Debasing algorithms

Marija Slavkovik
University of Bergen
@MSlavkovik
marija.slavkovik@uib.no

Fairness

- Fairness ≠ lacks of statistical bias
- There is no unique definition for fairness
- There is no unique source of bias

Fairness definitions

- **Group fairness**: do outcomes systematically differ between demographic groups?
- **Individual fairness**: like individuals should be treated alike
- **Process fairness**: how fair is it to use a given (predictive) feature
- **Fairness of use**: when the ML predictor works better for one group rather than other and as a consequence erodes position or opportunity to the unfavored group
- More: Arvind Narayanan tutorial FAT2018

  https://www.youtube.com/watch?v=jlXluYdnyyk

Why is algorithmic fairness difficult

- What is order for the spider is chaos for the fly - different things are fair to different stakeholders
- Bias can be implicit in society and as such embedded in the data
- Both people and algorithms can be biased, but people can change

De-biasing algorithms

- Algorithms that discover and mitigate algorithmic bias are called **de-biasing algorithms**
- More:

  Trusted AI and AI Fairness 360 Tutorial by Prasanna Sattigeri, September 18, 2019

  - **Material:** https://aif360.mybluemix.net/
  - https://www.youtube.com/watch?v=lXbG2u4lOYI&feature=youtu.be

How to evaluate fairness?

How to de-bias ML?

Source: https://fairmlbook.org/classification.html

Confusion matrix

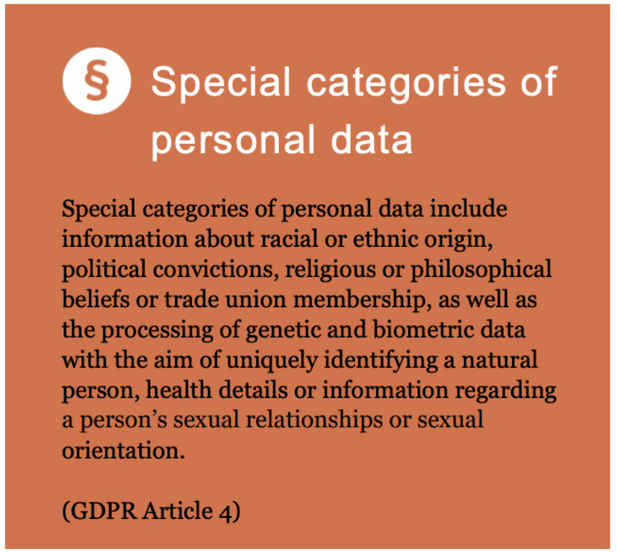
	is spam	is not spam
classified as spam	number of true positive	number of false positive
classified as not spam	number of false negative	number of true negative

How often was the classifier correct?

Sensitive features

GROUP FAIRNESS

- When the instance (features) describe a person, the features X contain or implicitly encode sensitive characteristics of that person.
- Let the letter A designate a discrete random variable that captures one or multiple sensitive characteristics
- Removing or ignoring sensitive attributes does not ensure the impartiality of the resulting classifier.
- Many fairness criteria have been proposed over the years, each aiming to formalize different desiderata.
- We consider the formal definitions of three representative fairness criteria that relate to many of the proposals that have been made.



Formal non-discrimination criteria

- Most of the proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the classifier or score R.

- Most of the formal nondiscrimination criteria fall into one of three different categories defined along the lines of different (conditional) independence:

Non-discrimination criteria		
Independence	**Separation**	**Sufficiency**
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

$R = \mathbb{E}[Y \mid X]$

sensitive feature

correct "label"

expectation of the target variable Y conditional on the features X we have observed

Independence

aka Demographic Parity, Statistical Parity, Group Fairness

Non-discrimination criteria		
Independence	**Separation**	**Sufficiency**
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Independence

aka Demographic Parity, Statistical Parity, Group Fairness

Non-discrimination criteria		
Independence	**Separation**	**Sufficiency**
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Independence requires the sensitive characteristic to be statistically independent of the score.

- In the case of binary classification, independence simplifies to the condition

$$\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\},$$

  for all groups a, b.

- Thinking of the event $R = 1$ "acceptance", the condition requires the acceptance rate to be the same in all groups.

$$\mathbb{P}\{R = 1 \mid A = a\} \geq \mathbb{P}\{R = 1 \mid A = b\} - \epsilon\,.$$

$$\frac{\mathbb{P}\{R = 1 \mid A = a\}}{\mathbb{P}\{R = 1 \mid A = b\}} \geq 1 - \epsilon\,.$$ ← 0.2

relates to the 80 percent rule in disparate impact law.

