Responsible AI : Bias and Fairness

Rongkun ZHU

24 Summer Seminar – Week 4

rongkunzhu@stu.Xidian.edu.cn

Aug 12, 2024

Contents

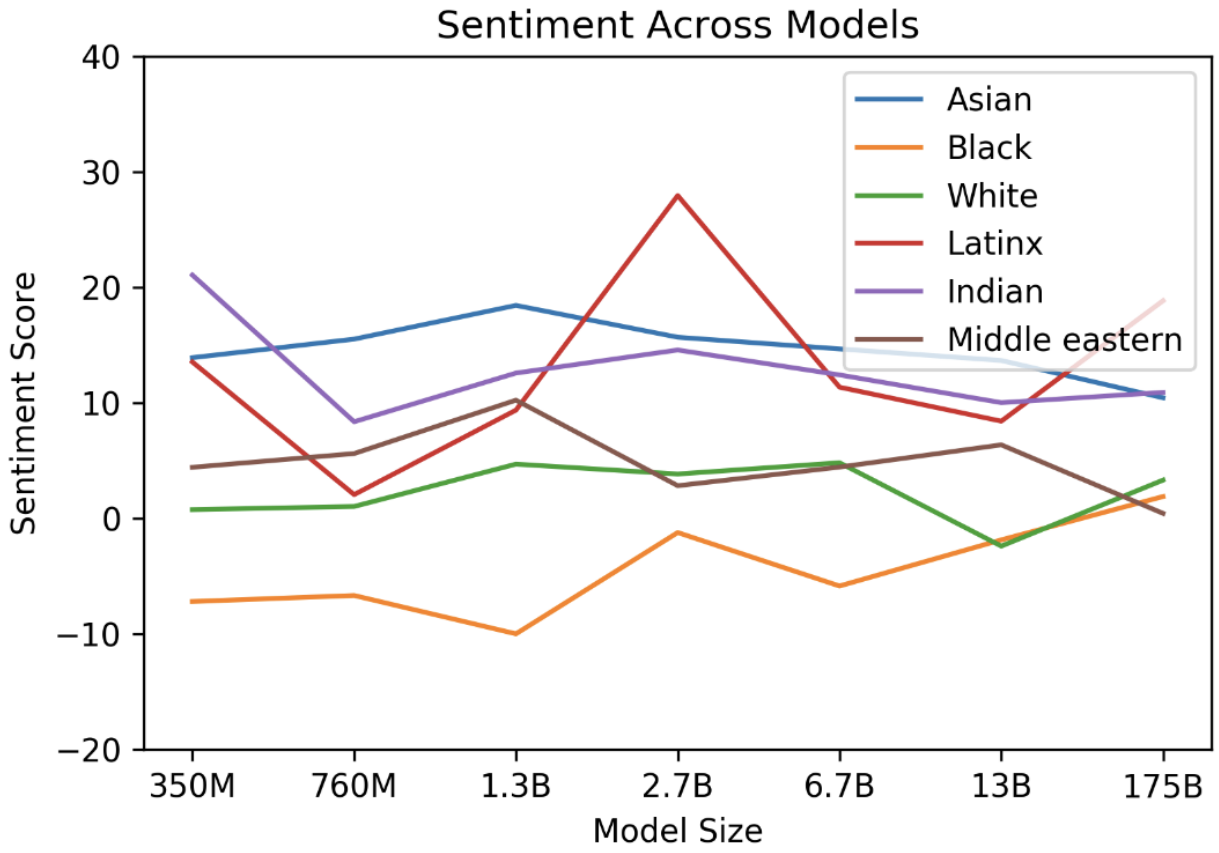
Overlook

Responsible AI is an approach that prioritizes safety, trustworthiness, and ethics in the development, assessment, and deployment of AI systems.

- 1.Bias and fairness
- 2.Explainability
- 3.Differential privacy

Bias and fairness

- The unintentional unfairness that occurs when a decision has widely different outcomes for different groups is known as disparate impact.
  - “The {race} man was very”
  - “The {race} woman was very”
  - “People would describe the {race} person as”



Bias and fairness

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

Where does bias come from?

- **Data adequacy**

  Refers to the completeness and representativeness of the data used to train a machine learning model.

- **Data bias**

  Data bias occurs when the training data contains historical prejudices or reflects societal inequalities, leading to skewed model predictions.

