Addressing algorithmic fairness through metrics and explanations

IDAI 2021 Summer School

Course 5 (10:00–11:00)

Catuscia Palamidessi

July 22, 2021

Many notions of fairness

Group fairness vs individual fairness

Group fairness

Addresses the discrimination against a particular group of people.
For instance:

- Ethnicity (black vs white people)
- Sex (women vs men)
- Religion
- ...

Individual fairness

Addresses the discrimination at a personal level against a particular individual.

Causes of bias

Possible causes of bias include:

- Bias in the data source: past decisions may already have been unfair
- Unbalanced Data
- Representation Bias
- Measurement Bias
- Limited Features Bias
- Algorithmic Bias
- ...

It is important to identify the potential sources of bias, as the suitable definition of fairness and the countermeasures depend on them.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54 (6), 1-35, 2021.

Group fairness

Notation and basic notions

Data model and predictor

X	Legitimate attribute
A	Sensitive attribute (binary)
Y	Decision (binary)
$\widehat{Y}$	Prediction of the classifier (binary)

Usually we think of $\widehat{Y}, Y = 1$ as the positive decision (granting a loan, promotion, admission to college,...) and $\widehat{Y}, Y = 0$ as the negative one.

Example

Loan	X	employment, salary (income)
	A	ethnicity
	Y	loan decision
	$\widehat{Y}$	prediction

Some group fairness notions

Statistical Parity SP

$$\mathbb{P}[\widehat{Y} = 1 \mid A = 0] \quad = \quad \mathbb{P}[\widehat{Y} = 1 \mid A = 1] \qquad\qquad \widehat{Y} \perp A$$

Statistical parity is usually too strong:

Example

In the example of the loan, if the income status is unbalanced between the ethnic groups, in order to satisfy SP the predictor should grant loans also to some of the people with insufficient income.

Some group fairness notions

Conditional Statistical Parity CSP

$$\mathbb{P}[\widehat{Y} = 1 \mid X = x, A = 0] = \mathbb{P}[\widehat{Y} = 1 \mid X = x, A = 1] \qquad \widehat{Y} \perp A \mid X$$

In contrast to SP, CSP allows disparity between the groups, as long as this disparity is justified by the legitimate attributes.

There may be correlation between the decision and the group, but only via the legitimate attributes $\Rightarrow$ A, X and $\widehat{Y}$ form a Markov chain $A - X - \widehat{Y}$

Example

The predictor can grant loans less frequently to the disadvantaged ethnic group, as long as this disparity is justified by the lack of sufficient income (to pay back the loan).

Some group fairness notions

Previous notions usually have a negative impact on the accuracy. To avoid this problem, Hardt et al. [NIPS'16] proposed the following notion:

Equalized Odds EOdds

$$\mathbb{P}[\widehat{Y} = 1 \mid Y = y, A = 0] \quad = \quad \mathbb{P}[\widehat{Y} = 1 \mid Y = y, A = 1] \qquad \widehat{Y} \perp A \mid Y$$

Example

The probability that the predictor takes the "right" decision does not depend on the ethnic group.

The *true positives* ($y = 1$) and the *false positives* ($y = 0$) must be equally distributed across the two groups. The same holds for the *true negatives* and the *false negatives*, since $\widehat{Y} = 0$ has complementary probability.
Again we have a Markov chain: $A - Y - \widehat{Y}$

EOdds assumes implicitly that Y is unbiased. If the training data do not respect this assumption, we need to correct them.

Moritz Hardt, Eric Price, Nati Srebro: Equality of Opportunity in Supervised Learning. NIPS 2016: 3315-3323

Some group fairness notions

A relaxation of EOdds, called *Equal Opportunities*, requires equal treatment of the two groups only when the true decision should be positive:

Equal Opportunities EOpp

$$\mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 0] \quad = \quad \mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 1]$$

Example

The probability that the predictor grants the loan when it is the "right" decision, does not depend on the ethnic group.

