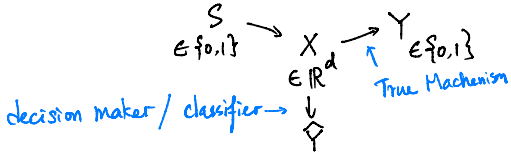
Methods for Machine Learning Fairness.

Recall:

- What is machine learning fairness?



Eg. What are $S, X, Y, \hat{Y}$ for the good/bad fish example?

S: a dummy variable indicating the color of the fish

X: "qualities" of a fish that determines it being good/bad

Y: a dummy variable that indicate whether a fish is good/bad

$\hat{Y}$: the prediction on whether a fish is good/bad.

Is the algorithm/classifier fair?

① $\mathbb{P}(\hat{Y}=1 \mid S=0) = \mathbb{P}(\hat{Y}=1 \mid S=1)$ parity

② $\mathbb{P}(\hat{Y}=1 \mid S=0, Y=y) = \mathbb{P}(\hat{Y}=1 \mid S=1, Y=y)$. $\forall y \in \{0,1\}$ equality of odds.

③ $\mathbb{P}(\hat{Y}=1 \mid S=0, X=x) = \mathbb{P}(\hat{Y}=1 \mid S=1, X=x)$, $\forall x \in \mathcal{X}$. explainable discrimination

④ $\mathbb{P}(\hat{Y}=Y \mid S=0) = \mathbb{P}(\hat{Y}=Y \mid S=1)$ Calibration.

ML Fairness Approaches:

- Pre-processing methods: modify training data (A6)
- In-processing methods: modify the learning algorithm (A1) – (A5), (A7).
- Post-processing methods: modify the prediction outcome.
- Causal reasoning methods.

I. In-processing Methods.

(i) Modify the cost functions/ constraints.

- Add a penalty / regularization term for being "unfair" / Add constraints to the loss minimization.

  * $\min_{\theta \in \Theta} L_\theta(D) + \lambda R_\theta(D)$ (A5)
  * $\min_{\theta \in \Theta} L_\theta(D)$ s.t. $R_\theta(D) \leq \tau$ (A2) (A4)
  * $\min_{\theta \in \Theta} R_\theta(D)$ s.t. $L_\theta(D) \leq \tau$. (A3)

where $D = (X, Y, S)$ is the dataset.

$L_\theta(D)$ is the classification loss, e.g. CE

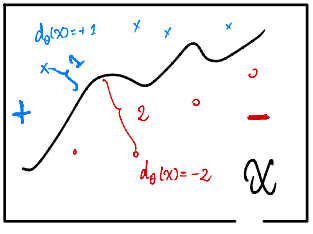
$R_\theta(D)$ is a measure of unfairness.

- How to quantify unfairness ?

a) Decision Boundary Covariance:

$$R_\theta = \left| \widehat{\mathrm{cov}}(S, d_\theta(X)) \right| \quad \text{(A2) (A3)}$$

$d_\theta(X)$ is the signed distance to decision boundary:



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

$\hat{Y} \perp\!\!\!\perp S \Rightarrow \mathrm{cov}(\hat{Y}, S) = 0$

A continuous version of $\mathrm{cov}(S, \hat{Y})$.

A more general version: $\mathrm{cov}(Z, g_\theta(Y, X))$

where $g_\theta(y, x) = 0 \wedge y d_\theta(x)$

or $\frac{1-y}{2} y\, d_\theta(x)$

or $\frac{1+y}{2} y\, d_\theta(x)$. (A4)

(b). Prejudice Index:

$$R_\theta := \sum_{Y,S} P(\hat{Y}, S) \log \frac{P(\hat{Y}, S)}{P(S)\, P(\hat{Y})} \approx \mathrm{MI}(\hat{Y}, S). \quad \text{(A5)}$$

In information theory, mutual information measures the amount of information two random variables share with each other.

