An overview of fairness methods

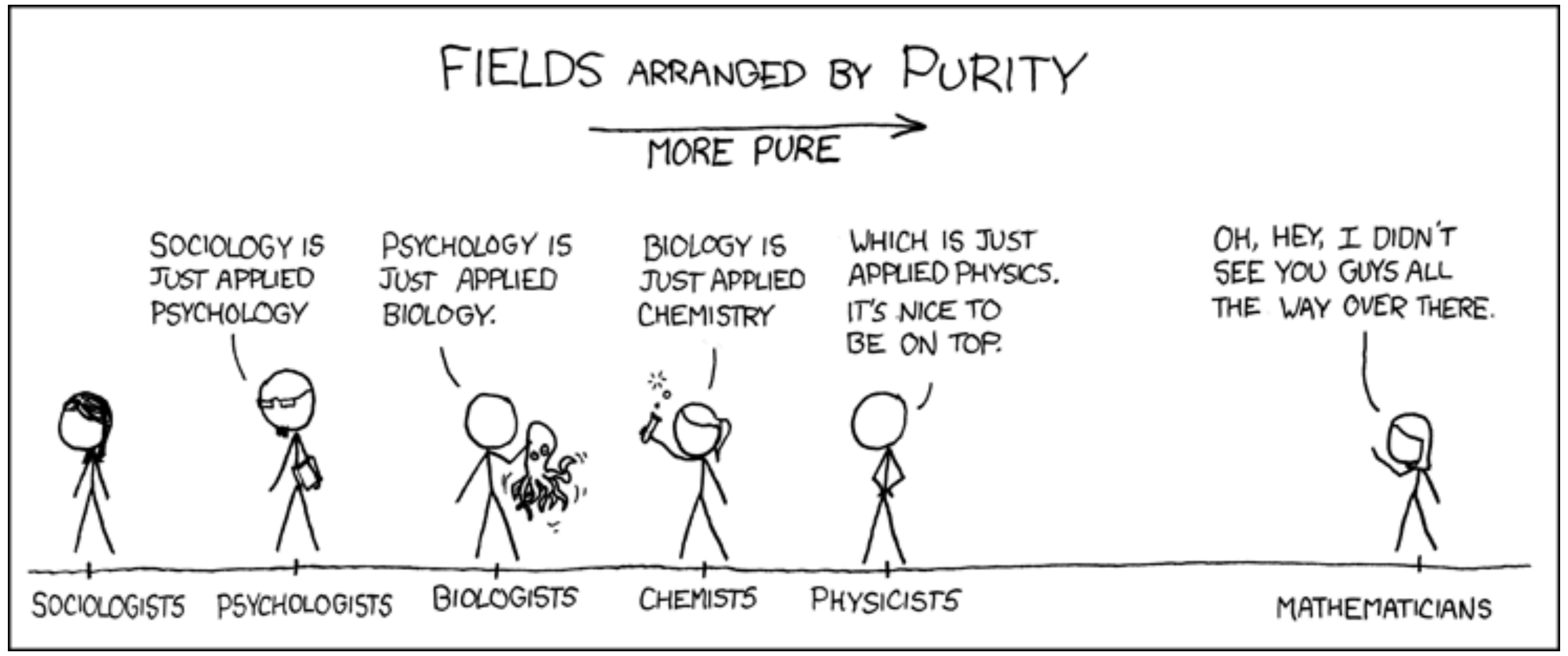
STAT GR5243 Applied Data Science

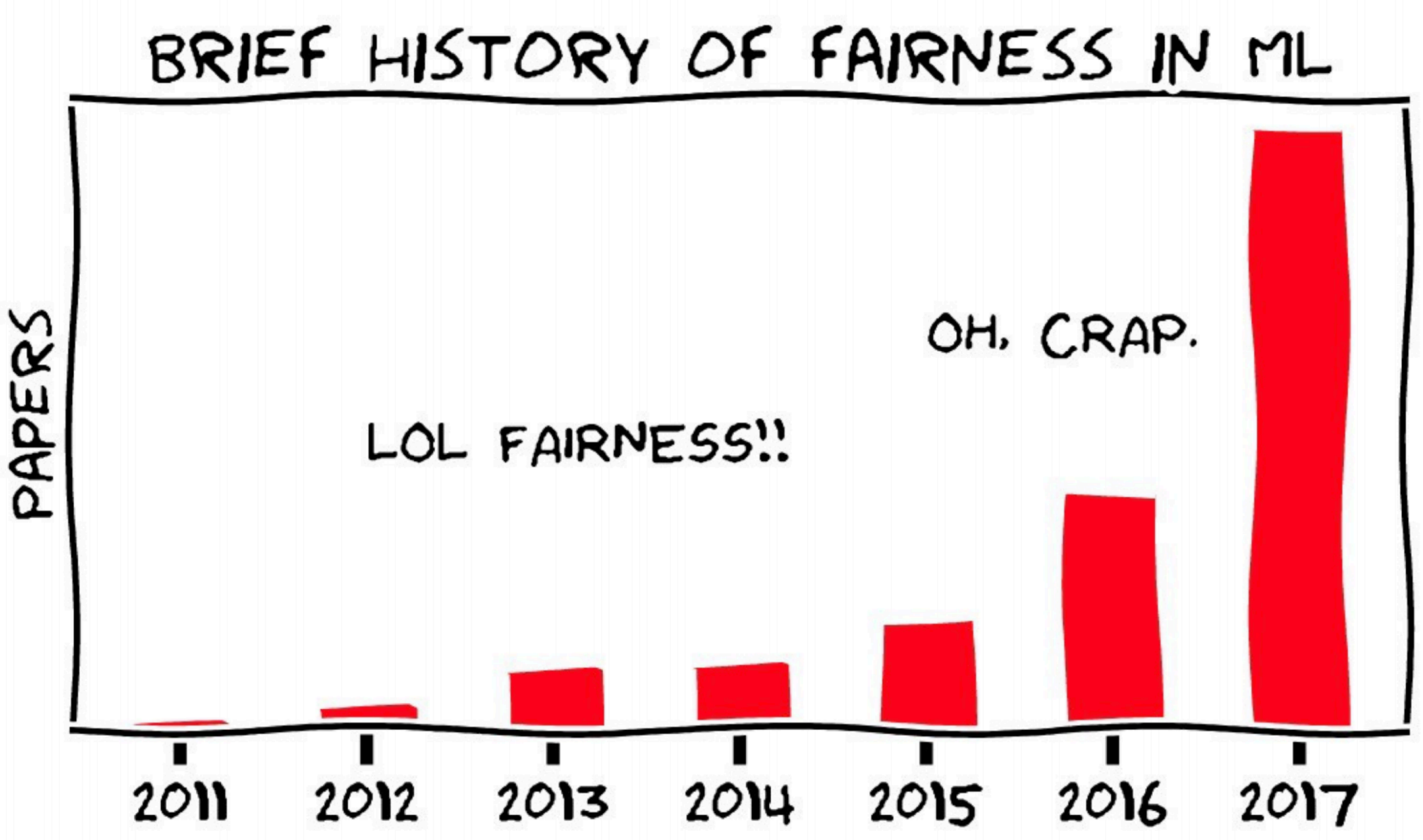
Claire He, Fall 2023

Motivation

Why should we care about fairness in ML?