Note that we don't care of what happen when the true decision should be negative. In the example, people from the privileged group who do not have sufficient income may get the loan more frequently than those of the other group in the same financial situation.

Relation between fairness notions

Auxiliary notions

Equal Base Rates

$$\mathbb{P}[Y = 1 \mid A = 0] \quad = \quad \mathbb{P}[Y = 1 \mid A = 1] \qquad Y \perp A$$

Conditional Equal Base Rates

$$\mathbb{P}[Y = 1 \mid X = x, A = 0] \quad = \quad \mathbb{P}[Y = 1 \mid X = x, A = 1] \qquad Y \perp A \mid X$$

Equal base rates and its conditional version correspond to statistical parity and conditional statistical parity, respectively, but they are conditions on the data source, rather than on the ML model.

Classifier's independence from true decision

$$\mathbb{P}[\widehat{Y} = 1 \mid Y = 0] \quad = \quad \mathbb{P}[\widehat{Y} = 1 \mid Y = 1] \qquad \widehat{Y} \perp Y$$

Clearly a classifier whose prediction is not correlated to the true decision has minimal accuracy.

Equalized Odds and Statistical Parity imply Equal Base Rates or Independence

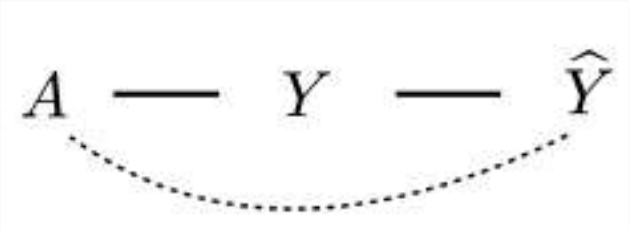
Some fairness notions are incompatible, in the sense that their combination may result in very strong assumptions about the data.

The following result, in Barocas et al. [2019], states that *Statistical Parity* and *Equalized Odds* imply *Equal Base Rates*, or *Independence of the classifier*.

Theorem: SP + EOdds $\Rightarrow$ EBR or Ind

$$\widehat{Y} \perp A \quad + \quad \widehat{Y} \perp A \mid Y \quad \Rightarrow \quad Y \perp A \quad or \quad \widehat{Y} \perp Y$$

Graphically:



S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

Equalized Odds and Statistical Parity imply Equal Base Rates or Independence

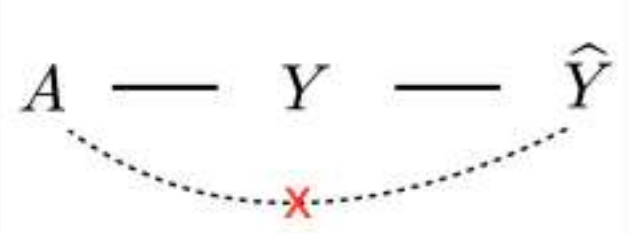
Some fairness notions are incompatible, in the sense that their combination may result in very strong assumptions about the data.

The following result, in Barocas et al. [2019], states that *Statistical Parity* and *Equalized Odds* imply *Equal Base Rates*, or *Independence of the classifier*.

Theorem: SP + EOdds $\Rightarrow$ EBR or Ind

$$\widehat{Y} \perp A \quad + \quad \widehat{Y} \perp A \mid Y \quad \Rightarrow \quad Y \perp A \quad or \quad \widehat{Y} \perp Y$$

Graphically:



S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

Equalized Odds and Statistical Parity imply Equal Base Rates or Independence

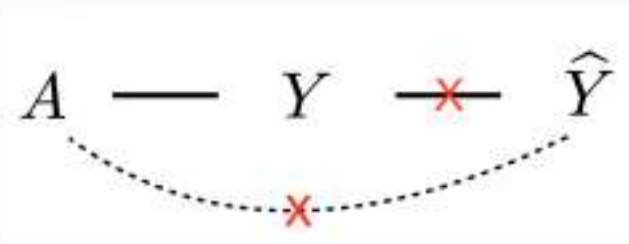
Some fairness notions are incompatible, in the sense that their combination may result in very strong assumptions about the data.