(ii). Modify the learning pipeline

Learning Fair Representation (AI)

$$X \in \mathcal{X} \longmapsto Z \in \{1, \cdots, K\} \longmapsto Y \in \{0, 1\}$$

Basic idea.

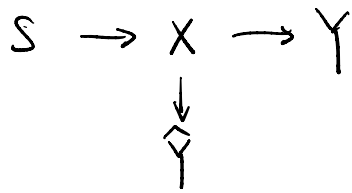
- The original features X:
  - contains useful info for classification
  - correlated with S, potentially leads to discrimination.
- Learn an intermediate representation Z that
  - keeps the useful bit of X and
  - discard the part that causes discrimination.
- Use Z to do the final classification.

The obj. function to minimize then contains three parts:

- classification loss, e.g. cross entropy
- representation loss. i.e. the loss of info by compressing.
- Unfairness loss: $\sum_{k=1}^{K} | P(Z=k|S=0) - P(Z=k|S=1) |$

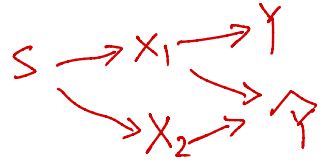
Feature Selection (A7).

Go back to the old model:



Question: Is "hiding" S from the classifier sufficient to ensure fairness?

YES, if the X that has causal effects on Y can be correctly identified

No, if  $X = (X_1, X_2)$.

+ features $\Rightarrow$ { + accuracy, − fairness

Aim: to find a best subset of X that strikes a balance between accuracy and fairness.

Again, it is about how to quantify them.

For any index set $I \in \{1, \cdots, n\}$, $X_I \doteq (X_i : i \in I)$. (# features)

$I = \{1, 3, 5\}$, $X_I = (X_1, X_3, X_5)$

- $v^{acc}(X_I)$ measures how accurate the prediction could be using X_I.
- $v^{D}(X_I)$ measures the discrimination level of using X_I for prediction.

What are "good" properties for accuracy and fairness metrics?

— Axioms that good metrics should satisfy.

- Axioms for accuracy metric v^{acc}:
  - Non-negativity: $v^{acc}(X_I) \geq 0, \quad \forall\, I \subset \{1, \cdots, n\}$
  - Monotonicity: $I_1 \subset I_2 \Rightarrow v^{acc}(X_{I_1}) \leq v^{acc}(X_{I_2})$
  - Blocking: $Y \perp\!\!\!\perp X_I \mid \{A, X_{I^c}\} \Rightarrow v^{acc}(X_I) = 0.$

- Axioms for discrimination metric v^D:
  - Non-negativity: $v^D(X_I) \geq 0, \quad \forall\, I \subset \{1, \cdots, n\}$
  - Monotonicity: $I_1 \subset I_2 \Rightarrow v^D(X_{I_1}) \leq v^D(X_{I_2})$
  - Y-independence: $Y \perp\!\!\!\perp X_I \Rightarrow v^D(X_I) = 0$
  - S-independence: $S \perp\!\!\!\perp X_I \Rightarrow v^D(X_I) = 0$
  - SY-independence: $S \perp\!\!\!\perp X_I \mid Y \Rightarrow v^D(X_I) = 0.$

II. Pre-processing Methods.

Handling Conditional Discrimination (Ab).

Recall that:

$$\mathbb{P}(Y=1 \mid S=1, X=x) = \mathbb{P}(Y=1 \mid S=0, X=x) \quad \cdots (*)$$

is a desired property for a fair dataset.

However, for many cases, (*) does not always hold for the training set.

The paper proposes two methods to manipulate the training data, to achieve the "de-biasing" purpose and make (*) holds.

- Local Massaging:

  relabel the data points close to the decision boundary

- Local Preferential Sampling.

  remove current samples and resample close to the decision boundary.

Why close to decision boundaries?

Eg. (Admission example – local massaging)