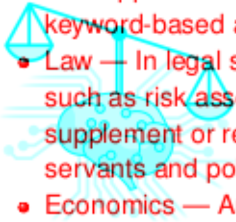
ETHICAL ASPECTS IN MACHINE LEARNING

- Machine learning methods are more and more applied in real-life application, especially for automated decision making:
  - Credit scoring and insurance applications — Should the credit/insurance be granted to a certain person or not?
  - Rating job applications — Machine learning models can help filter applications much more effectively than simple keyword-based approaches.
  - Law — In legal systems around the world, algorithmic tools such as risk assessment instruments (RAI), are being used to supplement or replace the human judgment of judges, civil servants and police officers in many contexts.
  - Economics — Automated trading systems buy and sell orders and automatically transmit the orders to market centers or exchanges.
  - . . .

Advanced Machine Learning

Fairness in Machine Learning



Learning goals

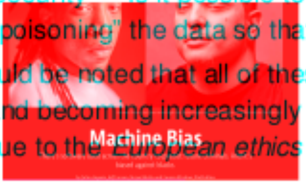
- Know why fairness aspects are relevant for Machine Learning
- Get familar with prevalent fairness criteria
- Know their disadvantages and their relationship

ETHICAL ASPECTS IN MACHINE LEARNING

- These are critical applications involving humans, which raises various ethical issues in different dimensions:
  - Accountability – Can we make sure that the system is functioning as intended?
  - Explainability/Transparency: Is it evident or explainable why one specific decision was made rather than another?
  - Fairness – Does the system disadvantage specific individuals or groups?
  - Privacy – Is the information (data) on the basis of which the system was developed secure against external access?
  - Security – Is it possible to attack the system, e.g. by "poisoning" the data so that undesirable effects occur?
- It should be noted that all of these aspects are intertwined in some way and becoming increasingly important from a legal perspective, e.g. due to the *European ethics guidelines for trustworthy AI.*

FAIRNESS IN MACHINE LEARNING: WHY BOTHER?

- In the recent past, there have been a number of automated decision making tools that have attracted attention for discriminatory behavior:
  - Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. But there is strong evidence that it is discriminating black defendants.
  - Amazon created a tool to trawl the web and spot potential candidates, rating them from one to five stars. But the algorithm learned to systematically downgrade women's CV's for technical jobs such as software developer.



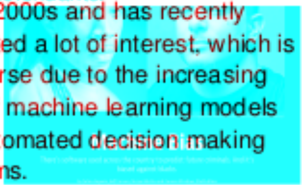
- These are critical applications involving humans, which raises various ethical issues in different dimensions:
  - Accountability — Can we make sure that the system is functioning as intended?
  - Explainability/Transparency: Is it evident or explainable why a specific decision was made rather than another?
  - Fairness — Does the system disadvantage specific individuals or groups?
  - Privacy — Is the information (data) on the basis of which the system was developed secure against external access?
  - Security — Is it possible to attack the system, e.g. by "poisoning" the data so that undesirable effects occur?
- It should be noted that all of these aspects are intertwined in some way and becoming increasingly important from a legal perspective, e.g. due to the *European ethics guidelines for trustworthy AI.*

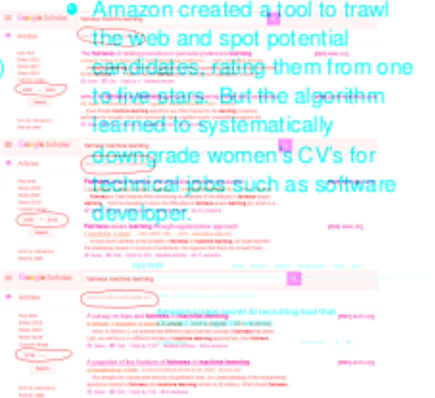
RESEARCH ON FAIRNESS

- 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.

[illegible]ARNING: WHY BOTHER?

- [illegible] ision [illegible] tracted attention for discriminatory behavior:
  - Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a tool used by U.S. courts to assess the likelihood of a defendant becoming a recidivist. But there is strong evidence that it is discriminating black defendants.
  - Amazon created a tool to trawl the web and spot potential candidates, rating them from one to five stars. But the algorithm learned to systematically downgrade women's CV's for technical jobs such as software developer.





FAIRNESS IN MACHINE LEARNING: ROUGH OVERVIEW

- The goal of fairness in Machine Learning is, roughly speaking, to identify and mitigate or even prevent biases of any kind in the decision making based on ML methods along all aspects of the pipeline.
- There are essentially two sources of bias, namely the available data and the ML model itself:
  - Data can be imbalanced or impoverished, e.g. more data on white recidivism outcomes than for blacks. The data can be biased, e.g. collected by a racist or chauvinistic hiring manager. Finally, inconsistencies in the data such as wrong labels or simply noise can lead to bias as well.
  - The prediction of the ML method can be imbalanced w.r.t. to the error. Moreover, the ML method might mimic the biases in the data and even compound injustices.