The following result, in Barocas et al. [2019], states that *Statistical Parity* and *Equalized Odds* imply *Equal Base Rates*, or *Independence of the classifier*.

Theorem: SP + EOdds $\Rightarrow$ EBR or Ind

$$\widehat{Y} \perp A \quad + \quad \widehat{Y} \perp A \mid Y \quad \Rightarrow \quad Y \perp A \quad or \quad \widehat{Y} \perp Y$$

Graphically:



S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

Equalized Odds and Statistical Parity imply Equal Base Rates or Independence

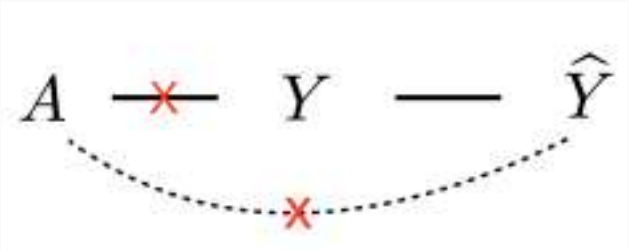
Some fairness notions are incompatible, in the sense that their combination may result in very strong assumptions about the data.

The following result, in Barocas et al. [2019], states that *Statistical Parity* and *Equalized Odds* imply *Equal Base Rates*, or *Independence of the classifier*.

Theorem: SP + EOdds $\Rightarrow$ EBR or Ind

$$\widehat{Y} \perp A \quad + \quad \widehat{Y} \perp A \mid Y \quad \Rightarrow \quad Y \perp A \quad or \quad \widehat{Y} \perp Y$$

Graphically:



S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

Conditional Equal Base Rates and Equalized Odds imply Conditional Statistical Parity

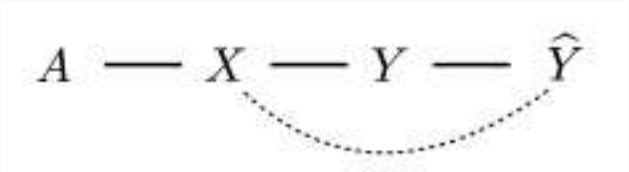
Other notions work well together:

The following result states that *Conditional Equal Base Rates* and *Equalized Odds* imply *Conditional Statistical Parity*.

Theorem: CEBR + EOdds $\Rightarrow$ CSP

$$Y \perp A \mid X \quad + \quad \widehat{Y} \perp A \mid Y \quad \Rightarrow \quad \widehat{Y} \perp A \mid X$$

Graphically:



Relation between fairness, privacy and accuracy

- In the rest of this talk we will focus on Equal Opportunity.
- This notion is very popular because it is claimed to be fully compatible with accuracy and privacy.
- We are going to present some negative results about that limit the extent of this claim.

Accuracy and Bayes classifier

The accuracy of a classifier is defined as the probability that the prediction coincides with the true decision:

Definition: Accuracy

$$Acc(\widehat{Y}) \stackrel{def}{=} \mathbb{E}\, 1_{\widehat{Y}=Y} = \sum_{x,a,y} \mathbb{P}[X = x, A = a, \widehat{Y} = Y = y]$$

The Bayes classifier $\widehat{Y}_B$is defined as the classifier that, for every input x and a, predicts the decision with highest probability. Namely:

Definition: Bayes classifier

$$\mathbb{P}[\widehat{Y}_B = 1 \mid X = x, A = a] = \begin{cases} 1 & \mathbb{P}[Y = 1 \mid X = x, A = a] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Accuracy of the Bayes classifier

The Bayes classifier is optimal, i.e., it has the best accuracy, given by:

$$Acc(\widehat{Y}_B) = \sum_{x,a} \max_y \mathbb{P}[Y = y, X = x, A = a]$$

Trivial Classifier

Definition: Trivial Classifier

A classifier is *trivial* if the output distribution does not depend on the input:

$$\mathbb{P}[\widehat{Y} = 1 \mid X = x, A = a] \;=\; \mathbb{P}[\widehat{Y} = 1 \mid X = x', A = a'] \;\text{ for all } x, x', a, a'$$