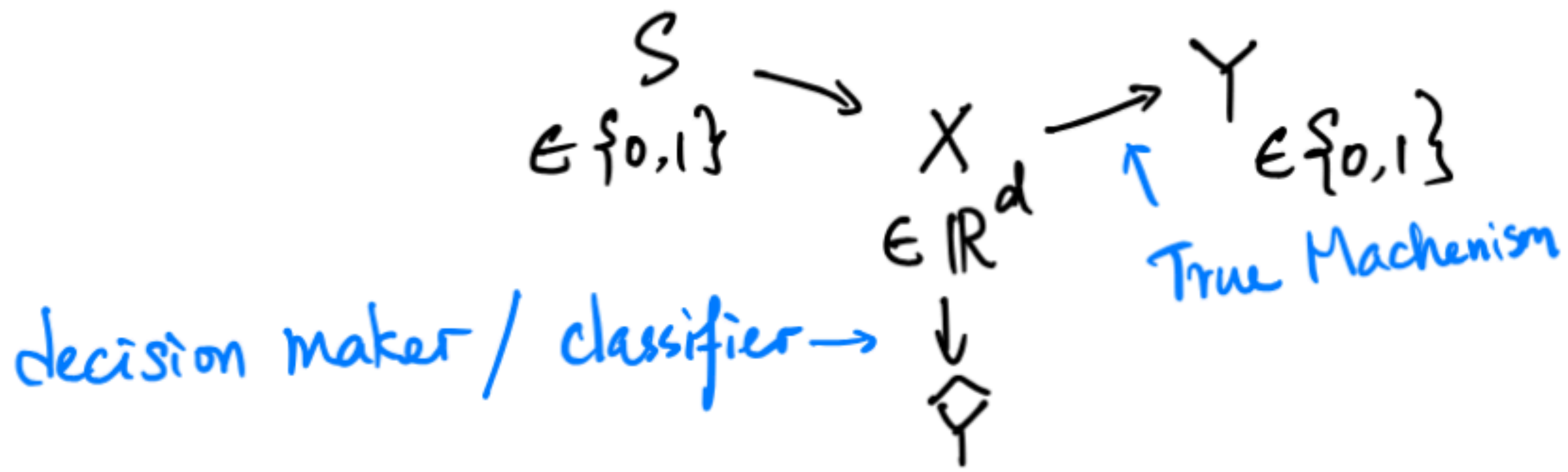
- Philosophical paradigm: science -> objectivity and fairness
- In reality: AI is often a decision-making aiding tool *informed* by domain knowledge/data/engineers/statisticians/data scientist (…)
- Where do we introduce/reproduce **bias, discrimination, …** « **unfairness** »?
- Need for a less confusing definition.





Introduction

What is machine learning fairness? Fish example (classification task)



- Let $Y \in \{0,1\}$ for Bad/Good,

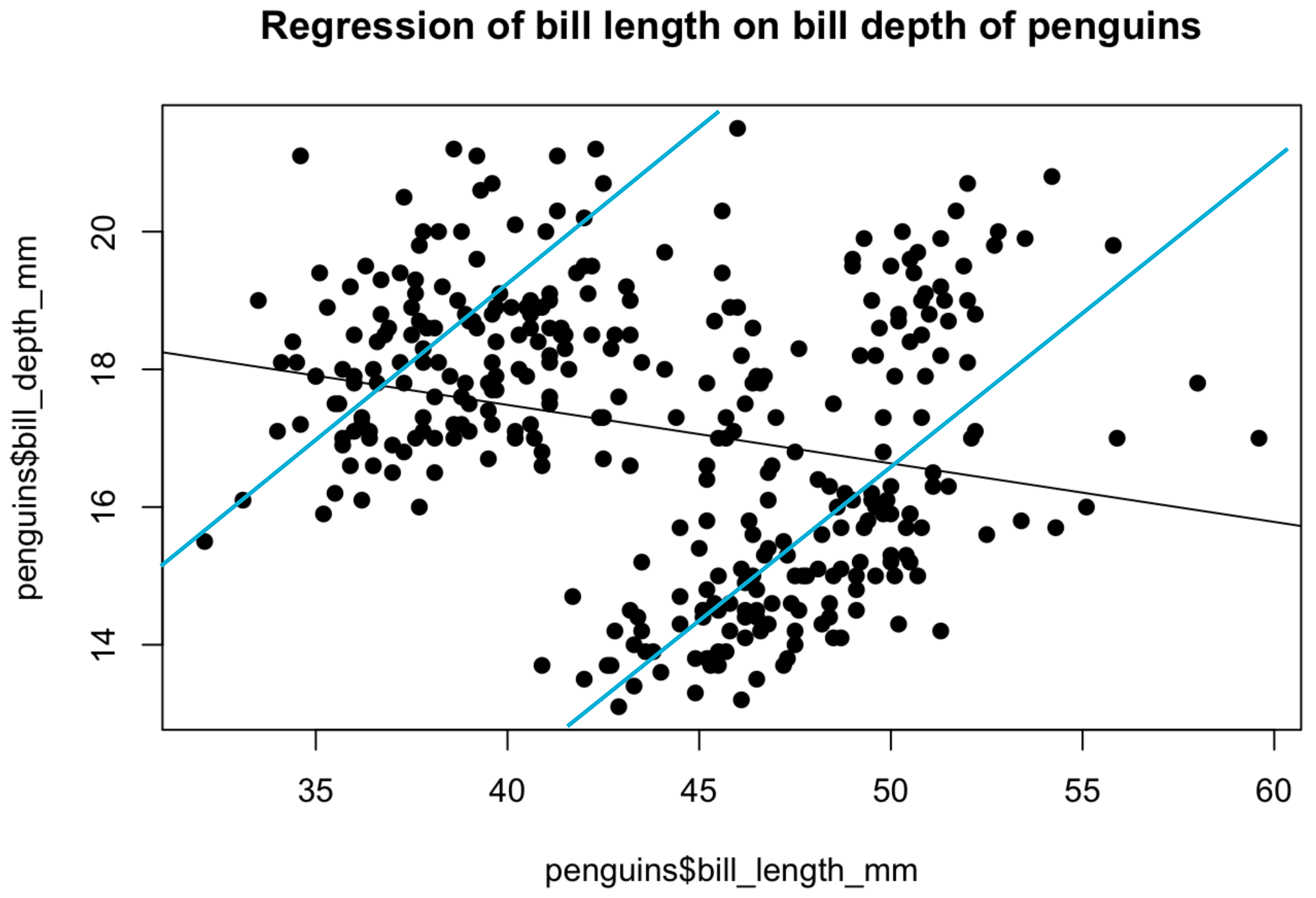
- $X \in \mathbb{R}^d$ our set of features, for Bad/Good classification we can imagine it includes « qualities » of the fish (aggressiveness in the tank to other fishes? Social fish? Small tank fish/big tank fish? …).

- $S \in \{0,1\}$ for blue/red color of the fish.