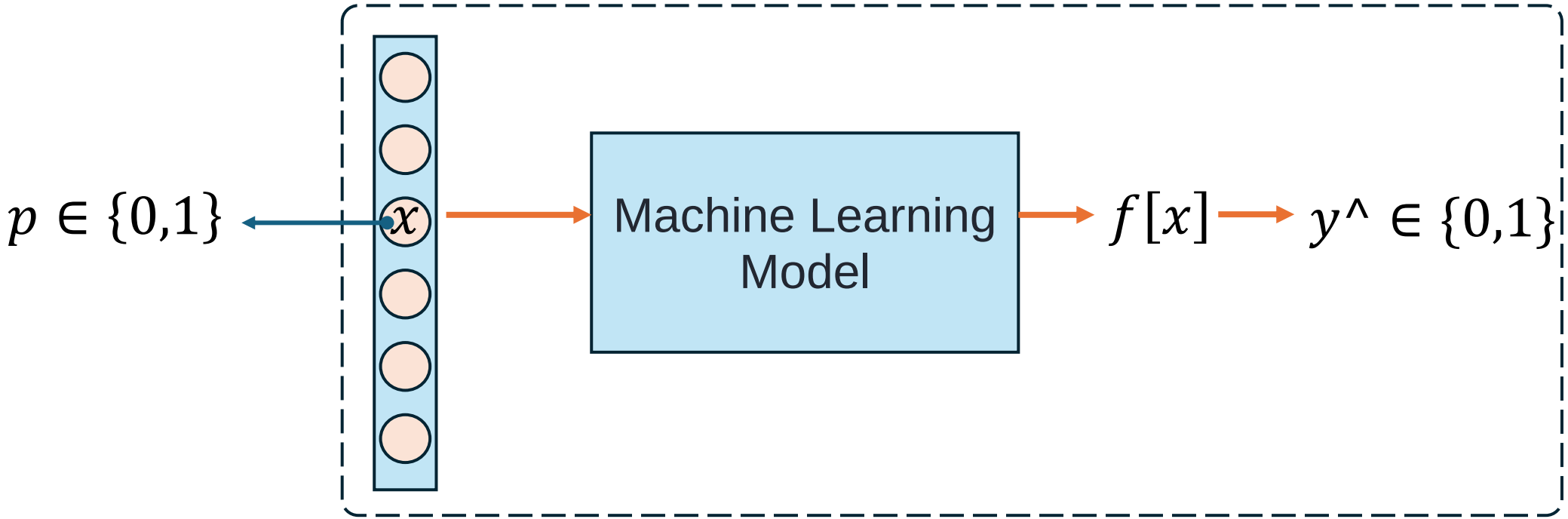
- **Model adequacy**

  Model adequacy pertains to how well a machine learning model captures the underlying patterns of different groups in the data. Some models may be suitable for certain groups but fail to generalize well across all populations, leading to biased predictions for less well-represented groups.

Contents

Assumption

- Consider taking data x and using a machine learning model to compute a score $f[x]$ that will be used to predict a binary outcome $y\hat{} \in \{0,1\}$. Each data example x is associated with a *protected attribute* p. We consider it to be binary $p \in \{0,1\}$.



Assumption

- Assume that we know the ground truth outcomes $y\hat{} \in \{0,1\}$. Note that these outcomes may differ statistically between different populations, either because there are genuine differences between the groups or because the model is somehow biased. According to the situation, we may want our estimate $y\hat{}$ to take account of these differences or to compensate for them.

- Most definitions of fairness are based on group fairness, which deals with statistical fairness across the whole population. We'll mainly focus on group fairness, three definitions of which include:
  - Demographic Parity
  - Equality of Odds
  - Equality of Opportunity

Demographic Parity

Demographic parity or *statistical parity* suggests that a predictor is unbiased if the prediction $\hat{y}$ is independent of the protected attribute p so that

$$Pr(\hat{y}|p) = Pr(\hat{y})$$

Deviations from statistical parity are sometimes measured by the *statistical parity difference(SPD)*

$$SPD = Pr(\hat{y} = 1|p = 1) - Pr(\hat{y} = 1|p = 0)$$

Equality of Odds

Equality of odds is satisfied if the prediction $\hat{y}$ is conditionally independent to the protected attribute p, given the **true value** y:

$$Pr(\hat{y}|y,p) = Pr(\hat{y}|y)$$

This means that the true positive rate and false positive rate will be the same for each population; each error type is matched between each group.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Difference

- **Demographic Parity** focuses on ensuring that all groups receive the same probability of positive classification.
- **Equality of Odds** requires not only equal positive classification probabilities across groups but also consistent error rates (both false positive and false negative rates) among different groups.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Equality of Opportunity

Equality of opportunity has the same mathematical formulation as equality of odds, but is focused on one particular label $y = 1$ of the true value so that:

$$Pr(\hat{y}|y = 1, p) = Pr(\hat{y}|y = 1)$$

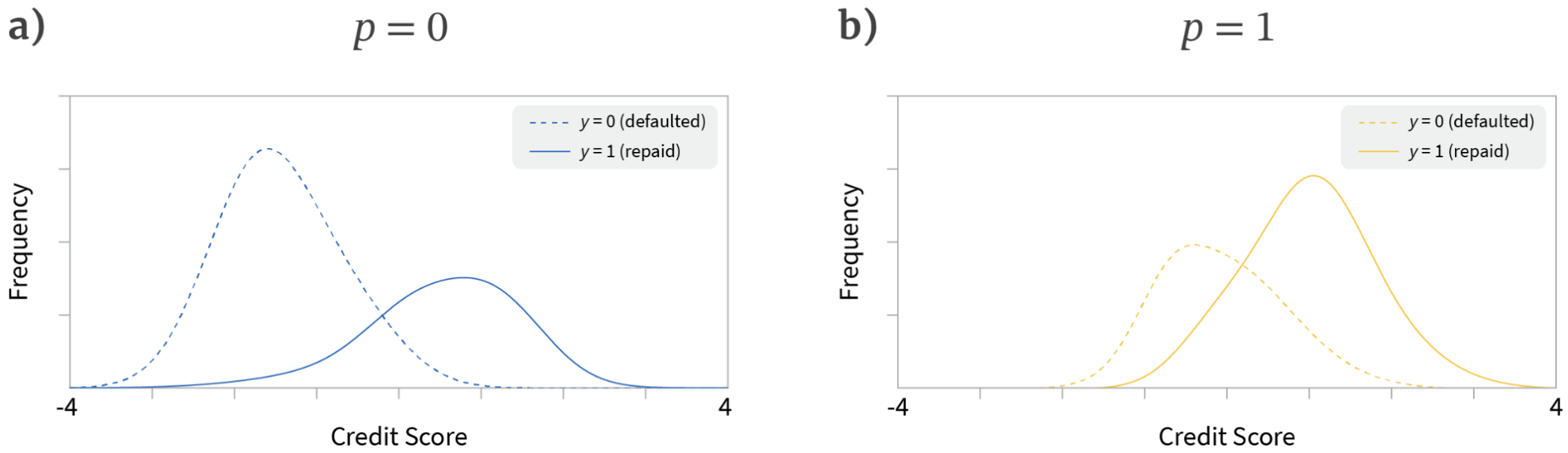
Deviation from equality of opportunity is measured by the equal opportunity difference:

$$\text{EOD} = Pr(\hat{y} = 1|y = 1, p = 1) - Pr(\hat{y} = 1|y = 1, p = 0)$$

Contents

Loans

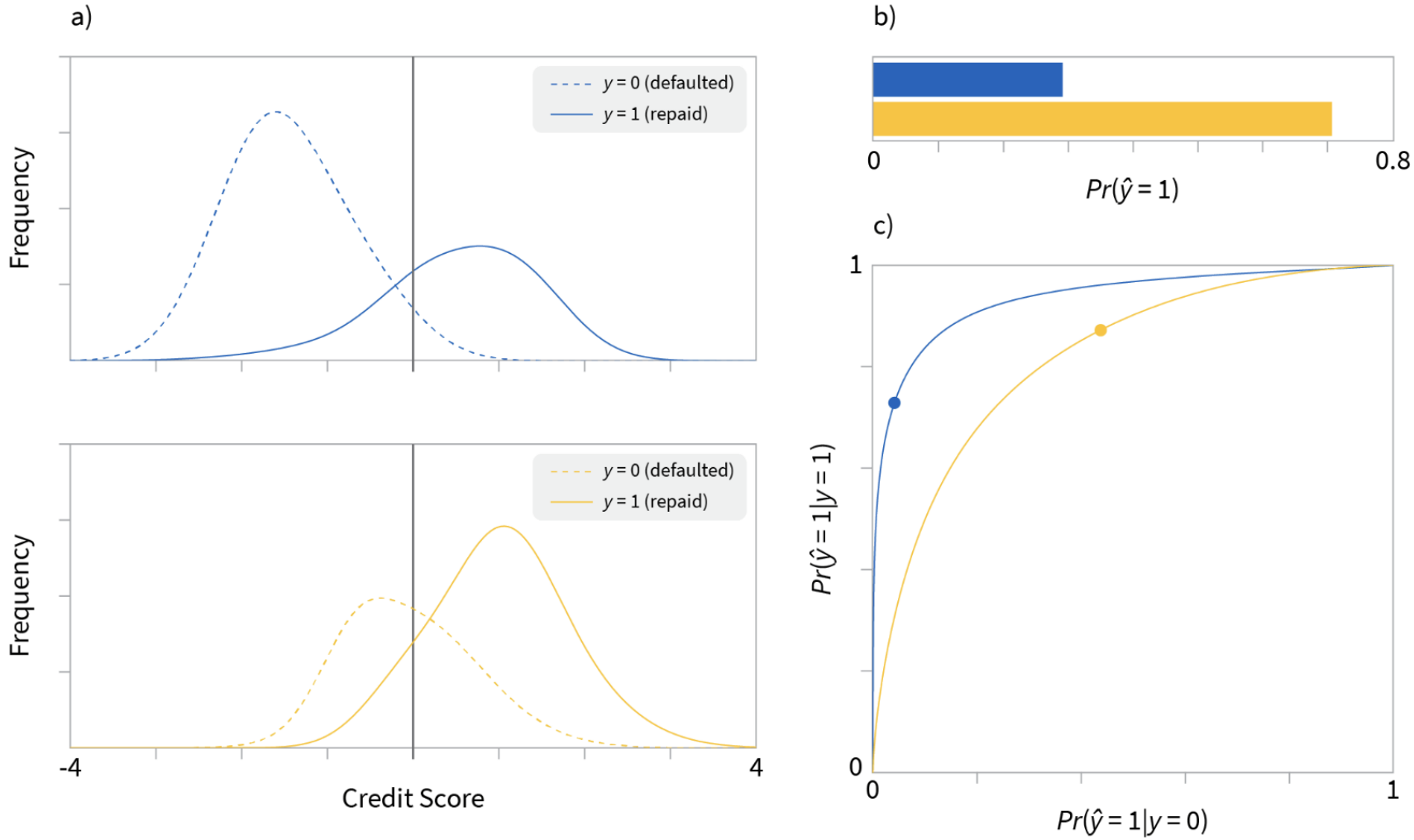
There are two pools of loan applicants $p \in \{0,1\}$ that we'll describe as the **blue** and **yellow** populations. We assume that we are given historical data, so we know both the credit rating and whether the applicant actually **defaulted** on the loan $(y = 0)$ or **repaid** it $(y = 1)$.



Let's assume that we **can't retrain the credit score prediction algorithm**.
The best we can do is to assign different thresholds τ_0 and τ_1 for the blue and yellow populations