FAIRNESS-AWARE BINARY CLASSIFICATION

- We will consider only the binary classification/prediction setting in this course, due to the following reasons:
  1. The majority of fairness-critical applications in real life are in fact binary classification/prediction tasks, e.g. credit application (granting vs. not granting a loan), job application (hire applicant or not), trading systems (buy or sell an order), . . .
  2. Quantifying fairness based on a binary outcome variable is mathematically more convenient, while the multi-class variant would introduce additional terms in the fairness quantities. Moreover, multiclass problems can (and are) often addressed by reducing them into multiple binary classification tasks, e.g. one-vs-rest, one-vs-one or error-correcting-codes approaches.
  3. The principles for fairness-aware regression tasks are often just modifications of the ones for binary classification.

[illegible] OVERVIEW

- The goal of fairness in Machine Learning is, roughly speaking, to identify and mitigate or even prevent biases of any kind in the decision making based on ML methods along all aspects of the pipeline.
- There are essentially two sources of bias, namely the available data and the ML model itself:
  - Data can be imbalanced or imperfect, e.g. more data on white recidivism outcomes than for blacks. The data can be biased, e.g. collected by a racist or chauvinistic hiring manager. Finally, inconsistencies in the data such as wrong labels or simply noise can lead to bias as well.
  - The prediction of the ML method can be imbalanced w.r.t. to the error. Moreover, the ML method might mimic the biases in the data and even compound injustices.

FAIRNESS-AWARE BINARY CLASSIFICATION: FORMAL SETTING

- We will consider only the binary classification/prediction setting in this course, due to the following reasons:
  1. The majority of fairness-critical applications in real life are in fact binary classification/prediction tasks, e.g. credit application (granting vs. not granting a loan), job application (hire applicant or not), trading systems (buy or sell an order), ...
  2. Quantifying fairness based on a binary outcome variable is mathematically more convenient, while the multi-class variant would introduce additional terms in the fairness quantities. Moreover, multi-class problems can (and are) often addressed by reducing them into multiple binary classification tasks, e.g. one-vs-rest, one-vs-one or error-correcting-codes approaches.
  3. The principles for fairness-aware regression tasks are often just modifications of the ones for binary classification.

We are provided with a data set $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right) \in (\mathcal{X} \times \mathcal{Y})^n$, 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,
- $\mathbf{x}_j = \left(x_j^{(1)}, \dots, x_j^{(n)}\right)^T$ 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, $\left(\mathbf{x}^{(1)}, y^{(1)}\right), \dots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)$ are 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 $\mathbf{x}$.

DECISION THEORY 101

- In binary classification, we typically call one class "positive" and the other "negative".
- The positive class is the more important, often smaller one.
- The confusion matrix gives an overview over the errors as well as correct decisions in a tabulated form:

		True Class y	
		$+$	$-$
Decision $\hat{y}$	$+$	True Positive (TP)	False Positive (FP)
	$-$	False Negative (FN)	True Negative (TN)

Here:

- **True Positive** (TP) means that we decide for +1 for a given instance that is really a +1 (correct decision).
- **False Positive** (FP) means that we decide for +1 for a given instance that is actually a -1 (incorrect decision).
- **False Negative** (FN) means that we decide for -1 for a given instance that is actually a +1 (incorrect decision).
- **True Negative** (TN) means that we decide for -1 for a given instance that is really a -1 (correct decision).

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 the ... which ... $f(\mathbf{x}) = \hat{y} \in \{-1, 1\}$.
Note that $\hat{y}$ is a random variable (can be constant), as it is essentially a function of the random input $\mathbf{x}$.

Input x
Data
Learning
Model
Decision

DECISION THEORY 101

The confusion matrix gives rise to common classification/decision criteria, which highlight different aspects of the decision making.

		True Class y +	True Class y −	
Decision $\hat{y}$	+	TP	FP	$\rho_{PPV} = \frac{TP}{TP+FP}$
	−	FN	TN	$\rho_{NPV} = \frac{TN}{FN+TN}$
		$\rho_{TPR} = \frac{TP}{TP+FN}$	$\rho_{TNR} = \frac{TN}{FP+TN}$	$\rho_{ACC} = \frac{TP+TN}{TOTAL}$

- True positive rate ρ_{TPR}: for how many of the true 1s did we decide for 1?
  - ⇝ Population counterpart: $\mathbb{P}(\hat{y} = 1 \mid y = 1)$
- True Negative rate ρ_{TNR}: for how many of the true -1s did we decide for -1?
  - ⇝ Population counterpart: $\mathbb{P}(\hat{y} = -1 \mid y = -1)$
- Positive predictive value ρ_{PPV}: if we decide for 1, how likely is it a true 1?
  - ⇝ Population counterpart: $\mathbb{P}(y = 1 \mid \hat{y} = 1)$
- Negative predictive value ρ_{NPV}: if we decide for -1, how likely is it a true -1?
  - ⇝ Population counterpart: $\mathbb{P}(y = -1 \mid \hat{y} = -1)$
- Accuracy ρ_{ACC}: for how many instances did we decide correctly?
  - ⇝ Population counterpart: $\mathbb{P}(\hat{y} = y)$

- In binary classification, we typically call one class "positive" and the other "negative".
- The positive class is the more important, often smaller one.
- The confusion matrix gives an overview over the errors as well as correct decisions in a tabulated form:

		True Class y +	True Class y −
Decision $\hat{y}$	+	True Positive (TP)	False Positive (FP)
	−	False Negative (FN)	True Negative (TN)

Here:

- **True Positive** (TP) means that we decide for +1 for a given instance that is really a +1 (correct decision).
- **False Positive** (FP) means that we decide for +1 for a given instance that is actually a -1 (incorrect decision).
- **False Negative** (FN) means that we decide for -1 for a given instance that is actually a +1 (incorrect decision).
- **True Negative** (TN) means that we decide for -1 for a given instance that is really a -1 (correct decision).