- We want to predict $\hat{Y}$ by learning a classifier to be as reflective of the true mechanism given features X that we can observe.

Introduction

Conditional probability as a metric?



Simpson's Paradox: $Y \sim X$ versus $Y \sim X + S$

$$y = \beta_0 + \beta_1 x + \epsilon \text{ vs } y = \delta_0 + \delta_1 x + \delta_2 z + \eta$$

$$y \,|\, (z = 0) = \delta_0 + \delta_1 x + \eta$$

$$y \,|\, (z = 1) = \delta_0 + \delta_1 x + \delta_2 + \eta$$ —> account for different Z

In the setting where $y \in \{0,1\}$, $P(Y \,|\, S = s) = f_s(X)$: probability within each population group z.

Recall Bayes rule $P(A, B) = P(A \,|\, B)P(B) \Rightarrow P(Y = y, S = s) = P(Y \,|\, S = s)P(S = s)$

Total probability $P(Y) = \sum_s P(Y \,|\, S = s)P(S = s)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$

B. Equality of odds: $P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y),\ \forall y \in \{0,1\}$

C. Explainable discrimination: $P(\hat{Y} = 1 \mid S = 0, X = x) = P(\hat{Y} = 1 \mid S = 1, X = x),\ \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y} = Y \mid S = 0) = P(\hat{Y} = Y \mid S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y),\ \forall y \in \{0,1\}$

C. Explainable discrimination: $P(\hat{Y} = 1 \mid S = 0, X = x) = P(\hat{Y} = 1 \mid S = 1, X = x),\ \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y} = Y \mid S = 0) = P(\hat{Y} = Y \mid S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y}=1\,|\,S=0) = P(\hat{Y}=1\,|\,S=1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y}=1\,|\,S=0, Y=y) = P(\hat{Y}=1\,|\,S=1, Y=y),\ \forall y \in \{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y}=1\,|\,S=0, X=x) = P(\hat{Y}=1\,|\,S=1, X=x),\ \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y}=Y\,|\,S=0) = P(\hat{Y}=Y\,|\,S=1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 | S = 0) = P(\hat{Y} = 1 | S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 | S = 0, Y = y) = P(\hat{Y} = 1 | S = 1, Y = y),\ \forall y \in \{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y} = 1 | S = 0, X = x) = P(\hat{Y} = 1 | S = 1, X = x),\ \forall x \in \mathbb{R}^d$ the probability of predicting the fish as good is the same regardless of color given the same observed features

D. Calibration: $P(\hat{Y} = Y | S = 0) = P(\hat{Y} = Y | S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \,|\, S = 0) = P(\hat{Y} = 1 \,|\, S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 \,|\, S = 0, Y = y) = P(\hat{Y} = 1 \,|\, S = 1, Y = y), \ \forall y \in \{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y} = 1 \,|\, S = 0, X = x) = P(\hat{Y} = 1 \,|\, S = 1, X = x), \ \forall x \in \mathbb{R}^d$ the probability of predicting the fish as good is the same regardless of color given the same observed features

D. Calibration: $P(\hat{Y} = Y \,|\, S = 0) = P(\hat{Y} = Y \,|\, S = 1)$ the probability of correct classification is the same regardless of the color

Introduction

The Impossibility Theorem

Kleinberg et al. (2016) showed that A, B and D (parity, equalized odds and calibration) can **not** be jointly optimized.

This means we will have to carefully choose and specify our metrics of fairness and that any AI system we build will necessarily violate some notion of fairness.

Our 4 papers introduce frameworks that aim for ensuring some level of **fairness** in ML tasks through different layers of ML workflow.

1. What is the fairness framework?

2. Where is the fairness introduced in the workflow?

ML fairness methods

An overview of some approaches to fairness

1. Pre-processing methods: modify training data
    A. Local massaging: relabeling points near the boundary
    B. Local preferential resampling: resample points close to the boundary
2. In-processing methods: modify the learning algorithm
    C. Through cost functions/constraints (regularization)
    D. Through the pipeline : adding a latent representation
    E. Through feature selection
3. Post-processing methods: modify the prediction outcome
4. Causal reasoning

Learning Fair Representations

Paper 1

- Fairness framework: **group/individual fairness**
  - Group: the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole
  - Individual: similar individuals should be treated similarly
  - Fairness metric: $P(Z = k \mid X, S = 0) = P(Z = k \mid X, S = 1)$
- Method: (2D) learning a latent representation (think dimension reduction methods like PCA)
  - X features, $S \in \{0,1\}$ protected set
  - $Z \sim Mult(n, v)$: with K « prototypes » associated to $(v_k)_{k=1,\ldots,K}$
    $X \in \mathscr{X} \longrightarrow Z \in \{1,\ldots,K\} \longrightarrow Y \in \{0,1\}$

Learning Fair Representations

Paper 1

$X \in \mathcal{X} \longrightarrow Z \in \{1,\ldots,K\} \longrightarrow Y \in \{0,1\}$

Idea:

- X informative but correlated with S —> discrimination
- Find an intermediate Z that keeps information, but is less correlated with S by adding an unfairness loss that ensures « parity » $P(Z=k \mid X, S=0) = P(Z=k \mid X, S=1)$ —> fair attribution of the prototypes
- Minimize simultaneously **reconstruction loss** $L_X = \lvert\lvert X - \hat{X} \rvert\rvert_2$ where $\hat{X} = f(Z)$**, cross entropy (classification) and unfairness loss** $L_Z = \sum_k \lvert P(Z=k \mid S=0) - P(Z=k \mid S=1) \rvert$