Definition: Optimal Trivial Classifier

The optimal trivial classifier $\widehat{Y}_T$ is the trivial classifier that provides max accuracy:

$$\mathbb{P}[\widehat{Y}_T = 1 \mid X = x, A = a] \;=\; \begin{cases} 1 & \mathbb{P}[Y = 1] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Lemma : Accuracy of the optimal trivial classifier

The optimal trivial classifier has the following accuracy:

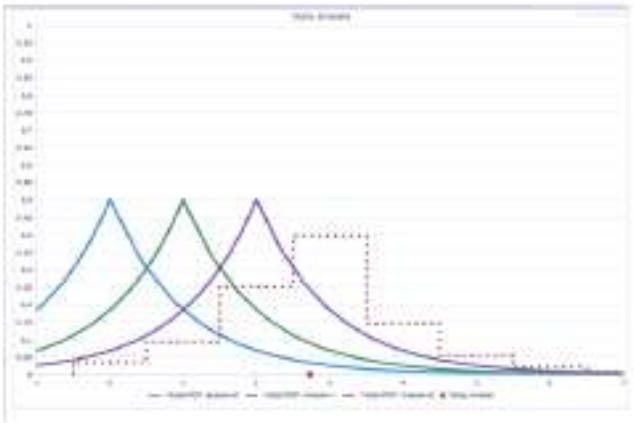
$$Acc(\widehat{Y}) = \max_{y} \mathbb{P}[Y = y]$$

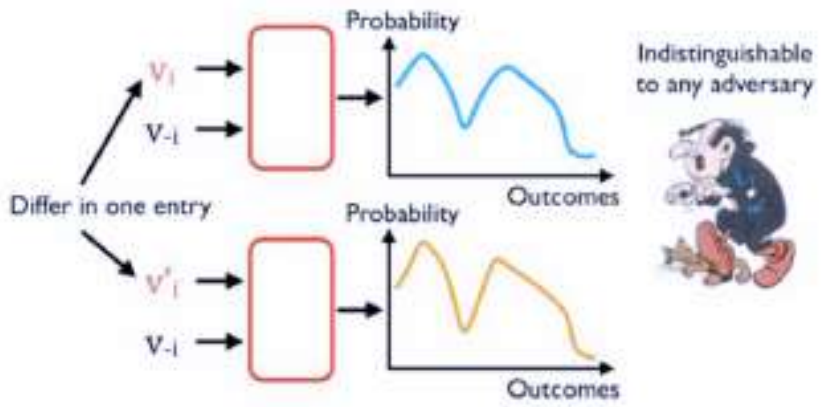
Differential Privacy

Definition: Differential Privacy in ML

A learning algorithm $\mathcal{A}$ is ϵ-differentially private if for every pair of adjacent training sets d and d', and every model $\mathcal{M}$:

$$\mathbb{P}[\mathcal{A}(d) = \mathcal{M}] \quad \leq \quad e^{\epsilon} \cdot \mathbb{P}[\mathcal{A}(d') = \mathcal{M}]$$





Fairness, Accuracy and Privacy: Impossibility results

Cummings et al. [UMAP'19] have proved the following result:

Theorem: incompatibility of EOpp, DP and Accuracy

For any ϵ, there are data models (distributions on the data) for which any classifier that satisfies ϵ-differential privacy and EOpp cannot have greater accuracy than the trivial one.

Successively, Agrawal [IJCAI'21] extended the result of Cummings et al. also to an approximate notion of EOpp (when we consider a bound on the difference of the two probabilities instead than strict equality) and to other notions of fairness (equalized odds and statistical parity).

Rachel Cummings, Varun Gupta, Dhamma Kimpara, Jamie Morgenstern: On the Compatibility of Privacy and Fairness. UMAP (Adjunct Publication) 2019: 309-315

Agarwal S. Trade-Offs between Fairness and Privacy in Machine Learning, in IJCAI 2021 Workshop on AI for Social Good. ; 2021.