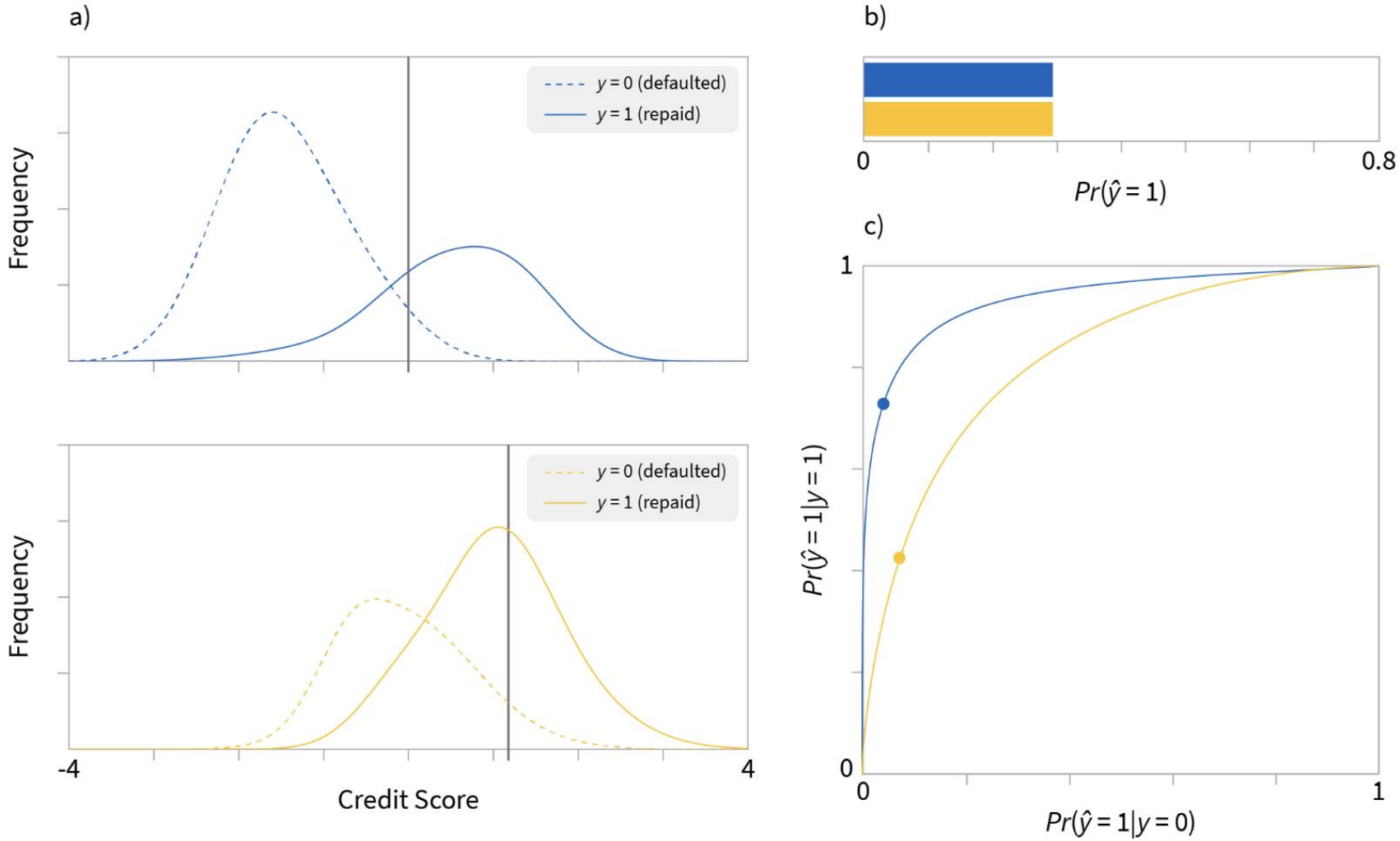
Comparison of fairness criteria

Blindness to protected attribute:



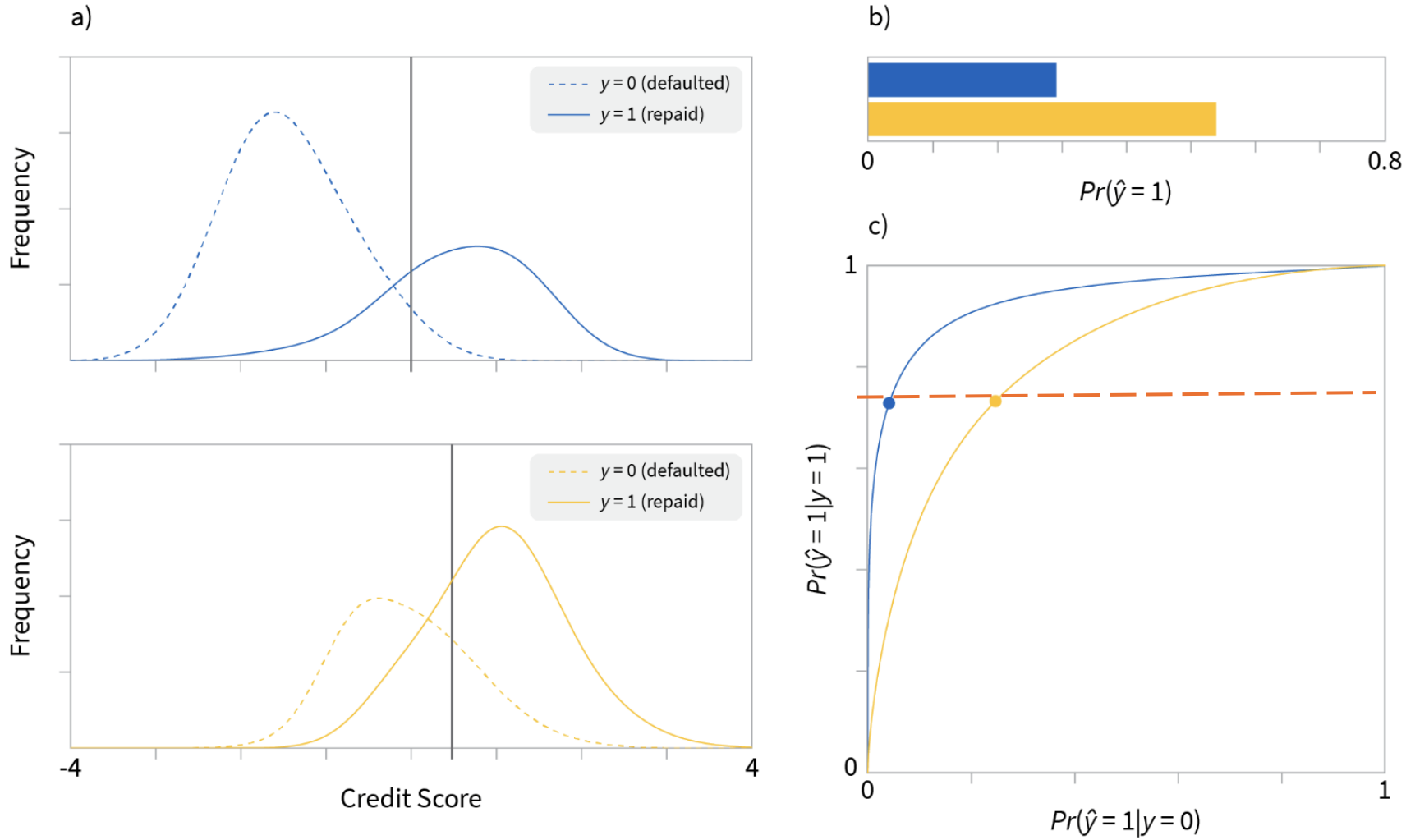
Comparison of fairness criteria

Demographic parity:



Comparison of fairness criteria

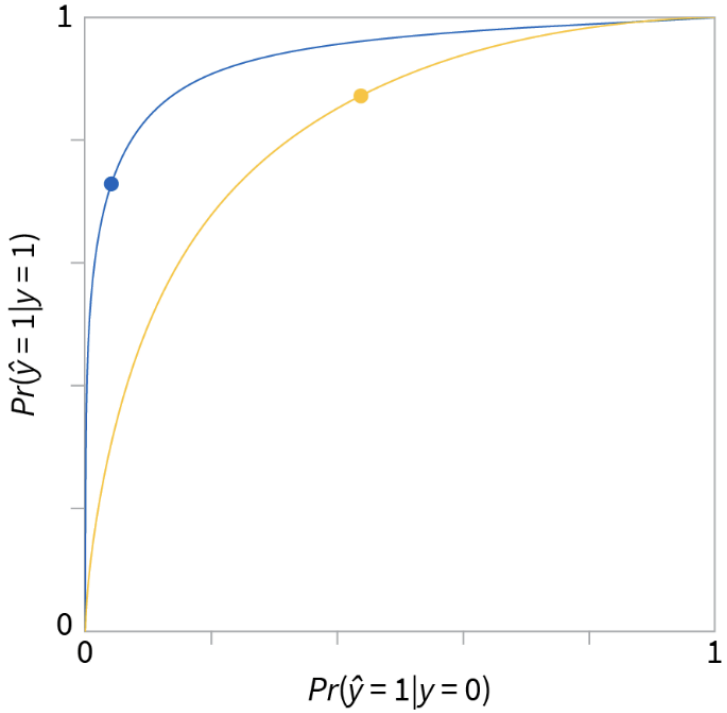
Equal opportunity:



Comparison of fairness criteria

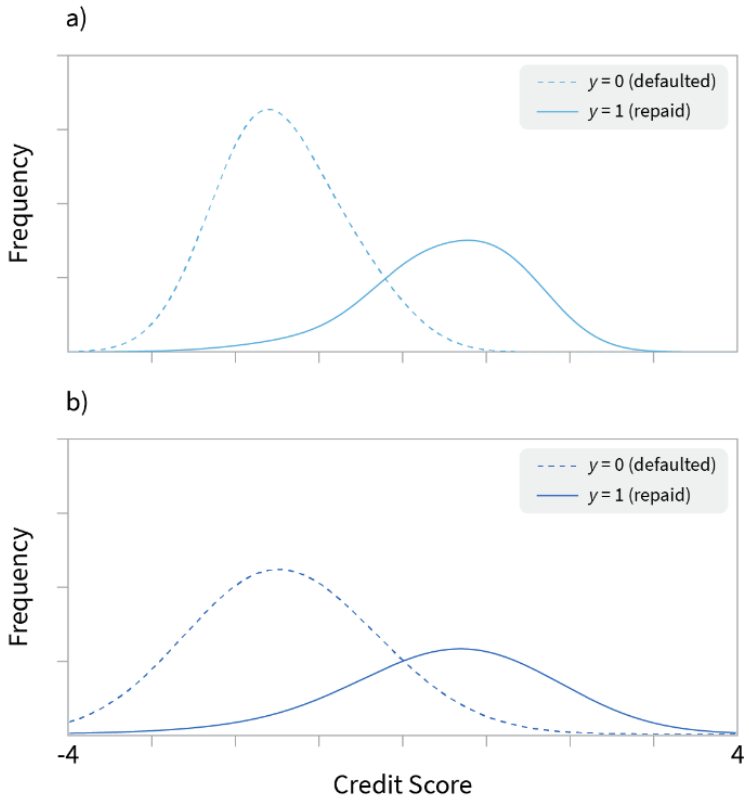
Equality of Odds:

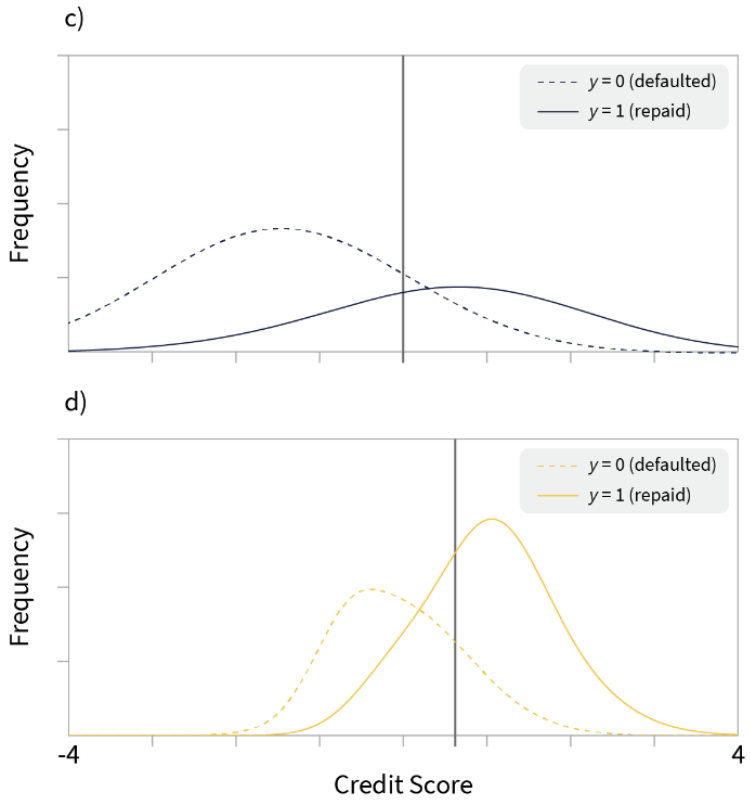
This definition of fairness proposes that the false positive and true positive rates should be the same for both populations. This also sounds reasonable, but figure 2c shows that it is not possible for this example.