Fairness constraints

Paper 2

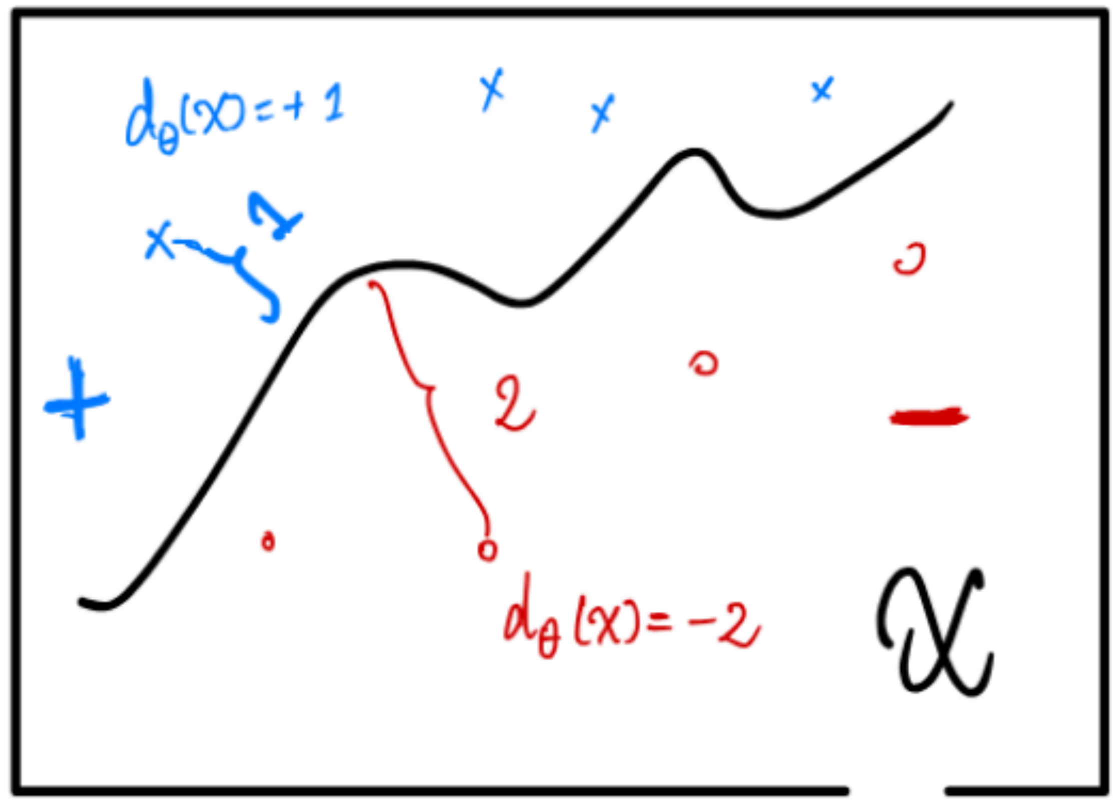
- Fairness framework: **Disparate treatment/impact**
  - **DT:** The decisions are (partly) based on the individual's sensitive attribute
  - **DI**: its outcomes disproportionately hurt (or, benefit) people with certain sensitive attribute values
- Method: (2C) modify the cost functions of convex margin-based clfs: penalty term for being « unfair »
  - $D = (X, Y, S)$ dataset
  - $L_\theta(D)$ classification loss (cross entropy f.e.)
  - $R_\theta(D)$ a measure of unfairness

Fairness constraints

Paper 2

- **DT:** The decisions are (partly) based on the individual's sensitive attribute —> don't use the sensitive attribute when making decisions ? —> use a s free loss $L_\theta(D) = f(Y|X, \theta)$

- **DI**: its outcomes disproportionately hurt (or, benefit) people with certain sensitive attribute values —> taking out sensitive attribute does not solve bias in training set entirely (indirect discrimination)… —> $R_\theta(D) = g(X, S, \theta)$



Define $R_\theta(D) = |Cov(s, d_\theta(x))|$ with the signed distance to decision boundary to quantify unfairness.

Note: $\hat{Y} \perp S \Rightarrow Cov(\hat{Y}, S) = 0 \Rightarrow P(\hat{Y} = 1 | S = 0) = P(\hat{Y} = 1 | S = 1)$

Aims to fullfil a sufficient condition on parity.

Fairness constraints

Paper 2

Convex margin-based classifier formulation

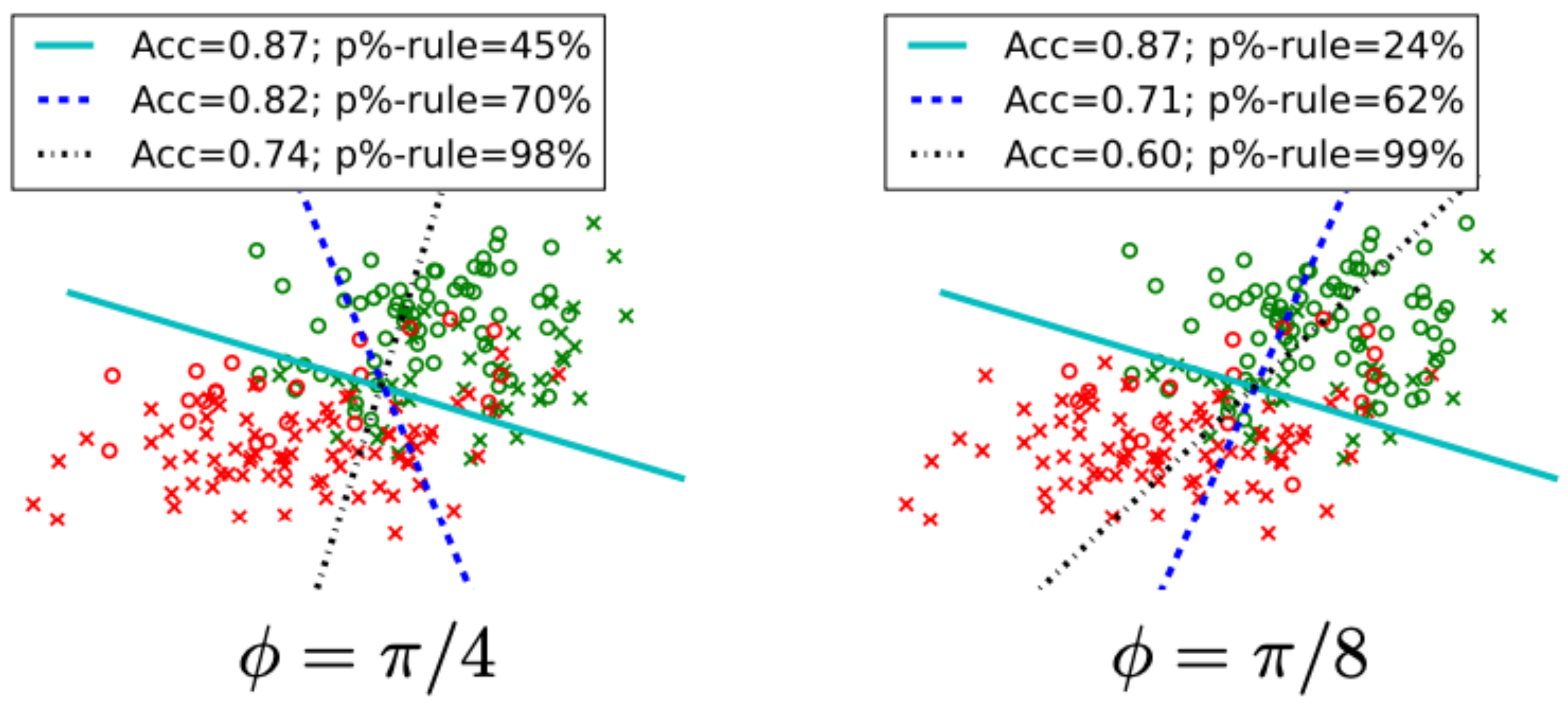
- Maximise accuracy under fairness constraint:

$$\min_{\theta} L_{\theta}(D) \text{ s.t. } R_{\theta}(D) \leq \tau$$

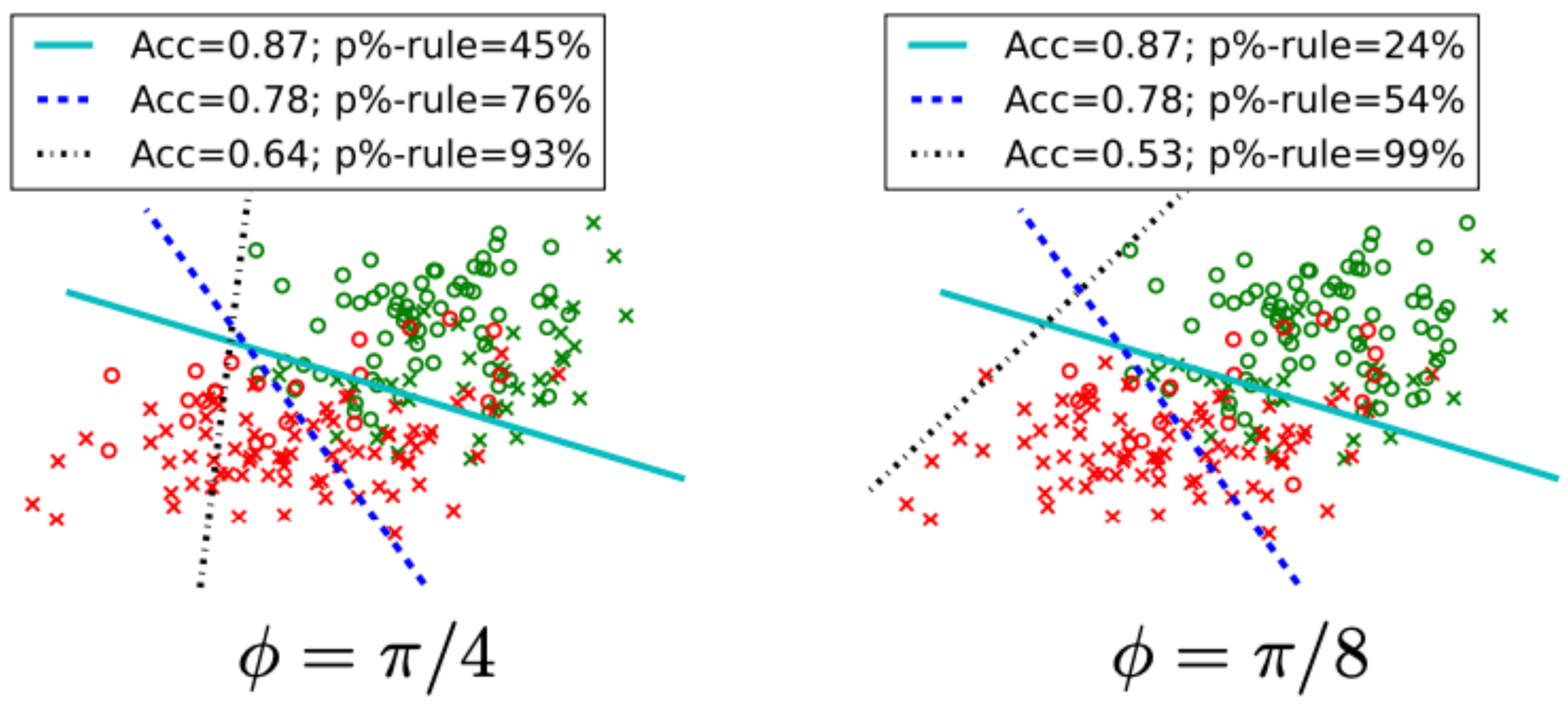
- Maximise fairness under accuracy constraint:

$$\min R_{\theta}(D) \text{ s.t. } L(\theta) \leq (1+\gamma)L(\theta^*)$$

Method applied to Logistic Regression and SVM (appendix)



(a) Maximizing accuracy under fairness constraints



(b) Maximizing fairness under accuracy constraints

Learning without Disparate Mistreatment

Paper 3

- Fairness framework: **disparate treatment, mistreatment, impact**
  - No disparate treatment: $P(\hat{y} \mid x, s) = P(\hat{y} \mid x)$
  - No disparate impact: $P(\hat{y} = 1 \mid s = 0) = P(\hat{y} = 1 \mid s = 1)$
  - No disparate **mistreatment:** if the misclassification rates for different groups of people having different values of the sensitive feature s are the same.
- Methods: (2C)
- Extension builds on the framework from the previous model, we use a continuous version of $Cov(S, \hat{Y}) \rightarrow Cov(s, g_\theta(y, X))$ where we choose g_θ to be some signed distance between misclassified users' feature vectors to the boundary.

Learning without Disparate Mistreatment

Paper 3

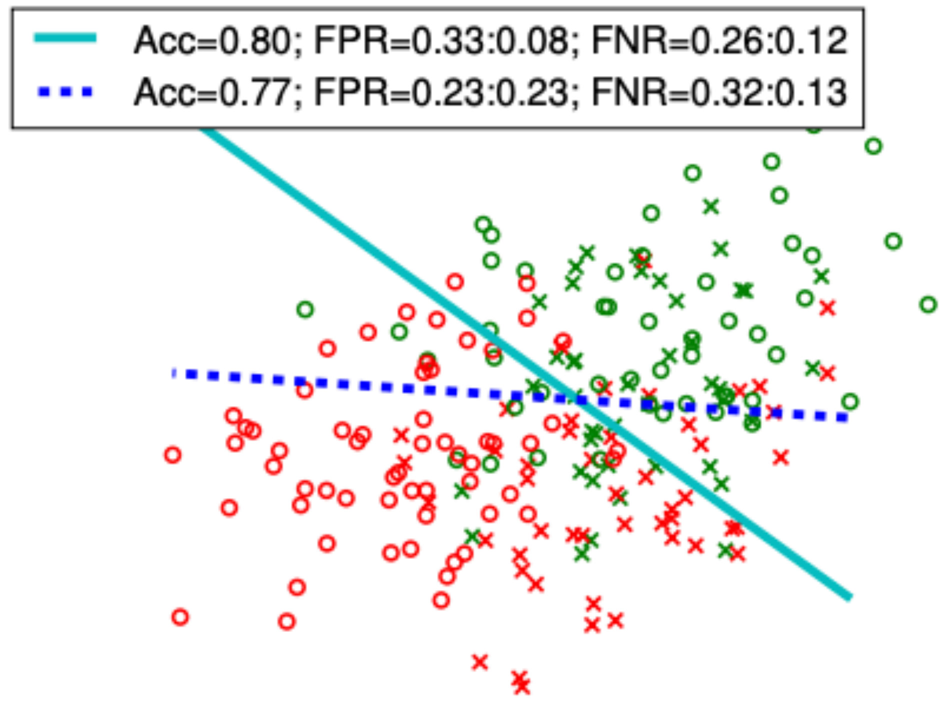
Optimization based classification method:

$$\min L_\theta(D) \ \text{ s.t. } \ M(D) < \epsilon$$

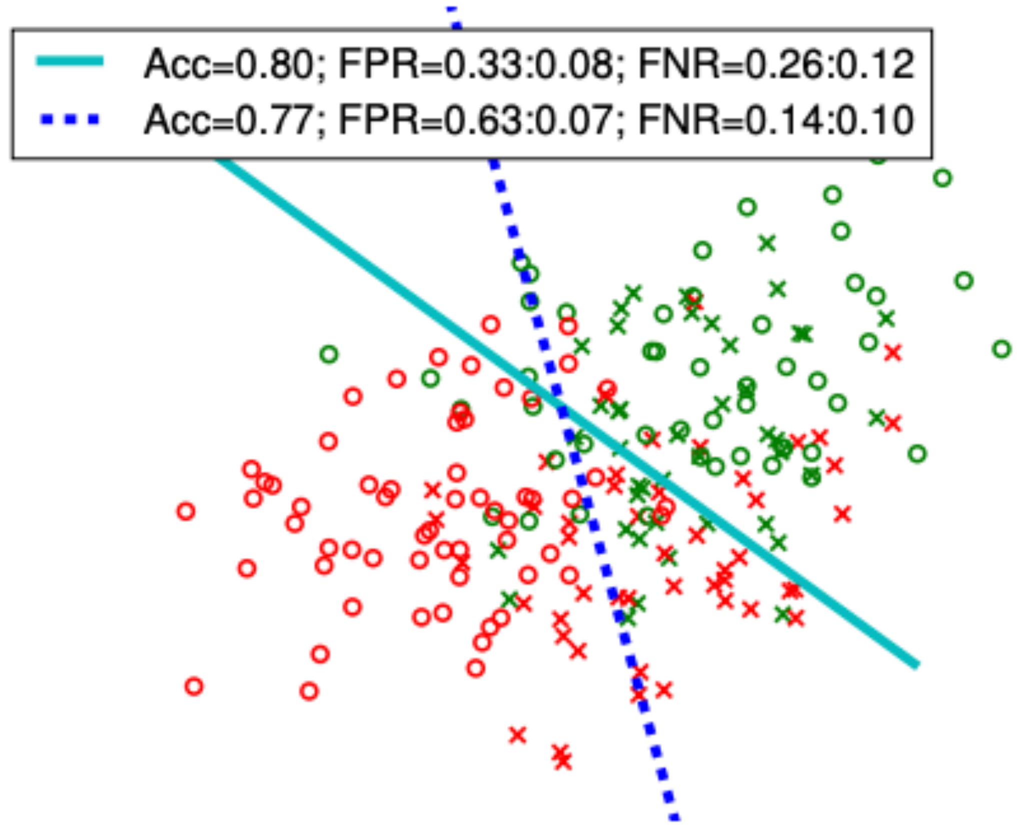
where $M(D) = Cov(s, g_\theta(y, X))$ is some metric of misclassification that brings unfairness.

For example

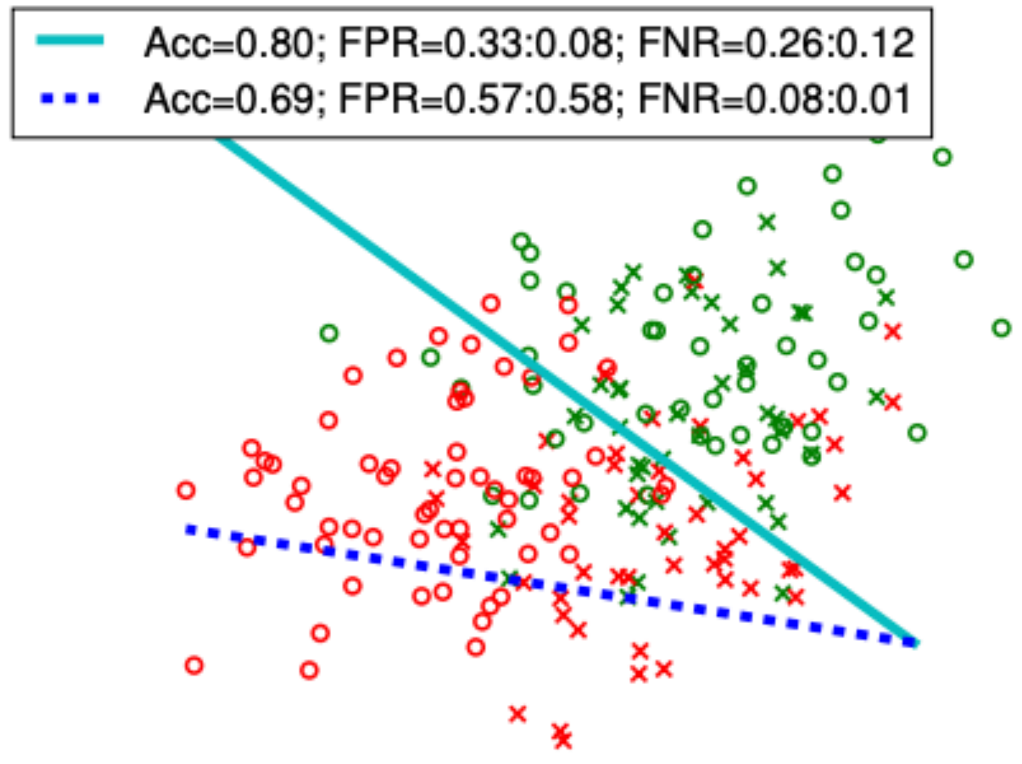
$$g_\theta(y,x) = \begin{cases} 0 \wedge y d_\theta(x) & \text{if control overall missclassification} \\ \dfrac{1-y}{2} y d_\theta(x) & \text{if control FPR} \\ \dfrac{1+y}{2} y d_\theta(x) & \text{if control FNR} \end{cases}$$



(a) FPR constraints



(b) FNR constraints



(c) Both constraints

Fairness with prejudice remove regularizer

Paper 5

- Fairness framework
  - Minimise the amount of « prejudice » through the sensitive information S shared in Y through
- Method: Prejudice remover regularizer (2C)
  - $\min L_\theta(D) + \eta R_\theta(D)$
  - Set $R_\theta(D) = \sum_{Y,S} \hat{P}(Y,S) \log \dfrac{\hat{P}(Y,S)}{\hat{P}(S)\hat{P}(Y)}$ as the Prejudice Index
  - constraint from information theory: recognize mutual information (see later slides)

Fairness-aware feature selection

Paper 4

- Fairness framework:
  - **group fairness/individual fairness**
  - **equalised odds and parity assumptions**
- Methods: (2E + 4)
  - Information theoretical metrics for feature selection
  - Graphical causal models

Fairness aware feature selection

Paper 4

- Information theoretic framework
    - UI: unique information
    - SI: shared information
    - CI: « synergetic » information (recoverable with both variables)
    - I: mutual information
- Construct measures with specific properties on our sets
- Two targets:
    - Accuracy of prediction
    - Discriminatory impact

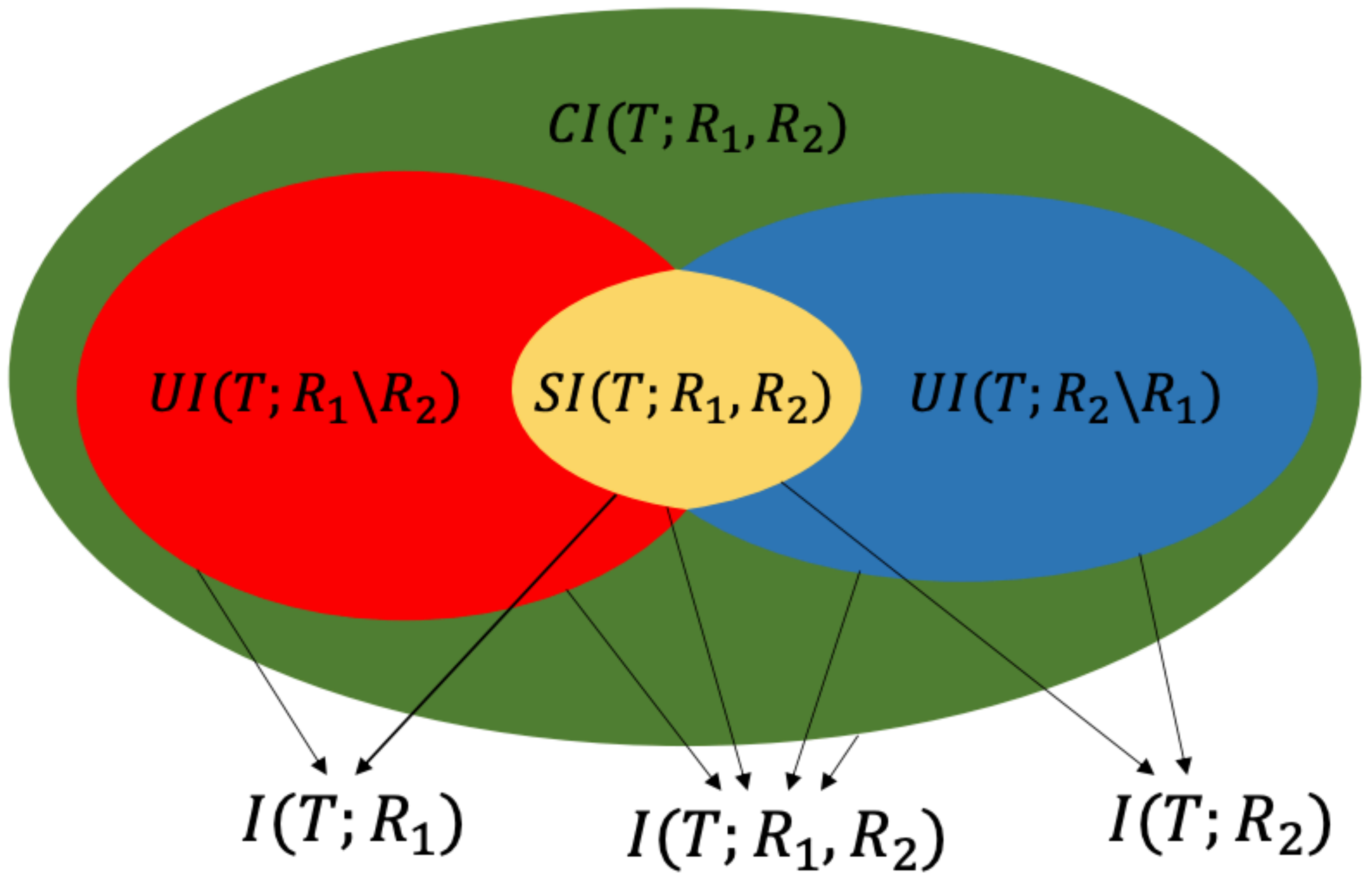


Figure 1: Decomposition of Information.

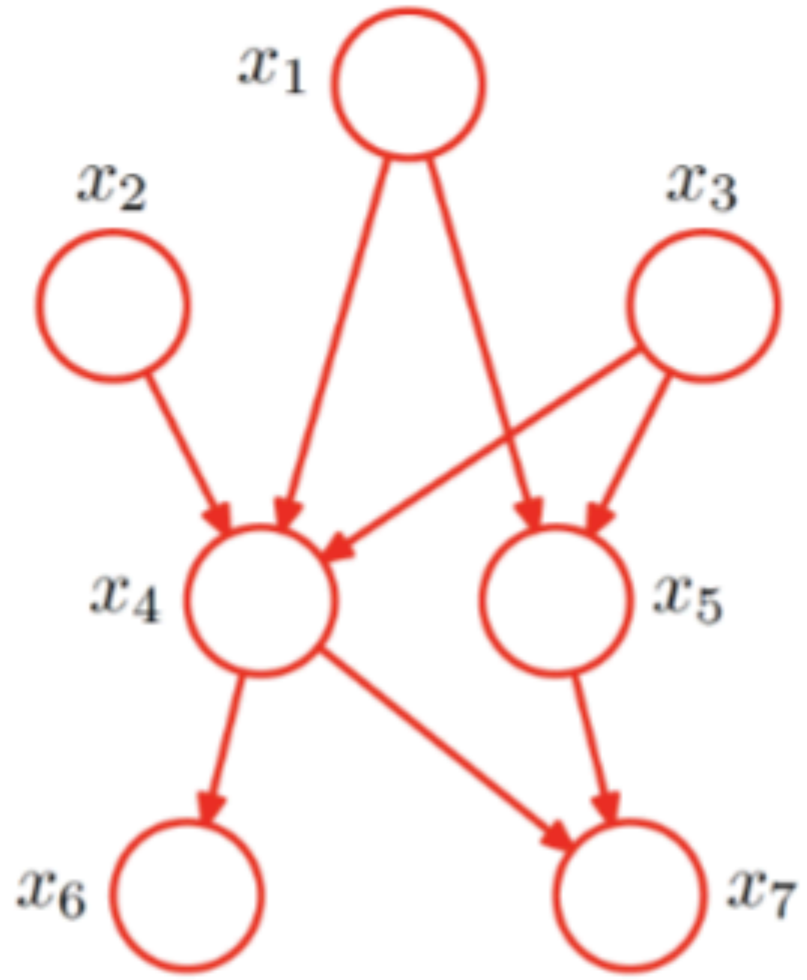
Fairness aware feature selection

Paper 4

$$\begin{array}{ccccc} A & \rightarrow & X^n & \rightarrow & Y \\ & & \downarrow & & \\ & & \hat{Y} & & \end{array}$$

Graphical model (causal)

- X^n parent of Y, $\hat{Y}$
- Y child of X^n
- A parent of X^n
- $A \perp \hat{Y}, Y \mid X^n$



Joint distribution $p(x_1, x_2, \ldots, x_7)$ equals

$$p(x_1) \cdot p(x_2) \cdot p(x_3) \cdot p(x_4|x_1, x_2, x_3) \cdot p(x_5|x_1, x_3) \cdot p(x_6|x_4) \cdot p(x_7|x_4, x_5)$$

Generally,

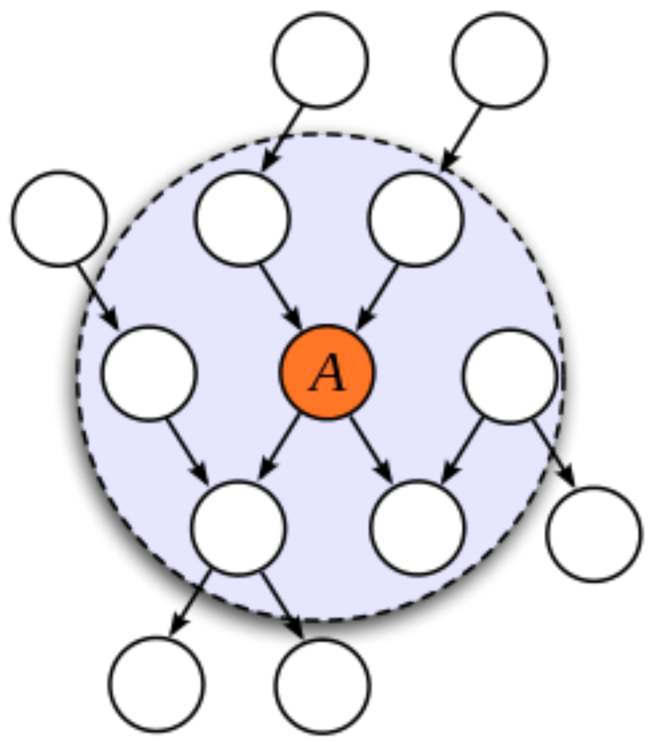
$$p(x_1, \ldots, x_K) = \prod_{k=1}^{K} p(x_k \mid \mathrm{pa}_k)$$

where pa_k denotes the set of parents of x_k

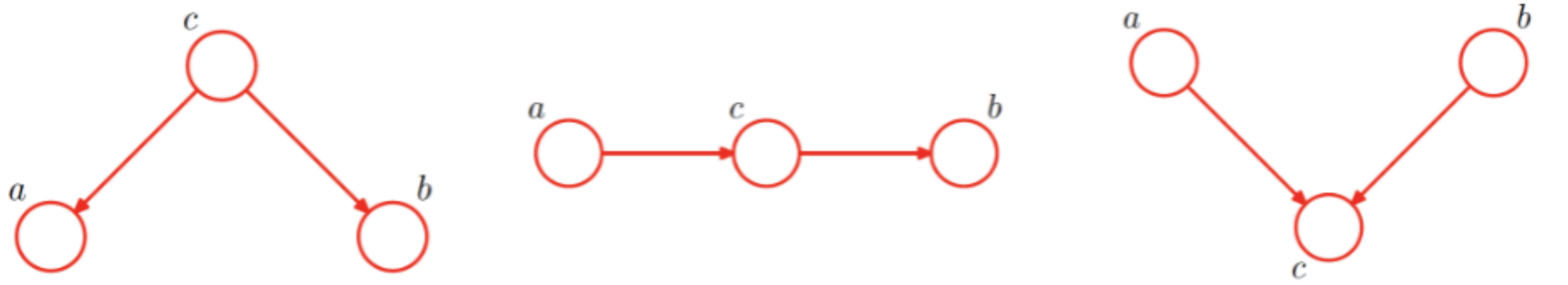
Fairness aware feature selection

Paper 4

- **Markov blanket** of a node x_i: set of **parents, children, co-parents of the node.**
- **Conditional distribution** of x_i | Markov blanket is independent of the rest of the graph



- Conditional independence relations can be conveniently represented by graphs
- Three graphs between three random variables a, b, c



1st graph: $a \not\perp\!\!\!\perp b$ marginally, but $a \perp\!\!\!\perp b \mid c$ (tail-to-tail)

2nd graph: $a \not\perp\!\!\!\perp b$ marginally, but $a \perp\!\!\!\perp b \mid c$ (head-to-tail)

3rd graph: $a \perp\!\!\!\perp b$ marginally, but $a \not\perp\!\!\!\perp b \mid c$ (head-to-head, "explaining away")

Fairness aware feature selection

Paper 4

Accuracy measure for a subset of features $X_S \subset X_{[n]}$ $\nu^{Acc}(X_S)$ has to satisfy

- **Non negativity** $\nu^{Acc} \geq 0$
- **Monotonicity** $S_1 \subset S_2 \Rightarrow \nu^{Acc}(X_{S_1}) \leq \nu^{Acc}(X_{S_2})$ adding feature does not decrease accuracy
- **Blocking** $Y \perp X_S \,|\, \{A, X_{S^c}\} \Leftrightarrow \nu^{Acc}(X_S) = 0$ accuracy measure should be non zero on Y's Markov blanket, and zero on remaining features

Then they define $\nu^{Acc}(X_S) = I(Y; X_S \,|\, \{A, X_{S^c}\})$ —> protect A, X_{S^c} being the attribute and sensitive features directly from the attribute.

Fairness aware feature selection
Paper 4

Discriminatory measure for a subset of features $X_S \subset X_{[n]}$ $\nu^{Acc}(X_S)$ has to satisfy

- **Non negativity** $\nu^D \geq 0$
- **Monotonicity** $S_1 \subset S_2 \Rightarrow \nu^D(X_{S_1}) \leq \nu^D(X_{S_2})$ adding feature does not decrease accuracy
- **Independences**
    - $Y \perp X_S \Rightarrow \nu^D(X_S) = 0$: sensitive features irrelevant to classification are not discriminatory
    - $A \perp X_S \Rightarrow \nu^D(X_S) = 0$: sensitive features not a proxy for protected attributes are not discriminatory
    - $A \perp X_S \,|\, Y \Rightarrow \nu^D(X_S) = 0$

Defined as $\nu^D(X_S) = SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A \,|\, Y)$: discriminatory in the sense that information is shared with the protected attribute!

Fairness aware feature selection

Paper 4

Measures are defined but not used as such: we need to take in account the aggregate effect of all subsets of features that include a certain feature:

Shapley function $\phi_i(v) \propto \sum_T \nu(T \cup \{i\}) - \nu(T)$ where we input $\nu \in \{\nu^{Acc}, \nu^D\}$

$$\varphi_i(v) = \frac{1}{\text{number of players}} \sum_{\text{coalitions including } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions excluding } i \text{ of this size}}$$

Define a **fairness utility score** for each feature $\mathcal{F}_i = \phi_i(\nu^{Acc}) - \alpha\phi_i(\nu^D)$

—> Strike a balance between accuracy and fairness

Handling Conditional Discrimination

Paper 6

- Fairness Framework:
    - **Explainable discrimination**
      $P(Y = 1 \mid S = 1, X = x) = P(Y = 1 \mid S = 0, X = x), \; \forall x$
- Methods (1A + 1B): debiasing training data
    - Local Massaging: relabel data close to the decision boundary
    - Local Preferential Sampling: remove current samples and resample close to the decision boundary
    - Why close to decision boundary? Remember SVM: « support vectors »…

Handling Conditional Discrimination

Paper 6

Algorithm 1: Local massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  rank males using $\mathcal{H}_i$;
  `relabel DELTA` (male) males that are the closest to the decision boundary from $+$ to $-$ (Algorithm 4);
  rank females using $\mathcal{H}_i$;
  `relabel DELTA` (female) females that are the closest to the decision boundary from $-$ to $+$
end

Algorithm 2: Local preferential sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: resampled dataset (a list of instances)

`PARTITION` $(\mathbf{X}, \mathbf{e})$ (see Algorithm 3);
for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  `rank` males using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (male) (see Algorithm 4) males $+$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (male) males $-$ that are the closest to the decision boundary;
  `rank` females using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (female) females $-$ that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (female) females $+$ that are the closest to the decision boundary;
end