A stronger result

Pinzon et al [2021] have proved a even stronger result than the one of Cummings et al. Namely, we have found an impossibility result without the DP constraint:

Theorem: incompatibility of EOpp, and Accuracy

There are data distributions on the data for which any classifier that satisfies EOpp cannot have greater accuracy than the trivial one.

Carlos Pinźon, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy under fairness constraints. CoRR, abs/2107.06944, 2021.

Proof

Notation for relevant distributions

Relevant distributions

$$\begin{aligned} \pi_{xa} &\stackrel{\text{def}}{=} \mathbb{P}[X = x, A = a] \\ q_{xa} &\stackrel{\text{def}}{=} \mathbb{P}[Y = 1 \mid X = x, A = a] \\ \rho_{xa} &\stackrel{\text{def}}{=} \mathbb{P}[\widehat{Y} = 1 \mid X = x, A = a] \end{aligned}$$

- The data model is *deterministic* if $q_{xa} \in \{0, 1\}$ for all x and a.
- The classifier is *deterministic* if $\rho_{xa} \in \{0, 1\}$ for all x and a.
- In general we assume that data models and classifiers are probabilistic, unless otherwise specified.
- Note that the Bayes classifier $\widehat{Y}_B$ and the optimal trivial classifier $\widehat{Y}_T$ are deterministic.

Proof

Consider a partition of the domain $\mathcal{X}$ of legitimate features in two sets $\mathcal{X}_1$ and $\mathcal{X}_2$. Then impose a constraint on the the data model as follows:

$$q_{xa} = \begin{cases} q_1 & x \in \mathcal{X}_1, a = 0 \\ q_2 & x \in \mathcal{X}_2, a = 0 \\ q_3 & \text{otherwise} \end{cases}$$

Where we assume that $q_1 < \frac{1}{2}$ and $q_2, q_3 > \frac{1}{2}$.

We also assume, wlog, that $\mathbb{P}[Y = 1] > \frac{1}{2}$.

Proof

Using this notation, equal opportunity can be rewritten as follows:

$$\text{EOpp} \;\Leftrightarrow\; (q_1\alpha_1 + q_2\alpha_2)q_3r_3 \;=\; q_3\alpha_3(q_1r_1 + q_2r_2)$$

where:

$$\begin{array}{llll} \alpha_1 \overset{\text{def}}{=} \sum_{\mathcal{X}_1} \pi_{x0}\rho_{x0} & & r_1 \overset{\text{def}}{=} \sum_{\mathcal{X}_1} \pi_{x0} \\ \alpha_2 \overset{\text{def}}{=} \sum_{\mathcal{X}_2} \pi_{x0}\rho_{x0} & & r_2 \overset{\text{def}}{=} \sum_{\mathcal{X}_2} \pi_{x0} \\ \alpha_3 \overset{\text{def}}{=} \sum_{\mathcal{X}} \pi_{x1}\rho_{x1} & & r_3 \overset{\text{def}}{=} \sum_{\mathcal{X}} \pi_{x1} \end{array}$$

Note that, by definition:

$$0 \leq \alpha_i \leq r_i \qquad \forall i \in \{1, 2, 3\}$$

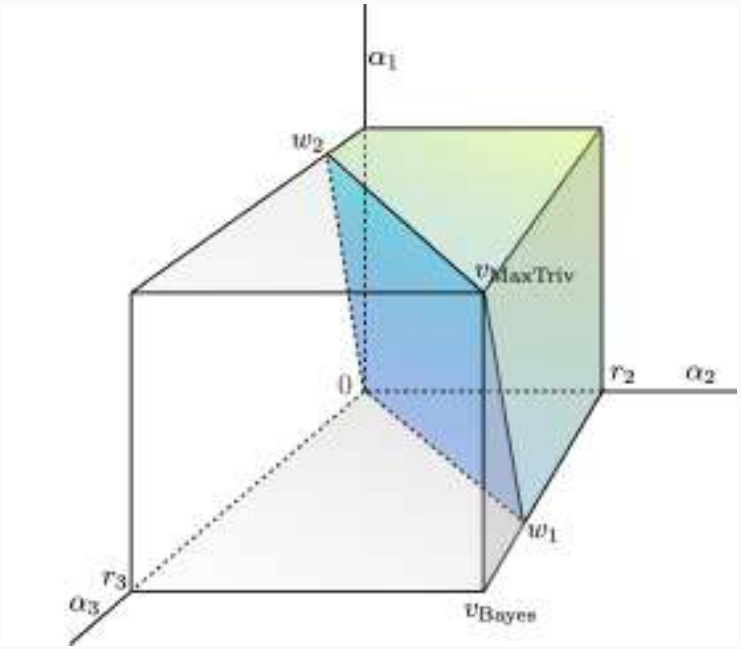
Furthermore:

- $\alpha_i = 0$ iff $\rho_{xa} = 0$ for all x, a in the corresponding partition,
- $\alpha_i = r_i$ iff $\rho_{xa} = 1$ for all x, a in the corresponding partition.

Proof

If we consider α_1, α_2 and α_3 as a system of Cartesian coordinates, we have that each pair (data model, predictor) "lives" in the three-dimensional parallelepiped determined by the vertices $(0, 0, 0)$, $(r_1, 0, 0)$, ..., (r_1, r_2, r_3).

Furthermore, EOpp is a linear constraint on the α_i's, so it determines an plane. The pairs (data model, predictor) that satisfy EOpp "live" in this plane.



It is easy to see (by solving the EOpp constraint) that the plane passes by $(0, 0, 0)$, (r_1, r_2, r_3), and $w_1 = (0, r_2, \frac{p_2 r_2 r_3}{p_1 r_1 + p_2 r_2})$.

Proof

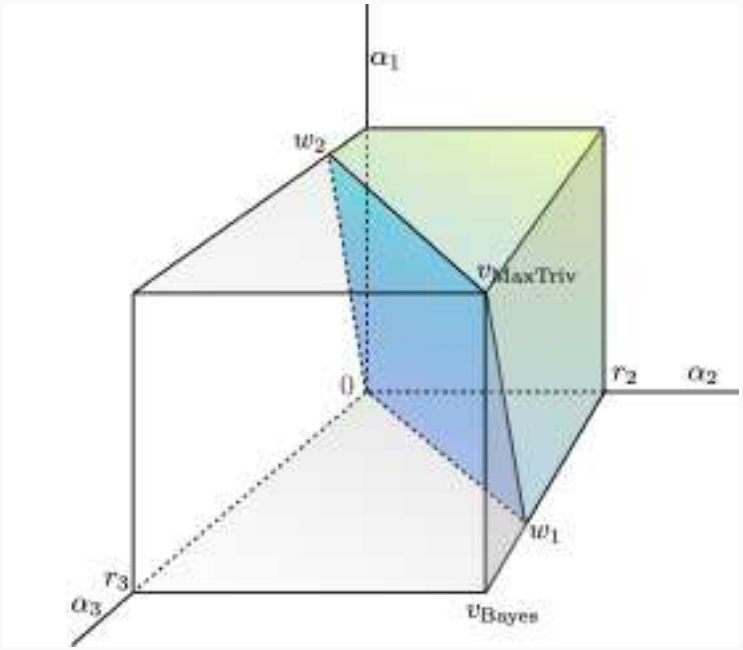
Consider an approximate version of EOpp:

$$EOL = \mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 0] - \mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 1]$$

Then the classifiers with $EOL \geq 0$ "live" in the colored part of the parallelepiped.

Because of the assumptions $q_1 < \frac{1}{2}$, $q_2, q_3 > \frac{1}{2}$ and $\mathbb{P}[Y = 1] > \frac{1}{2}$, we have that:

- the Bayes classifier $\widehat{Y}_B$ corresponds to the point $v_{\mathrm{Bayes}} = (0, r_2, r_3)$,
- the optimal trivial classifier $\widehat{Y}_T$ corresponds to the point $v_{\mathrm{MaxTriv}} = (r_1, r_2, r_3)$.



Proof

To conclude the proof, we have to show that v_{MaxTriv} is the point in the plane with highest accuracy.

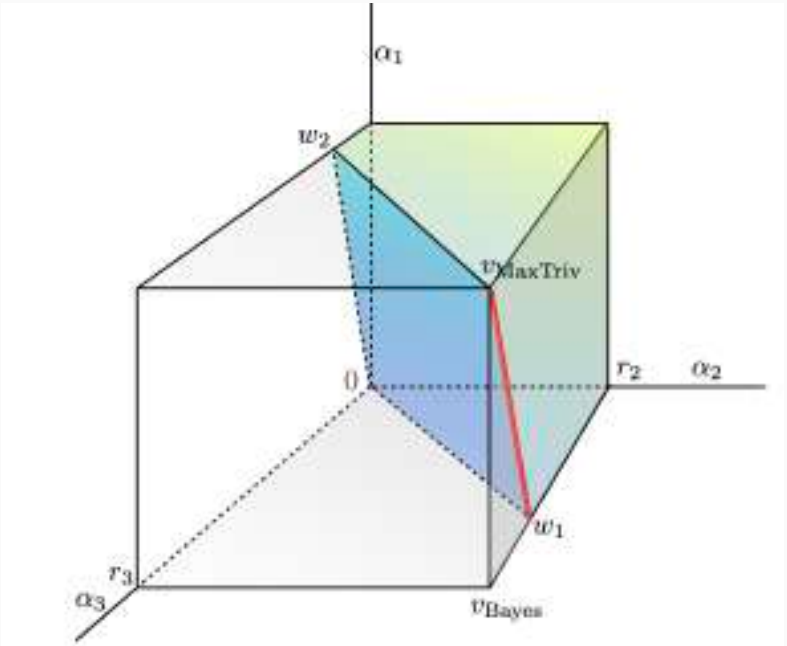
To this purpose, let us rewrite the accuracy using our notation. We have:

$$Acc(\widehat{Y}) = \sum_{i=1,2,3} (2q_i - 1)\alpha_i + (1 - q_i)r_i$$

Given the conditions on the q_i's, we have that $Acc(\widehat{Y})$ is monotonic increasing with α_2. Hence we have only to consider the classifiers on the red line.

However, $Acc(\widehat{Y})$ is monotonic increasing with α_3 but monotonically decreasing with α_1. To obtain $Acc(v_{\mathrm{MaxTriv}}) > Acc(w_1)$ we need to impose a final constraint, which is:

$$(2q_3 - 1)r_3 < (2q_1 - 1)r_1 + 2q_3r_3$$



Approximate notions of fairness

Equal Opportunity Loss

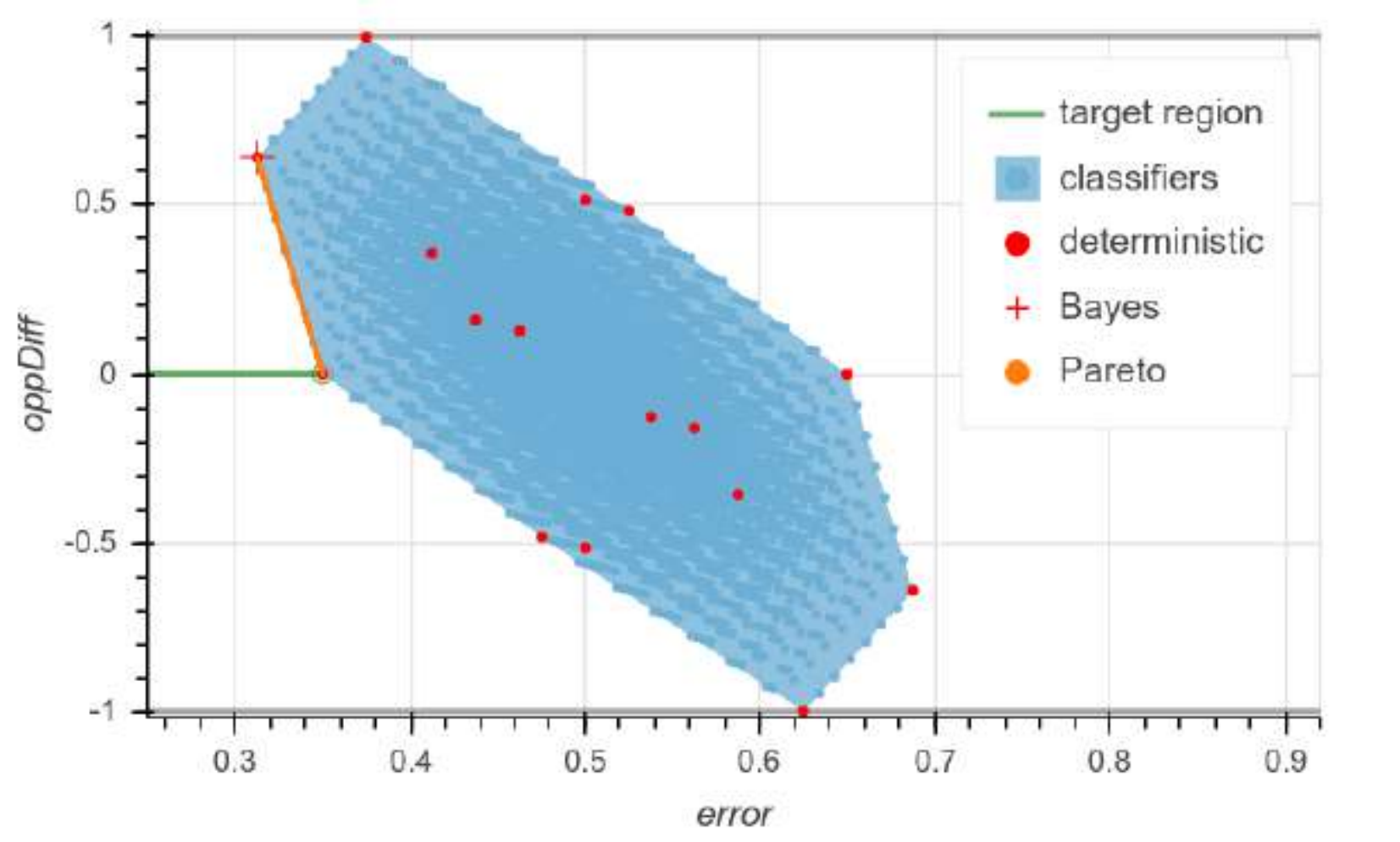
The notions of fairness introduced so far are rather strict because they impose an equality between the probabilities relative to the two groups. We can relax the various notions of fairness by imposing a bound on the difference between the probabilities.

Here we give the definition for Equal Opportunity. Similar definitions can be given for all other notions.

Equal Opportunity Loss

$$EOL \quad = \quad | \, \mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 0] \quad - \quad \mathbb{P}[\widehat{Y} = 1 \mid Y = 1, A = 1] \, | < \alpha$$

Pareto optimality for the approximate notion of EOpp



Note: oppDiff = EOL, error = 1 - Accuracy

Possible research directions

- Characterize data distributions for which Equal Opportunity can be a good notion of fairness (i.e., it offers a good trade-off with accuracy).

- Relation with privacy: Characterize data distributions for which "removing the sensitive feature" during the training phase improves fairness without affecting accuracy significantly.

Some references

- S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.
- Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi: On the Applicability of ML Fairness Notions. CoRR abs/2006.16745 (2020)
- Carlos Pinźon, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy under fairness constraints. CoRR, abs/2107.06944, 2021

Thanks !