Limitations of independence

- Imagine a company that in group a hires diligently selected applicants at some rate $p > 0$.
- In group b, the company hires carelessly selected applicants at the same rate p.
- Even though the acceptance rates in both groups are identical, it is far more likely that unqualified applicants are selected in one group than in the other.
- It will appear in hindsight that members of group b performed worse than members of group a, thus establishing a negative track record for group b.

- If there is a lot of training data on group a, compared to group b, then the predictions will be better.

16: This problem was identified and called *self-fulfilling prophecy* in, C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness Through Awareness," in *Proc. 3rd ITCS*, 2012, 214–26. One might object that enforcing demographic parity in this scenario might still create valuable additional training data which could then improve predictions in the future after re-training the classifier on these additional data points.

Separation

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- The separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable.

- In the case where R is a binary classifier, separation is equivalent to requiring for all groups a, b the two constraints

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}\,.$$

- Recall that $\mathbb{P}\{R = 1 \mid Y = 1\}$ is called the **true positive rate** of the classifier. It is the rate at which the classifier correctly recognizes positive instances.

- The false positive rate $\mathbb{P}\{R = 1 \mid Y = 0\}$ highlights the rate at which the classifier mistakenly assigns positive outcomes to negative instances.

Sufficiency

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Sufficiency formalizes that the score already subsumes the sensitive characteristic for the purpose of predicting the target
- One says that R satisfies sufficiency when the sensitive attribute A and target variable Y are clear from the context.

- In the binary case where $Y \in \{0, 1\}$, a random variable R is sufficient for A if and only if for all groups a, b and all values r in the support of R, we have

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}$$

- When R has only two values we recognize this condition as requiring a parity of positive/negative predictive values across all groups.

Relations

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Proposition. Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.

Proposition. Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.

Proposition. Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.

Relations

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

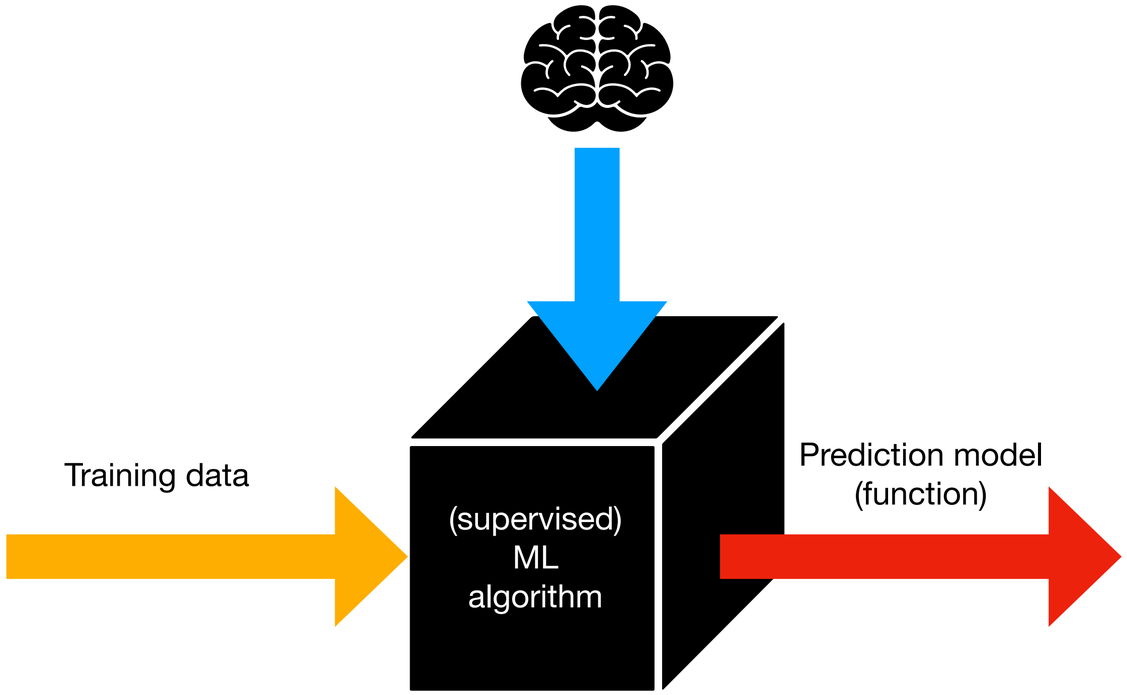
Proposition. Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.