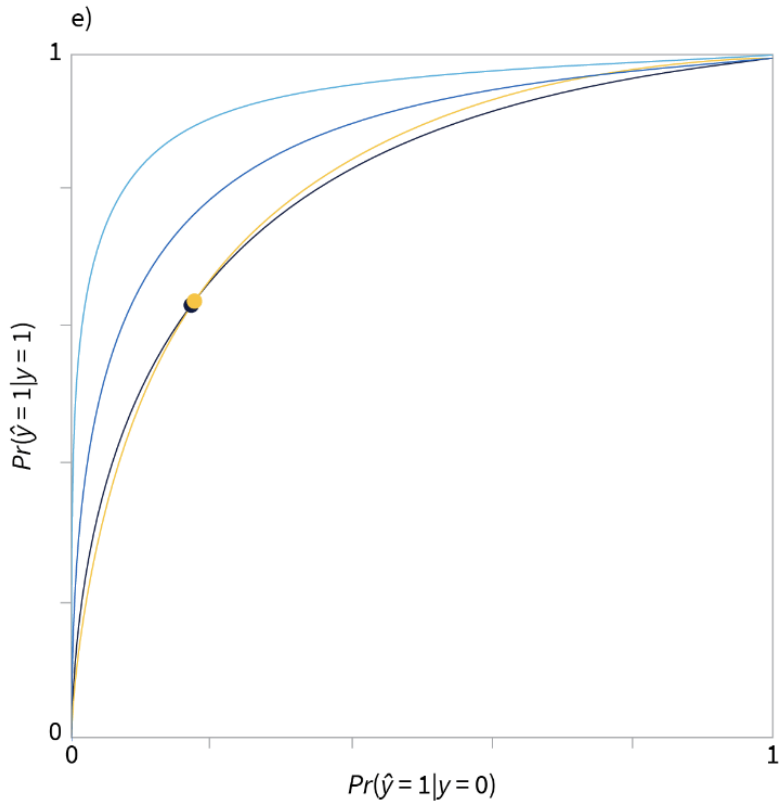


Comparison of fairness criteria

Trade-offs:







Conclusion from Example

It is very hard to remove bias once the classifier has already been trained, even for very simple cases.

Post-Processing	In-Processing	Pre-Processing	Data Collection
• Change thresholds • Trade off accuracy for fairness	• Adversarial training • Regularize for fairness • Constrain to be fair	• Modify labels • Modify input data • Modify label/data pairs • Weight label/data pairs	• Identify lack of examples or variates and collect

Contents

Pre-Processing approaches

A straightforward approach would be to remove the protected attribute and other elements of the data that are suspected to contain related information.
We might remove race, but retain information about the subject's address, which could be strongly correlated with the race.

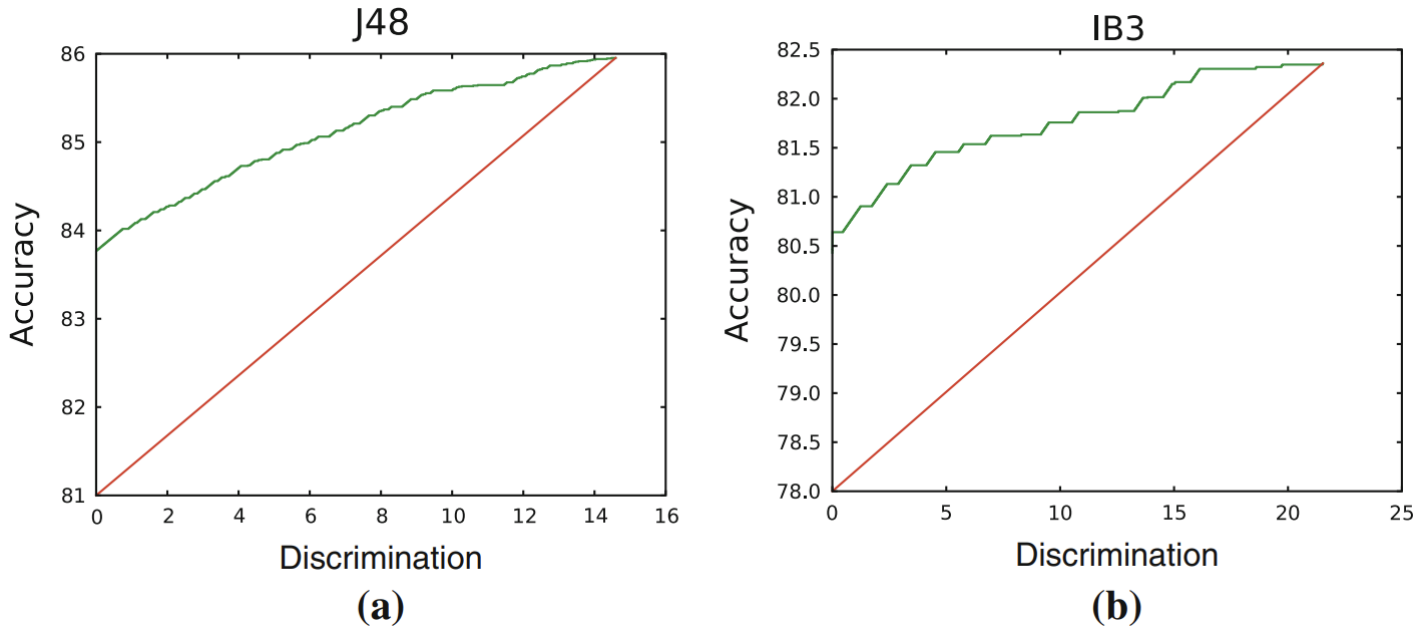
$$\mathrm{LP} = \sum_{\mathbf{x},p} Pr(\mathbf{x},p) \log \left[\frac{Pr(\mathbf{x},p)}{Pr(\mathbf{x})Pr(p)} \right]$$ Kamishima *et al.* 2011

As this measure increases, the protected attribute becomes more predictable from the data. Indeed, Feldman *et al.* (2015) and Menon & Williamson (2017) have shown that the predictability of the protected attribute puts mathematical bounds on the potential discrimination of a classifier.