DECISION THEORY 101

The confusion matrix gives rise to common classification/decision criteria, which highlight different aspects of the decision making

		True Class y		
		$+$	$-$	
Decision	$+$	TP	FP	$\rho_{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}}$
$\hat{y}$	$-$	FN	TN	$\rho_{NPV} = \frac{\text{TN}}{\text{FN}+\text{TN}}$
		$\rho_{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}$	$\rho_{TNR} = \frac{\text{TN}}{\text{FP}+\text{TN}}$	$\rho_{ACC} = \frac{\text{TP}+\text{TN}}{\text{TOTAL}}$

- False positive rate $\rho_{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}}$: for how many of the true -1s did we decide for +1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} = +1 \mid y = -1)$
- False Negative rate $\rho_{FNR} = \frac{\text{FN}}{\text{TP}+\text{FN}}$: for how many of the true 1s did we decide for -1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} = -1 \mid y = 1)$
- Error $\rho_{err} = 1 - \rho_{ACC}$: for how many instances did we decide incorrectly?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} \neq y)$

- True positive rate ρ_{TPR}: for how many of the true 1s did we decide for 1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} = 1 \mid y = 1)$
- True Negative rate ρ_{TNR}: for how many of the true -1s did we decide for -1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} = -1 \mid y = -1)$
- Positive predictive value ρ_{PPV}: if we decide for 1, how likely is it a true 1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(y = 1 \mid \hat{y} = 1)$
- Negative predictive value ρ_{NPV}: if we decide for -1, how likely is it a true -1?
  - $\leadsto$ Population counterpart: $\mathbb{P}(y = -1 \mid \hat{y} = -1)$
- Accuracy ρ_{ACC}: for how many instances did we decide correctly?
  - $\leadsto$ Population counterpart: $\mathbb{P}(\hat{y} = y)$

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 $\mathbf{A}$ 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 $\mathbf{A}$.

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

The confusion matrix gives rise to common classification/decision criteria, which highlight different aspects of the decision making

		True Class y		
		$+$	$-$	
Decision $\hat{y}$	$+$	TP	FP	$\rho_{PPV} = \frac{TP}{TP+FP}$
	$-$	FN	TN	$\rho_{NPV} = \frac{TN}{FN+TN}$
		$\rho_{TPR} = \frac{TP}{TP+FN}$	$\rho_{TNR} = \frac{TN}{FP+TN}$	$\rho_{ACC} = \frac{TP+TN}{TOTAL}$

- False positive rate $\rho_{FPR} = \frac{FP}{FP+TN}$: for how many of the true -1's did we decide for +1?

  ⇝ Population counterpart: $\mathbb{P}(\hat{y} = +1 \mid y = -1)$
- False Negative rate $\rho_{FNR} = \frac{FN}{TP+FN}$: for how many of the true 1s did we decide for -1?

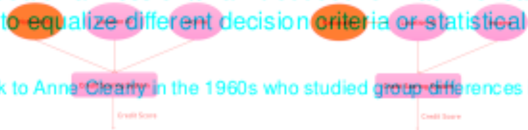
  ⇝ Population counterpart: $\mathbb{P}(\hat{y} = -1 \mid y = +1)$
- Error $\rho_{err} = 1 - \rho_{ACC}$ for how many instances did we decide incorrectly?

  ⇝ Population counterpart: $\mathbb{P}(\hat{y} \neq y)$

REMOVING SENSITIVE FEATURES

- A straightforward (and also naive) approach is to simply ignore or remove all sensitive features at prediction time. This approach is often called *fairness through unawareness*. However, in many cases other non-sensitive features are slightly correlated with the sensitive one(s). For example:
  - gender with hobbies or interests (job application),
  - race and zip code (law systems),
  - nationality and location id (credit application),
- Thus, an ML model trained on data including the sensitive features might combine the corresponding correlated non-sensitive features to make essentially the same decision, as it still seeks to maximize accuracy.

- Note that we assume that the attribute/feature observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ are random observations of the random vector $\mathbf{x} = (x_1, \dots, x_p)^\top$ with distribution $\mathbb{P}_x$. Accordingly, the j-th attribute/feature vector $\mathbf{x}_j = \left(x_j^{(1)}, \dots, x_j^{(n)}\right)$ 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 $\mathbf{A}$ 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 $\mathbf{A}$.



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

INDEPENDENCE AS A FAIRNESS CRITERION

- A quite natural fairness criterion is given by ensuring (stochastic) independence between the decision $\hat{y}$ and the sensitive attributes/features $\mathbf{A}$:

$$\hat{y} \perp\!\!\!\perp \mathbf{A}$$

- This is equivalent to ensuring an equal "acceptance rate" among all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

$$\mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \tilde{\mathbf{a}})$$

- This criterion is also known as statistical/demographic parity, group fairness, equal positive rates or Darlington's fourth criterion.
- One can relax the criterion by introducing a fixed tolerance parameter $\epsilon > 0$ and only require that for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$ it holds that

$$\left|\mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \mathbf{a}) - \mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \tilde{\mathbf{a}})\right| \leq \epsilon$$

DOWNSIDES OF INDEPENDENCE AS A FAIRNESS CRITERION

- Independence does not take into account that the outcome y might be correlated with $\mathbf{A}$, which means that the different realizations of $\mathbf{A}$ 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.

- A quite natural fairness criterion is given by ensuring (stochastic) independence between the decision $\hat{y}$ and the sensitive attributes/features $\mathbf{A}$:

$$\hat{y} \perp \mathbf{A}$$

- This is equivalent to ensuring an equal acceptance rate among all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

$$\mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \tilde{\mathbf{a}})$$

- This criterion is also known as statistical/demographic parity, group fairness, equal positive rates or Darlington's fourth criterion.
- One can relax the criterion by introducing a fixed tolerance parameter $\epsilon > 0$ and only require that for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$ it holds that

$$\left|\mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \mathbf{a}) - \mathbb{P}(\hat{y} = 1 \mid \mathbf{A} = \tilde{\mathbf{a}})\right| \le \epsilon$$

ACHIEVING INDEPENDENCE VIA REPRESENTATION LEARNING

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

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

E AS A FAIRNESS

- 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.

SEPARATION AS A FAIRNESS CRITERION

- As we discussed above, the independence criterion does not take correlation between y and $\mathbf{A}$ into account. As an alternative fairness criterion one can consider *separation*, which ensures (stochastic) independence between the decision $\hat{y}$ and the sensitive attributes $\mathbf{A}$ given y :

$$\hat{y} \perp\!\!\!\perp \mathbf{A} \mid y$$

- This is equivalent to equalize the (population) error rates for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

$$\mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \tilde{\mathbf{a}})$$

(equal false positive rates)

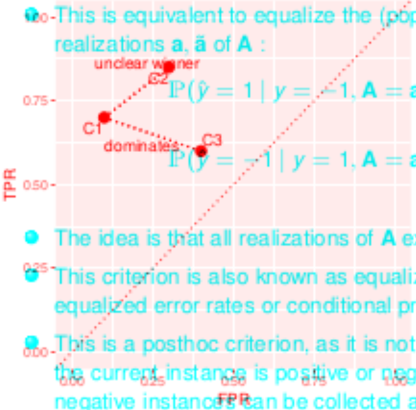
$$\mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \tilde{\mathbf{a}})$$

(equal false negative rates)

- The idea is that all realizations of $\mathbf{A}$ experience the same FPR and FNR.
- This criterion is also known as equalized odds, avoiding disparate mistreatment, equalized error rates or conditional procedure accuracy.
- This is a posthoc criterion, as it is not known at the time of the decision whether the current instance is positive or negative. Only in hindsight the positive and negative instances can be collected and compared with the decisions made.

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

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

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

ESS CRITERION

- between y and $\mathbf{A}$ into account. As an alternative fairness criterion one can consider *separation*, which ensures (stochastic) independence between the decision $\hat{y}$ and the sensitive attributes $\mathbf{A}$ given y :

$$\hat{y} \perp\!\!\!\perp \mathbf{A} \mid y$$

- This is equivalent to equalize the (population) error rates for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

$$\mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \tilde{\mathbf{a}})$$

(equal false positive rates)

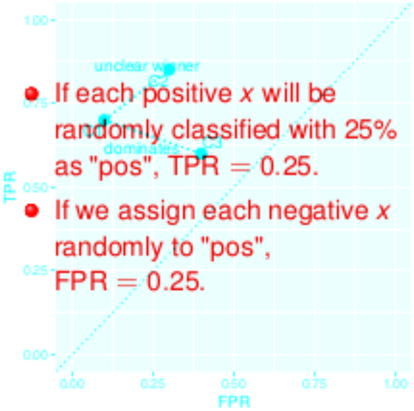
$$\mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \tilde{\mathbf{a}})$$

(equal false negative rates)

- The idea is that all realizations of $\mathbf{A}$ experience the same FPR and FNR.
- This criterion is also known as equalized odds, avoiding disparate mistreatment, equalized error rates or conditional procedure accuracy.
- This is a posthoc criterion, as it is not known at the time of the decision whether the current instance is positive or negative. Only in hindsight the positive and negative instances can be collected and compared with the decisions made.

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.

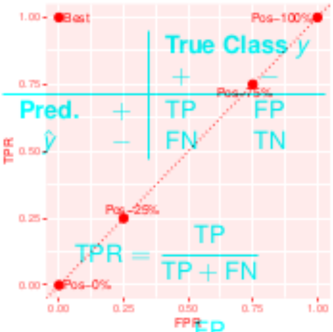


		True Class y	
		$+$	$-$
Pred. $\hat{y}$	$+$	TP	FP
	$-$	FN	TN

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

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

- The best classifier lies on the top-left corner, where FPR equals 0 and TPR is maximal.
- The diagonal is worst as it corresponds to a classifier producing random labels (with different proportions).
- If each positive x will be randomly classified with 25% as "pos", $\text{TPR} = 0.25$.
- If we assign each negative x randomly to "pos", $\text{FPR} = 0.25$.



INTERLUDE: ROC CURVES FOR SCORING CLASSIFIERS

- Many binary classification methods use a score (function) $s : \mathcal{X} \to \mathbb{R}$ and a threshold value c to make the prediction (decision):

$$f(\mathbf{x}) = 2 \cdot \mathbb{1}_{[s(\mathbf{x}) \geq c]} - 1.$$

- The choice of threshold affects the TPR and FPR, so it is interesting to examine the effects of different thresholds on these.
- A ROC curve is a visual tool to help in finding good threshold values.

- The best classifier lies on the top-left corner, where FPR equals 0 and TPR is maximal.
- The diagonal is worst as it corresponds to a classifier producing random labels (with different proportions).
- If each positive x will be randomly classified with 25% as "pos", $\text{TPR} = 0.25$.
- If we assign each negative x randomly to "pos", $\text{FPR} = 0.25$.

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.)

- Many binary classification methods use a score (function) $s : \mathcal{X} \to \mathbb{R}$ and a threshold value c to make the prediction (decision):

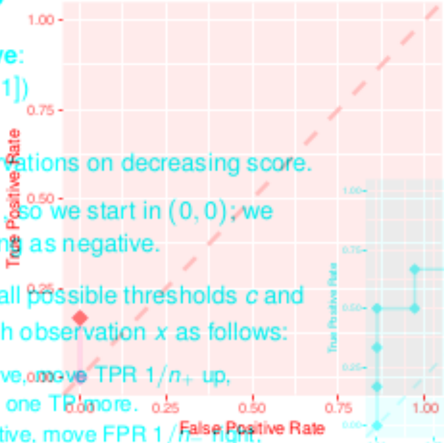
$$f(x) = 2 \cdot \mathbb{1}_{[s(x) \geq c]} - 1.$$

- The choice of threshold affects the TPR and FPR, so it is interesting to examine the effects of different thresholds on these.
- A ROC curve is a visual tool to help in finding good threshold values.

DRAWING ROC CURVES: EXAMPLE

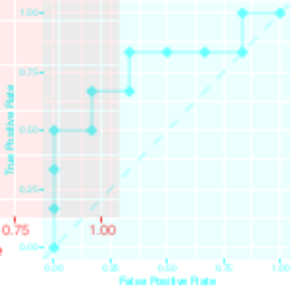
SCORING CLASSIFIERS

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



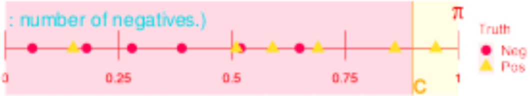
To draw a ROC curve:
(W.l.o.g. $s : \mathcal{X} \to [0, 1]$)

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.



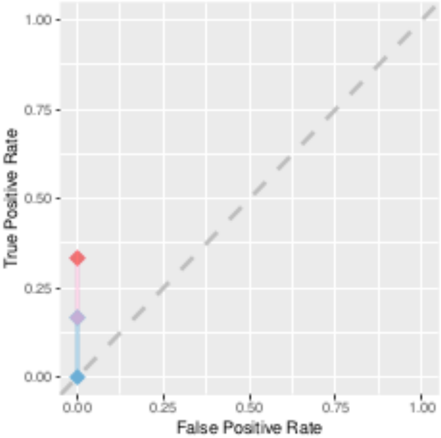
(n_+: number of positives, n_-: number of negatives.)

$c = 0.9$
$\to$ TPR = 0.167
$\to$ FPR = 0



DRAWING ROC CURVES: EXAMPLE

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



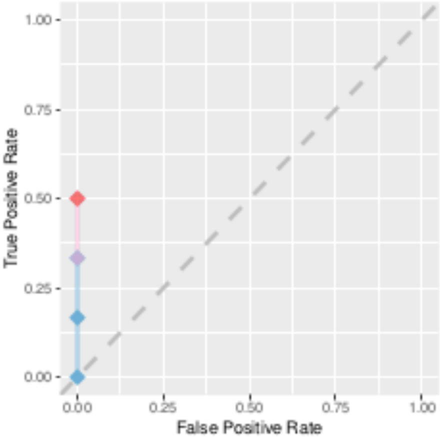
$c = 0.85$

$\rightarrow$ TPR = 0.333

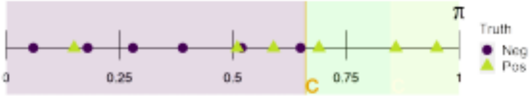
$\rightarrow$ FPR = 0

DRAWING ROC CURVES: EXAMPLE

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06

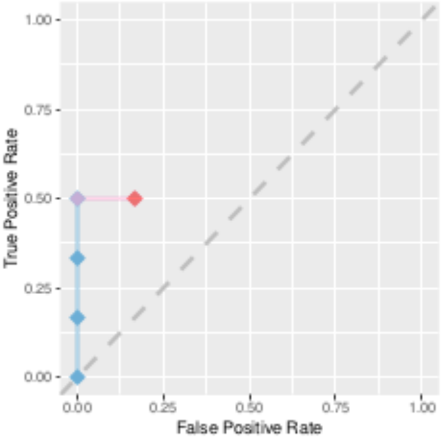


$c = 0.66$
$\rightarrow$ TPR = 0.533
$\rightarrow$ FPR = 0

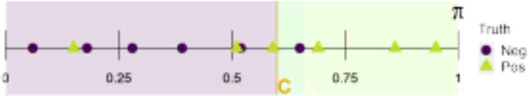


DRAWING ROC CURVES: EXAMPLE

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06

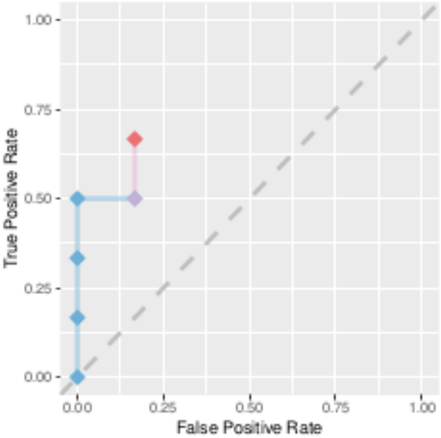


$c = 0.66$
$\rightarrow$ TPR = 0.5
$\rightarrow$ FPR = 0.167



DRAWING ROC CURVES: EXAMPLE

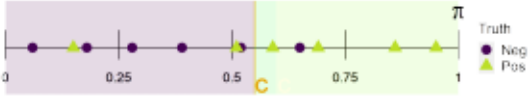
#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



$c = 0.55$

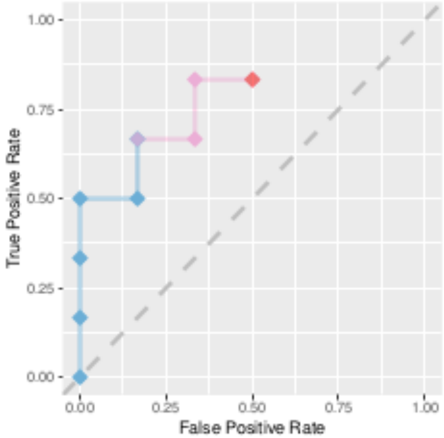
$\rightarrow$ TPR = 0.667

$\rightarrow$ FPR = 0.167



DRAWING ROC CURVES: EXAMPLE

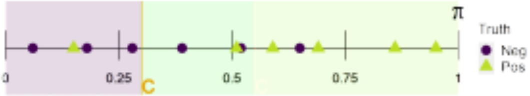
#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



$c = 0.35$

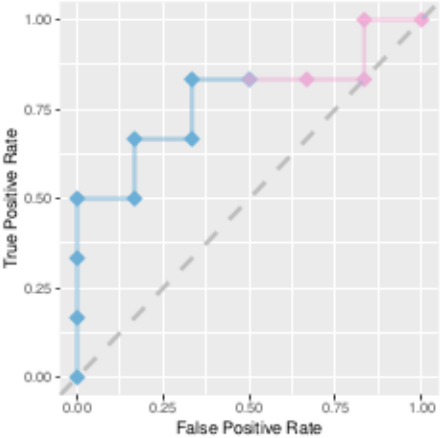
$\rightarrow$ TPR = 0.833

$\rightarrow$ FPR = 0.567



DRAWING ROC CURVES: EXAMPLE

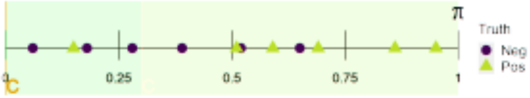
#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



$c = 0.3$

$\rightarrow$ TPR = 0.833

$\rightarrow$ FPR = 0.5



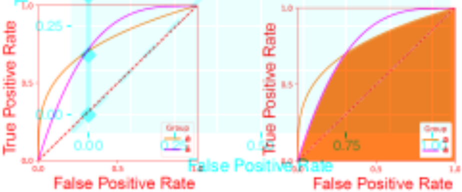
SEPARATION AND ROC CURVES

- Separation, i.e.,

$$\mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \tilde{\mathbf{a}}) \qquad \text{(equal FPRs)}$$

$$\mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \tilde{\mathbf{a}}) \qquad \text{(equal FNRs)}$$

  means that all ROC curves of a classifier restricted on realizations of **A** should be the same. This implies that the ROC curve of the score-based classifier conditional on realizations of **A** must be "under" all ROC curves.



- In practice, we should never obtain a classifier below the diagonal.
- Inverting the predicted labels ($-1 \mapsto 1$ and $1 \mapsto -1$) will result in a reflection at the diagonal $\Rightarrow$ $\text{TPR}_{\text{new}} = 1 - \text{TPR}$ and $\text{FPR}_{\text{new}} = 1 - \text{FPR}$.

DOWNSIDES OF SEPARATION AS A FAIRNESS CRITERION

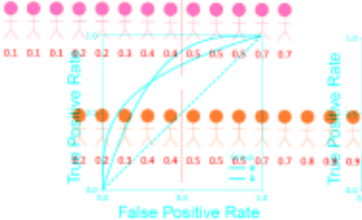
- Separation, i.e.,

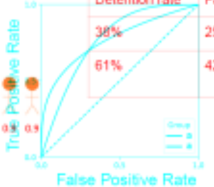
$$\mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = 1 \mid y = -1, \mathbf{A} = \bar{\mathbf{a}}) \quad \text{(equal FPRs)}$$

$$\mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \mathbf{a}) = \mathbb{P}(\hat{y} = -1 \mid y = 1, \mathbf{A} = \bar{\mathbf{a}}) \quad \text{(equal FNRs)}$$

means that all ROC curves of a classifier restricted on realizations of **A** should be the same. This implies that the ROC curve of the score-based classifier conditional on realizations of **A** must be "under" all ROC curves.

- 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 $c = 0.5$.



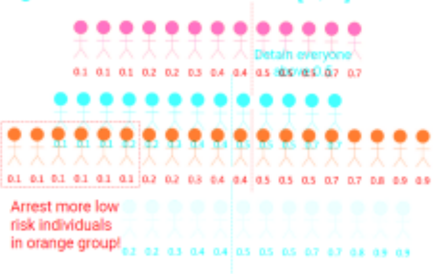


Detention rate	False pos. rate
38%	25%
61%	42%

- The classifier is not satisfying separation as FPR and FNR are not the same among the two groups.
- In practice, we should never obtain a classifier below the diagonal.
- Inverting the predicted labels ($-1 \mapsto 1$ and $1 \mapsto -1$) will result in a reflection at the diagonal $\Rightarrow$ $\text{TPR}_{new} = 1 - \text{TPR}$ and $\text{FPR}_{new} = 1 - \text{FPR}$.

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.
- 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%	~~42%~~ 26%

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

- Thus, as with achieving independence, separation can also lead to undesirable outcomes.
- The classifier is not satisfying separation as FPR and FNR are not the same among the two groups.

SUFFICIENCY AS A FAIRNESS CRITERION

- Another idea to specify a fairness criterion for score-based classifiers is that the corresponding score random variable $\mathbf{S} = s(\mathbf{x})$ already subsumes the sensitive attributes $\mathbf{A}$ for the prediction:

$$y \perp\!\!\!\perp \mathbf{A} \mid \mathbf{S}$$

- This is equivalent to require that the fraction of positive instances assigned some score s is the same for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

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

- This criterion is also known as Cleary's model, conditional use accuracy or calibration within groups.

- This is an a priori guarantee: The decision maker sees the score value and knows based on this what the frequency of positives is.

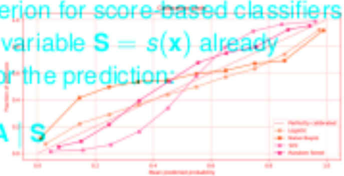
SUFFICIENCY AND CALIBRATION

- Sufficiency is very closely related to the concept of *calibration* of a probabilistic classifier, i.e., a classifier such that $s : \mathcal{X} \to [0, 1]$. More specifically, a probabilistic classifier is called *calibrated* if for all $s \in [0, 1]$

$$\mathbb{P}(y = 1 \mid \mathbf{S} = s) = s.$$

- Note that:
  1. This condition means that the set of all instances assigned a score value s also account for a proportion s of positive instances.
  2. It is a condition over all features and in particular on the sensitive ones. Consequently, it does not mean that at the level of a single value of $\mathbf{A}$ a score of s corresponds to a probability s of a positive outcome.
- The notion of calibration can be specified also on the group level, that is, a probabilistic classifier is called *calibrated on the group level* if for all $s \in [0, 1]$ and all possible realizations $\mathbf{a}$ of $\mathbf{A}$:

$$\mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \mathbf{a}) = s.$$



- Another idea to specify a fairness criterion for score-based classifiers is that the corresponding score random variable $\mathbf{S} = s(\mathbf{x})$ already subsumes the sensitive attribute $\mathbf{A}$ for the prediction:

$$y \perp\!\!\!\perp \mathbf{A} \mid \mathbf{S}$$

- This is equivalent to require that the fraction of positive instances assigned some score s is the same for all possible realizations $\mathbf{a}, \tilde{\mathbf{a}}$ of $\mathbf{A}$:

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

- This criterion is also known as Cleary's model, conditional use accuracy or calibration within groups.
- This is an a priori guarantee: The decision maker sees the score value and knows based on this what the frequency of positives is.

SUFFICIENCY AND CALIBRATION / 2

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

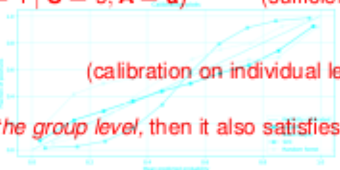
vs.

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

- 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.

- Sufficiency is very closely related to the concept of *calibration* of a probabilistic classifier, i.e., a classifier such that $s : \mathcal{X} \to [0,1]$. More specifically, a probabilistic classifier is called *calibrated* if for all $s \in [0,1]$

$$\mathbb{P}(y=1 \mid \mathbf{S}=s) = s.$$

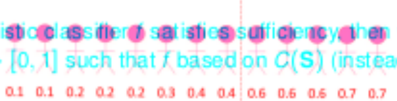


- Note that:
  1. This condition means that the set of all instances assigned a score value s also account for a proportion s of positive instances.
  2. It is a condition over all features and in particular on the sensitive ones. Consequently, it does not mean that at the level of a single value of $\mathbf{A}$ a score of s corresponds to a probability s of a positive outcome.
- The notion of calibration can be specified also on the group level, that is, a probabilistic classifier is called *calibrated on the group level* if for all $s \in [0,1]$ and all possible realizations $\mathbf{a}$ of $\mathbf{A}$:

$$\mathbb{P}(y=1 \mid \mathbf{S}=s, \mathbf{A}=\mathbf{a}) = s.$$

DOWNSIDES OF SUFFICIENCY AS A FAIRNESS CRITERION

- Consider a group of blue people and assume we are interested in deciding whether we should detain (positive class) a person and use a scoring classifier with scores in $[0, 1]$ and a threshold $c = 0.5$. Suppose we know the true probability that a person will reoffend and the scores are equal to these.



- Assume that there are two groups among the blue people:

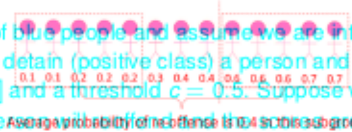
$$\mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \mathbf{a}) = \mathbb{P}(y = 1 \mid \mathbf{S} = s, \mathbf{A} = \tilde{\mathbf{a}}) \qquad \text{(sufficiency)}$$

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

- 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

- Consider a group of blue people and assume we are interested in deciding whether we should detain (positive class) a person and use a scoring classifier with scores in [0, 1] and a threshold $c = 0.5$. Suppose we know the true probability that a person will reoffend and the scores are equal to these.



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

- Assume that there are two groups among the blue people:

RELATIONSHIPS BETWEEN THE FAIRNESS CRITERIA

- We have considered three fairness criteria:

$$\hat{y} \perp\!\!\!\perp \mathbf{A} \qquad \text{(Independence)}$$

$$\hat{y} \perp\!\!\!\perp \mathbf{A} \mid y \qquad \text{(Separation)}$$

$$y \perp\!\!\!\perp \mathbf{A} \mid \mathbf{S} \qquad \text{(Sufficiency)}$$

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

- A tempting question is how these criteria relate to each other.
- **(Informal) Theorem.** Any two of these criteria are mutually exclusive in general.
- As a consequence, we cannot impose multiple of these criteria as hard constraints on the classifier.
- A possible solution to this issue is to consider relaxed version of these criteria as constraints.

FINAL REMARKS

- Fairness is a challenging issue as also philosophers and social scientists have been trying to define it for decades.
- Due to the increased use of ML methods in automated decision making there is a need to think about fairness in more detail.
- Fairness criteria such as independence, sufficiency and separation are a statistical objective way to incorporate fairness aspects into ML methods. However, on their own they are neither equivalent to a “proof of fairness” nor are they prefect objective functions for this purpose.
- In summary, there are three ways to tackle the question: “how to satisfy fairness criteria?”
  1. Pre-processing phase: Adjust the feature space to be uncorrelated with the sensitive attribute.
  2. Training phase: Build the constraint into the optimization process for the classifier.
  3. Post-processing phase: Adjust a learned classifier so that it is uncorrelated to the sensitive attribute.

RELATIONSHIPS BETWEEN THE FAIRNESS CRITERIA

- We have considered three fairness criteria:

$$\hat{y} \perp \mathbf{A} \qquad \text{(Independence)}$$
$$\hat{y} \perp \mathbf{A} \mid y \qquad \text{(Separation)}$$
$$y \perp \mathbf{A} \mid \mathbf{S} \qquad \text{(Sufficiency)}$$

- A tempting question is how these criteria relate to each other.
- **(Informal) Theorem.** Any two of these criteria are mutually exclusive in general.
- As a consequence, we cannot impose multiple of these criteria as hard constraints on the classifier.
- A possible solution to this issue is to consider relaxed version of these criteria as constraints.

FINAL REMARKS

- Fairness is a challenging issue as also philosophers and social scientists have been trying to define it for decades.
- Due to the increased use of ML methods in automated decision making there is a need to think about fairness in more detail.
- Fairness criteria such as independence, sufficiency and separation are a statistical objective way to incorporate fairness aspects into ML methods. However, on their own they are neither equivalent to a “proof of fairness” nor are they prefect objective functions for this purpose.
- In summary, there are three ways to tackle the question: “how to satisfy fairness criteria?”
  1. Pre-processing phase: Adjust the feature space to be uncorrelated with the sensitive attribute.
  2. Training phase: Build the constraint into the optimization process for the classifier.
  3. Post-processing phase: Adjust a learned classifier so that it is uncorrelated to the sensitive attribute.