Proposition. Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.

Proposition. Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.

= you have to choose which "fairness" you will insist upon

Bias mitigation



- **Pre-processing**: Adjust the feature space to be uncorrelated with the sensitive attribute.
- **At training time(in-processing)**: Work the constraint into the optimisation process that constructs a classifier from training data.
- **Post-processing**: Adjust the prediction model so as to be uncorrelated with the sensitive attribute.

Preprocessing

- family of techniques to transform a feature space into a representation that as a whole is independent of the sensitive attribute.
- model agnostic
- because the information content cannot be increased by doing local operations on it, as a consequence of the transformations, any classifier trained on the transformed data will satisfy independence

Techniques 1/2

- **Suppression**. Find the attributes that correlate most with the sensitive attribute A. To reduce the bias of the class labels and the attribute A, we remove A and these most correlated attributes.
- **Massaging the dataset.** Change the labels of some objects in the dataset in order to remove the bias from the input data. A good selection of which labels to change is essential.
- **Reweighing.** Instead of changing the labels, the tuples in the training dataset are assigned weights. By carefully choosing the weights, the training dataset can be made bias-free w.r.t. A without having to change any of the labels. The weights on the tuples can be used directly in any method based on frequency counts.
- **Sampling.** For those methods that cannot directly work with weights, the related sampling method can be used instead. We calculate sample sizes for the 4 combinations of A- and Class-values that would make the dataset bias-free. Then, we apply stratified sampling on the four groups; two of the groups will be under-sampled and two over-sampled.

Techniques 2/2

- **Optimized preprocessing (Calmon et al., 2017)** learns a probabilistic transformation that edits the features and labels in the data set with group fairness, individual distortion, and data fidelity constraints and objectives.
- **Learning fair representations (Zemel et al., 2013)** finds a latent representation that encodes the data well but obfuscates information about protected attributes. They maximise the mutual information between the latent representation and the features X while minimising the mutual information between A and the latent representation.
- **Disparate impact remover (Feldman et al., 2015)** edits feature values to increase group fairness while preserving rank-ordering within groups.
- Source: https://arxiv.org/pdf/1810.01943.pdf

Latent representation is a representation that only captures the most relevant characteristics of the input.

How does pre-processing impact explainability?

In-processing

- **Adversarial debiasing** (Zhang et al., 2018) learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.
- **Prejudice remover** (Kamishima et al., 2012) adds a discrimination-aware regularization term to the learning objective.
- Source: https://arxiv.org/pdf/1810.01943.pdf

Post-processing

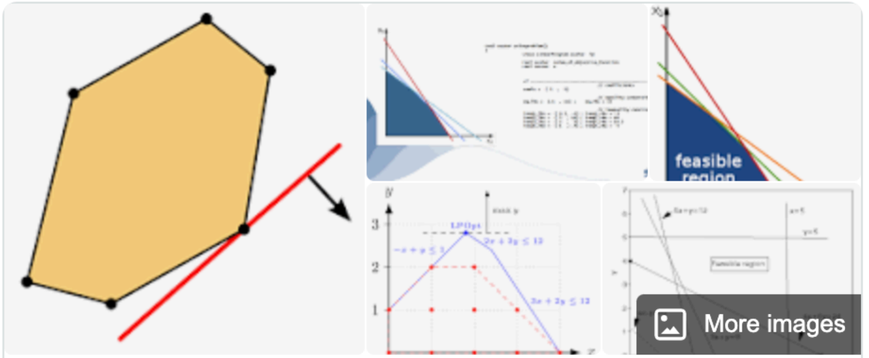
- Not applied to every prediction model (model specific)
- Cannot correct for every fairness criteria
- Impacts accuracy of predictions (trade-offs are necessary)

Post-processing

- **Equalized odds postprocessing** (Hardt et al., 2016) solves a linear program to find probabilities with which to change output labels to optimize equalized odds.

Definition 2.1 (Equalized odds). We say that a predictor $\widehat{Y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\widehat{Y}$ and A are independent conditional on Y.

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$



Linear programming

Linear programming is a method to achieve the best outcome in a mathematical model whose requirements are represented by linear relationships. Linear programming is a special case of mathematical programming. Wikipedia

Post-processing

- **Reject option classification** (Kamiran et al., 2012) gives favourable outcomes to unprivileged groups and unfavourable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty