Advanced Machine Learning

Fairness in Machine Learning

Learning goals

- Know why fairness aspects are relevant for Machine Learning
- Get familiar with prevalent fairness criteria
- Know their disadvantages and their relationship

RESEARCH ON FAIRNESS

- The question of what fairness actually is goes back thousands of years to antiquity. Even back then, philosophers such as Aristotle asked themselves this question.
- The academic research on fairness started with the pioneering works in educational testing (Clearly, 1968) and economics (Becker 1957, Phelps 1972, Arrow 1973).
- In computer science, the research essentially started in the early 2000s and has recently attracted a lot of interest, which is of course due to the increasing use of machine learning models for automated decision making systems.

FAIRNESS-AWARE BINARY CLASSIFICATION: FORMAL SETTING

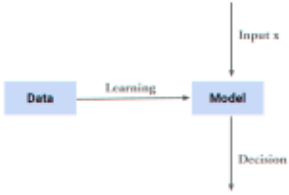
We are provided with a data set $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)} \right) \right) \in (\mathcal{X} \times \mathcal{Y})^n$, where

- $\mathcal{X}$ is the input/feature/attribute space with $p = \dim(\mathcal{X})$,
- $\mathcal{Y}$ the output / target / label space (for now $\mathcal{Y} = \{-1, 1\}$),
- the tuple $\left(\mathbf{x}^{(i)}, y^{(i)} \right) \in \mathcal{X} \times \mathcal{Y}$ is the i-th observation,
- $\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(n)} \right)^T$ the j-th feature vector.

So we have observed n objects, described by p features.

- We assume the observed data $\mathcal{D}$ to be generated by a process that can be characterized by some probability distribution $\mathbb{P}_{xy}$, defined on $\mathcal{X} \times \mathcal{Y}$.
- In particular, $\left(\left(\mathbf{x}^{(1)}, y^{(1)} \right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)} \right) \right)$ is i.i.d. with $\left(\mathbf{x}^{(i)}, y^{(i)} \right) \sim \mathbb{P}_{xy}$.
- We denote the random variables (vectors) following this distribution by lowercase $\mathbf{x}$ and y.

The ultimate goal for a machine learning model f is then loosely speaking "to predict y from $\mathbf{x}$", which leads to a decision $f(\mathbf{x}) = \hat{y} \in \{-1, 1\}$. Note that $\hat{y}$ is a random variable (can be constant), as it is essentially a function of the random input $\mathbf{x}$.



SENSITIVE ATTRIBUTES/FEATURES

- The aspect of fairness usually arises due to the presence of sensitive attributes/features among the attributes/features $\mathbf{x}_1, \ldots, \mathbf{x}_p$, e.g. age, gender, nationality, race, ...
- Note that we assume that the attribute/feature observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ are random observations of the random vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$ with distribution $\mathbb{P}_x$. Accordingly, the j-th attribute/feature vector $\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(n)}\right)^\top$ is a collection of random observations of the random variable x_j with distribution $\mathbb{P}_{x_j}$, which is the marginal distribution of $\mathbb{P}_x$ for the j-th attribute/feature.
- We introduce the random variable **A** to capture all sensitive attributes/features, which typically has discrete values.
- The basic idea of fairness criteria introduced for machine learning methods is to equalize different decision criteria or statistical quantities involving **A**.

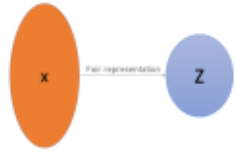
  This goes back to Anne Clearly in the 1960s who studied group differences in educational testing.

DOWNSIDES OF INDEPENDENCE AS A FAIRNESS CRITERION

- Independence does not take into account that the outcome y might be correlated with **A**, which means that the different realizations of **A** have different underlying distributions for y.
- Not considering this dependency can lead to decisions which are fair through the lens of the independence criterion, but not for the groups themselves.
- Moreover, independence does not rule out the possibility of unfair practices. For example, consider a job hiring process involving different groups of people. Assume that we
  - make thoughtful and good decisions in one specific group with accepting people from that group with a rate $p \in (0, 1)$,
  - make poor and bad decisions in all other groups with the same acceptance rate $p \in (0, 1)$, respectively.

ACHIEVING INDEPENDENCE VIA REPRESENTATION LEARNING

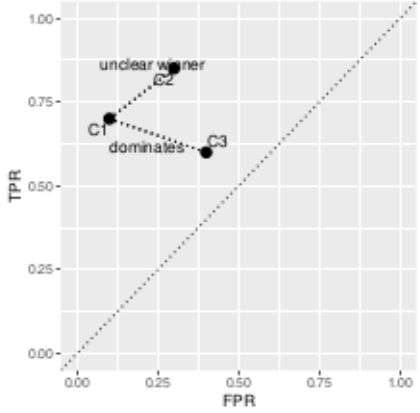
- One common idea to satisfy the independence criterion is by finding a "fair representation" $\mathbf{Z}$ of the data $\mathbf{x}$, i.e., one such that $\mathbf{Z} \perp\!\!\!\perp \mathbf{A}$ holds. Then, the ML method f uses $\mathbf{Z}$ instead of $\mathbf{x}$ for the decision: $\hat{y} = f(\mathbf{Z})$



- The idea goes back to Zernel et al. (2013), where three requirements on the representation are formulated:
  1. Information about $\mathbf{x}$ should be preserved $\Leftrightarrow$ Mutual information between $\mathbf{x}$ and $\mathbf{Z}$ is high.
  2. The sensitive attributes/features $\mathbf{A}$ are obfuscated $\Leftrightarrow$ Mutual information between $\mathbf{A}$ and $\mathbf{Z}$ is low.
  3. Accuracy of the model f using $\mathbf{Z}$ is (still) high $\Leftrightarrow$ Mutual information between y and $\mathbf{Z}$ is high.

INTERLUDE: ROC SPACE

- For comparing classifiers, we characterize them by their TPR and FPR values and plot them in a coordinate system.
- We could also use two different ROC metrics (decision criteria) which define a trade-off, for instance, TPR and PPV.



		True Class y	
		$+$	$-$
Pred. $\hat{y}$	$+$	TP	FP
	$-$	FN	TN

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

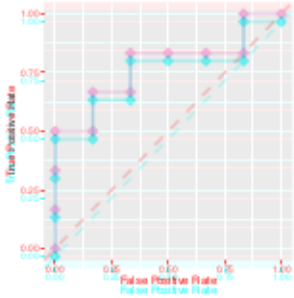
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

INTERLUDE: ROC CURVES FOR SCORING CLASSIFIERS / 2

To draw a ROC curve:
(W.l.o.g. $s : \mathcal{X} \to [0, 1]$)

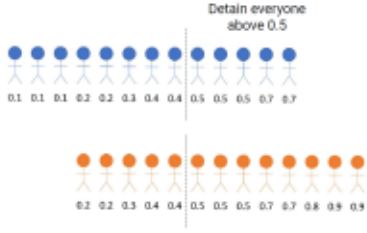
1. Rank test observations on decreasing score.
2. Start with $c = 1$, so we start in $(0, 0)$; we predict everything as negative.
3. Iterate through all possible thresholds c and proceed for each observation x as follows:
   - If x is positive, move TPR $1/n_+$ up, as we have one TP more.
   - If x is negative, move FPR $1/n_-$ right, as we have one FP more.

(n_+ : number of positives, n_- : number of negatives.)



DOWNSIDES OF SEPARATION AS A FAIRNESS CRITERION

- Consider two groups of people: blue and orange. We are interested to decide whether we should detain (positive class) a person and use a scoring classifier with scores in $[0, 1]$ and a threshold $c = 0.5$.

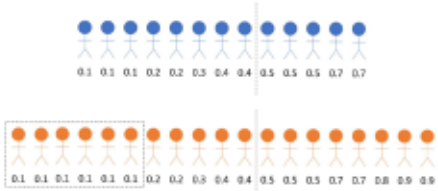


Detention rate	False pos. rate
38%	25%
61%	42%

- The classifier is not satisfying separation as FPR and FNR are not the same among the two groups.

DOWNSIDES OF SEPARATION AS A FAIRNESS CRITERION / 2

- In order to achieve separation we would need to arrest more low risk individuals in the orange group.



Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

- Thus, as with achieving independence, separation can also lead to undesirable outcomes.

SUFFICIENCY AND CALIBRATION / 2

$$\mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \mathbf{a}) = \mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \tilde{\mathbf{a}}) \qquad \text{(sufficiency)}$$

vs.

$$\mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \mathbf{a}) = s \qquad \text{(calibration on individual level)}$$

- If a probabilistic classifier is *calibrated on the group level*, then it also satisfies sufficiency.
- If a probabilistic classifier f satisfies sufficiency, then we can find a function $C : [0, 1] \to [0, 1]$ such that f based on $C(\mathbf{S})$ (instead of $\mathbf{S}$) is calibrated on the group level.
- Sufficiency is only slightly weaker, but it is fair to say that both properties are essentially equivalent.

DOWNSIDES OF SUFFICIENCY AS A FAIRNESS CRITERION / 2

If we calibrate the classifier, we have no detentions any more!