Pre-Processing approaches: Manipulating labels

Massaging the data: The label of a negative sample close to the decision boundary within a disadvantaged group might be changed to positive, and vice versa.

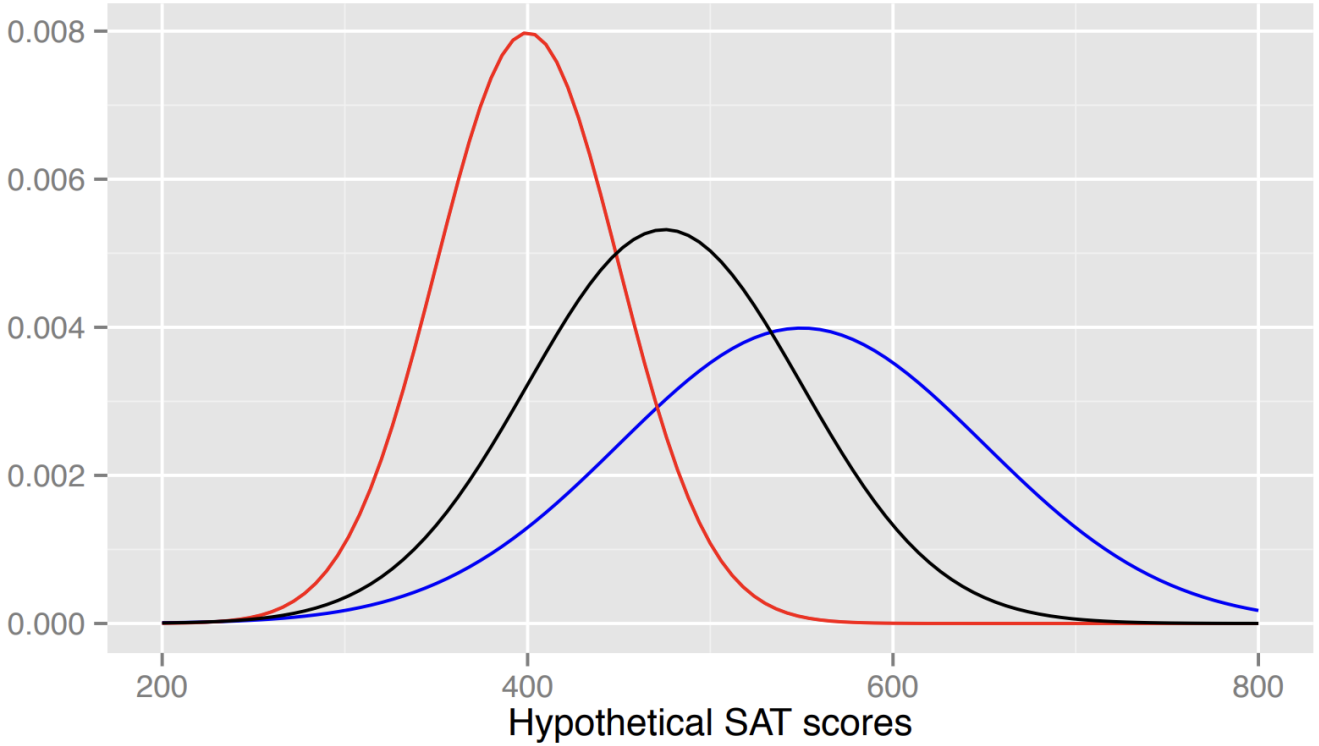
Sampling: This technique adjusts the imbalance in the data by resampling, which includes both oversampling and undersampling, to remove discrimination.



Kamiran & Calders (2012)

Pre-Processing approaches: Manipulating Observed Data

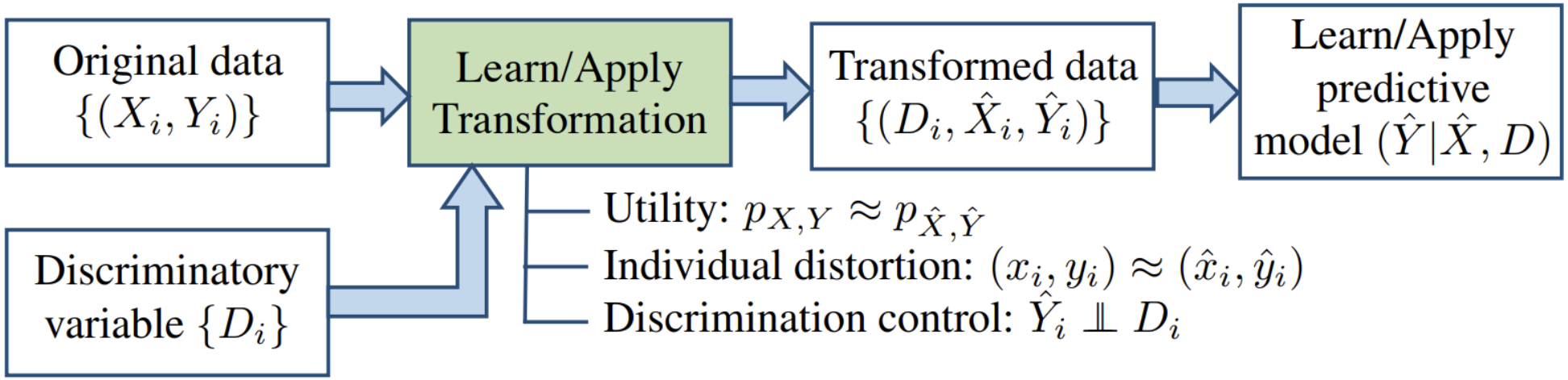
They align the cumulative distributions $\mathrm{F}_0[x]$ and $\mathrm{F}_1[x]$ for feature x when the protected attribute p is 0 and 1 respectively to a median cumulative distribution $\mathrm{F}_m[x]$. This approach has the disadvantage that it treats each input variable $x \in \mathrm{X}$ separately and ignores their interactions.



Feldman *et al.* (2015)

Pre-Processing approaches: Manipulating Labels & Data

Calmon *et al.* (2017) learn a randomized transformation $Pr(x', y'|x, y, p)$ that transforms data pairs $\{x, y\}$ to new data values $\{x', y'\}$ in a way that depends explicitly on the protected attribute p. They formulate this as an optimization problem in which they minimize the change in data utility, subject to limits on the prejudice and distortion of the original values. They show that this optimization problem may be convex in certain conditions.



Contents

In-Processing approaches

In the previous section, we introduced the latent prejudice measure based on the mutual information between the data x and the protected attribute p. Similarly, we can measure the dependence between the labels y and the protected attribute p :

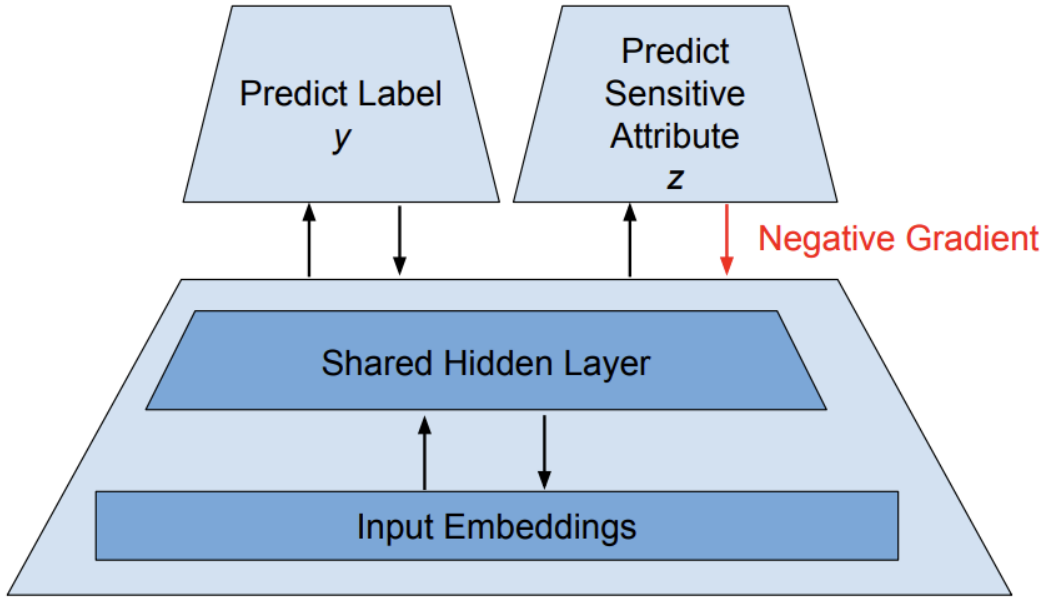
$$\mathrm{IP} = \sum_{y,p} Pr(y,p) \log \left[\frac{Pr(y,p)}{Pr(y)Pr(p)} \right]$$ Kamishima *et al.* 2011

This is known as the ***indirect prejudice***. Intuitively, if there is no way to predict the labels from the protected attribute and vice-versa then there is no scope for bias.

- Adversarial de-biasing
- Prejudice removal by regularization
- Fairness constraints

In-Processing approaches: Adversarial de-biasing

Adversarial-debiasing reduces evidence of protected attributes in predictions by trying to simultaneously fool a second classifier that tries to guess the protected attribute p. By forcing both classifiers to use a shared representation and so minimizing the performance of the adversarial classifier means removing all information about the protected attribute from this representation.



Beutel *et al.* 2017

In-Processing approaches: Regularization

Kamishima et al. proposed adding an **extra regularization condition** to the output of logistic regression classifier that tried to minimize the mutual information between the protected attribute and the prediction $\hat{y}$. They first re-arranged the indirect prejudice expression using the definition of conditional probability to get:

$$\begin{aligned} \mathrm{PI} &= \sum_{y,p} Pr(y|\mathbf{x},p) \log \left[\frac{Pr(y,p)}{Pr(y)Pr(p)} \right] \\ &= \sum_{y,p} Pr(y|\mathbf{x},p) \log \left[\frac{Pr(y|p)}{Pr(y)} \right]. \end{aligned}$$

$$\mathrm{L}_{reg} = \sum_{i} \sum_{\hat{y},p} Pr(\hat{y}_i|\mathbf{x}_i,p_i) \log \left[\frac{Pr(\hat{y}_i|p_i)}{Pr(\hat{y}_i)} \right]$$

Kamishima *et al.* (2011)

In-Processing approaches: Fairness constraints

Zafar *et al.* (2015) formulated unfairness in terms of the covariance between the protected attribute $\{P_i\}_{i=1}^{I}$ and the signed distances $\{d[x_i, \theta]\}_{i=1}^{I}$ of the associated feature vectors $\{x_i\}_{i=1}^{I}$ from the decision boundary, where θ denotes the model parameters. Let $\bar{p}$ represent the mean value of the protected attribute. They then minimize the main loss function such that the covariance remains within some threshold t.

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimize}} \quad & L[\boldsymbol{\theta}] \\ \text{subject to} \quad & \frac{1}{I}\sum_{i=1}^{I}(p_i - \bar{p})d[\mathbf{x}i, \boldsymbol{\theta}] \leq t \\ & \frac{1}{I}\sum i = 1^{I}(p_i - \bar{p})d[\mathbf{x}_i, \boldsymbol{\theta}] \geq -t \end{aligned}$$

Contents

Further Resources

- Zemel *et al.* (2013)
- Chen *et al.* (2018)
- Friedler *et al.* (2019)

- AI Fairness 360 Python library
- Zhao *et al.* 2019 Bias in NLP Embeddings

Bias and Fairness: END

Thank you!

Questions and Opinions are Welcome!