Fairness of decision algorithm in machine learning

Alessandro La Farciola, Leevi Rönty, Camilla Beneventano

EPFL

14 December 2022

Fairness in machine learning

- Machine learning increasingly affects decision in domains protected by anti-discrimination law: ensure fairness in machine learning!

Fairness in machine learning

- Machine learning increasingly affects decision in domains protected by anti-discrimination law: ensure fairness in machine learning!
- Naïve approach: ignoring all protected attributes, such as race, color, sex, ...

Fairness in machine learning

- Machine learning increasingly affects decision in domains protected by anti-discrimination law: ensure fairness in machine learning!
- Naïve approach: ignoring all protected attributes, such as race, color, sex, ...

Definition (Demograpghic parity)

We say that a binary predictor $\hat{Y} \in \{0, 1\}$ *satisfies* **demographic parity** *with respect to a binary procted attribute* $A \in \{0, 1\}$ *and* Y *if*

$$Pr\{\hat{Y} = 1 | A = 0\} = Pr\{\hat{Y} = 1 | A = 1\}.$$

Equalized odds and equal opportunity

Definition (Equalized odds)

We say that a predictor $\hat{Y}$ satisfies **equalized odds** *with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y, i.e. if*

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = Pr\{\hat{Y} = 1 | A = 1, Y = y\}, \quad y \in \{0, 1\}.$$

Equalized odds and equal opportunity

Definition (Equalized odds)

We say that a predictor $\hat{Y}$ *satisfies* **equalized odds** *with respect to protected attribute* A *and outcome* Y*, if* $\hat{Y}$ *and* A *are independent conditional on* Y*, i.e. if*

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = Pr\{\hat{Y} = 1 | A = 1, Y = y\}, \quad y \in \{0, 1\}.$$

Definition (Equal opportunity)

We say that a binary predictor $\hat{Y}$ *satisfies* **equal opportunity** *with respect to* A *and* Y *if*

$$Pr\{\hat{Y} = 1 | A = 0, Y = 1\} = Pr\{\hat{Y} = 1 | A = 1, Y = 1\}.$$

Derived predictor

- Consider $R \in [0,1]$ a **score function** and the predictor $\hat{Y} = \mathbb{1}\{R > t\}$;
- Find $\tilde{Y} = \mathbb{1}\{R > t_A\}$ using different thresholds for different values of A;

Definition (Derived predictor)

*A predictor $\tilde{Y}$ is **derived** from a random variable R and the protected attribute A if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\tilde{Y}$ is independent of X conditional on (R, A).*

It is always possible to construct a trivial predictor satisfying equalized odds or equal opportunity, but our goal is to derive predictors $\tilde{Y}$ that minimize the expected loss $\mathbb{E}[\ell(\tilde{Y}, Y)]$, where $\ell \colon \{0,1\}^2 \to \mathbb{R}$ is a loss function.

Deriving from a score function

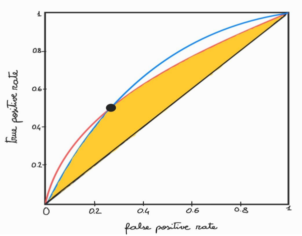
Consider the A-conditional ROC curves

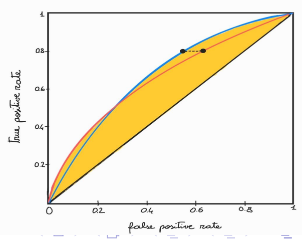
$$C_a(t) := (Pr\{\hat{R} > t | A = a, Y = 0\}, Pr\{\hat{R} > t | A = a, Y = 1\});$$

and the convex hull of the image of the conditional ROC curve

$$D_a = convhull\{C_a(t) \ : \ t \in [0,1]\}.$$

- for equal odds we must have that for all classes the resulting true positive rate and false positive rate must be in $D_0 \cap D_1$
- equal opportunity means that $C_0(t_0)$ and $C_1(t_1)$ agree in the second component.





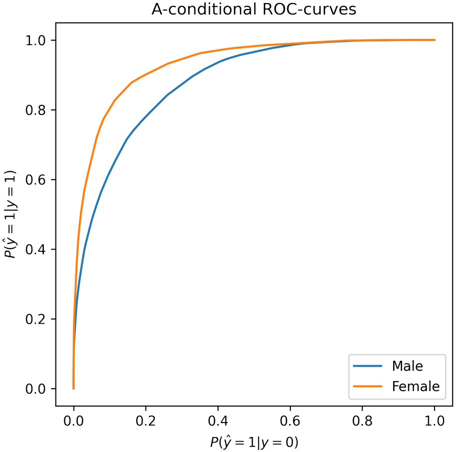
Experiments: dataset

	Age	Workclass	Education	Education.num	Occupation	Relationship	Race	Sex	Capital.gain	Capital.loss	Hours.per.week	label
1	39	State-gov	Bachelors	13	Adm-clerical	Not-in-family	White	Male	2174	0	40	<=50K
2	50	Self-emp-not-inc	Bachelors	13	Exec-managerial	Husband	White	Male	0	0	13	<=50K
3	38	Private	HS-grad	9	Handlers-cleaners	Not-in-family	White	Male	0	0	40	<=50K
4	53	Private	11th	7	Handlers-cleaners	Husband	Black	Male	0	0	40	<=50K
5	28	Private	Bachelors	13	Prof-specialty	Wife	Black	Female	0	0	40	<=50K
6	37	Private	Masters	14	Exec-managerial	Wife	White	Female	0	0	40	<=50K
7	49	Private	9th	5	Other-service	Not-in-family	Black	Female	0	0	16	<=50K
8	52	Self-emp-not-inc	HS-grad	9	Exec-managerial	Husband	White	Male	0	0	45	>50K
9	31	Private	Masters	14	Prof-specialty	Not-in-family	White	Female	14084	0	50	>50K
10	42	Private	Bachelors	13	Exec-managerial	Husband	White	Male	5178	0	40	>50K
11	37	Private	Some-college	10	Exec-managerial	Husband	Black	Male	0	0	80	>50K
12	30	State-gov	Bachelors	13	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	>50K
13	23	Private	Bachelors	13	Adm-clerical	Own-child	White	Female	0	0	30	<=50K
14	32	Private	Assoc-acdm	12	Sales	Not-in-family	Black	Male	0	0	50	<=50K
15	40	Private	Assoc-voc	11	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	>50K
16	34	Private	7th-8th	4	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	<=50K
17	25	Self-emp-not-inc	HS-grad	9	Farming-fishing	Own-child	White	Male	0	0	35	<=50K
18	32	Private	HS-grad	9	Machine-op-inspct	Unmarried	White	Male	0	0	40	<=50K
19	38	Private	11th	7	Sales	Husband	White	Male	0	0	50	<=50K
20	43	Self-emp-not-inc	Masters	14	Exec-managerial	Unmarried	White	Female	0	0	45	>50K

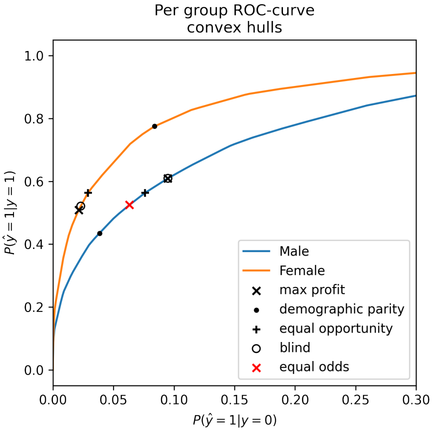
Experiments: predictors

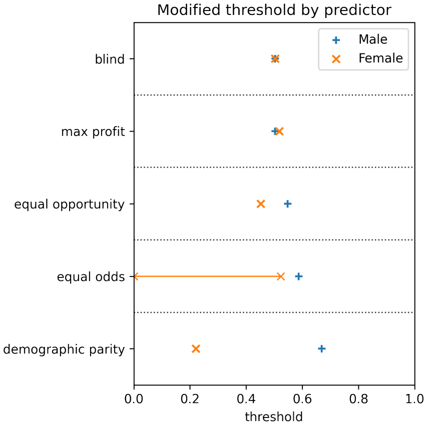
- **Sex Blind**: same threshold for all classes
- **Max profit**: no fairness constraints, best threshold for each class
- **Demographic parity**: max profit subject to positive prediction rate per class being equal
- **Equal opportunity**: max profit subject to true positive rate per class being equal
- **Equalized odds**: same fraction of true positives and false positives for each group, different thresholds (possibly randomized)

Experiments: results

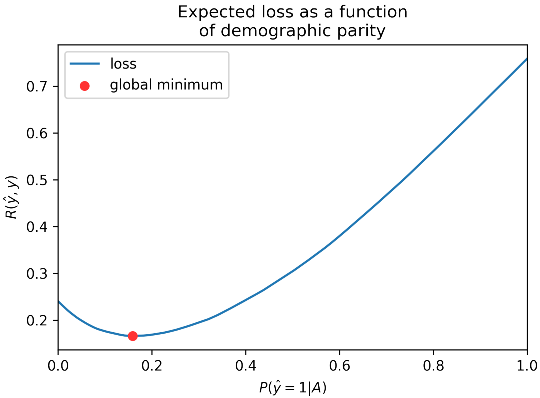


Experiments: results





How to choose demographic parity



Loss is a convex function of positive prediction rate $\Rightarrow$ we can choose optimal prediction rate with ternary search.

Experiments: errors

predictor	train loss	test loss
blind	0.1482	0.1487
max profit	0.1481	0.1488
demographic parity	0.1664	0.1695
equal opportunity	0.1490	0.1491
equal odds	0.1624	0.1692

References:

- Hardt, M., Price, E., Srebro N., (2016). *Equality of opportunity in supervised learning*. In Advances in Neural Information Processing Systems (pp. 3315-3323).
- Adult. (1996). UCI Machine Learning Repository.

Thank you for the attention!