | S = 1, Y = y') \\ = \quad \mathbb{P}(\hat{Y} \leq y | Y = y'), \ \forall (y, y') \in \mathbb{R}^2.\end{aligned}$$

This implies (but is not equivalent to)

$$\mathbb{E}[\hat{Y} | S = 0, Y = y'] = \mathbb{E}[\hat{Y} | S = 1, Y = y'] = \mathbb{E}[\hat{Y} | Y = y'], \ \forall y' \in \mathbb{R}.$$

In the case of binary Y, we often think of the outcome $Y = 0$ as the "advantaged" outcome: "not defaulting on a loan", "admission to a college", "receiving a promotion"...

Relaxation of EO: non-discrimination only within the "advantaged" outcome group.

$\Rightarrow$ *equal opportunity*, Hardt et al. (2016).

When Y and $\hat{Y}$ are binary and if we privilege $Y = 0$, this means

$$\mathbb{E}[\hat{Y}|S = 0, Y = 0] = \mathbb{E}[\hat{Y}|Y = 0].$$

Lack of Disparate Mistreatment (LDM)

(c) In the case of discrete outcomes, the *lack of disparate mistreatment* (Zafar et al., 2017) is defined as

$$\mathbb{P}(\hat{Y} \neq Y | S = 0) = \mathbb{P}(\hat{Y} \neq Y | S = 1).$$

For any type of outcome, the latter definition of LDM may be

$$\mathbb{E}[|Y - \hat{Y}|^{\alpha} \mid S = 0] = \mathbb{E}[|Y - \hat{Y}|^{\alpha} \mid S = 1],$$

for some constant $\alpha > 0$, or even (stronger)

$$\mathbb{P}(Y - \hat{Y} \leq y \mid S = 0) = \mathbb{P}(Y - \hat{Y} \leq y \mid S = 1), \; y \in \mathbb{R}.$$

+ other concepts: approximate fairness, fairness with probability $1 - \varepsilon$, etc.

Ways of “correcting” fairness biases

Temptation: remove S and re-calibrate the model. This does not work in general !

Challenge: improve the fairness of a ML algorithm ...without damaging its predictive power too much!

Many attempts in the literature.

Three families of proposed solutions.

Ways of "correcting" fairness biases

(1) pre-processing: modify the training data so that the outcome of (potentially) any machine learning algorithm applied to that data will be fair.

Ex.: change labels and/or attributes, remove or weight observations, etc.

Pros: a definitive and discreet solution.

Cons: "data is the past truth". Potential unexpected future problems!

Ref.: Kamiran (2009), Dwork et al. (2012), Kamiran and Calders (2012), Feldman et al. (2015), Calmon et al. (2017)

Ways of “correcting” fairness biases

(2) *algorithm modification techniques*: modify an existing algorithm or create a new one that will be fair under any inputs.

Typically, add constraints during the calibration stage (regularization).

Pros: theoretically attractive

Cons: complicate existing models, potential numerical problems

Ref.: Calders and Verwer (2010), Kamishima et al. (2012), Zemel et al. (2013), Zafar et al. (2017), Friedler et al (2018)

Ways of “correcting” fairness biases

(3) *post-processing*: take the outputs of some ML models and conveniently modify their predictions to be fair.

Ex.: modification of decision thresholds, randomization.

Pros: use existing ML algorithms.

Cons: not omnibus (depend on the initial classifers/predictors)

Ref.: Zliobaite (2015), Hardt et al. (2016), Woodworth et al. (2017), Agarwal et al. (2019), Chzhen et al. (2020)