Fair Multiple Decision Making Through Soft Interventions

Yaowei Hu[†], Yongkai Wu[‡], Lu Zhang[†], Xintao Wu[†]

[†]University of Arkansas, [‡]Clemson University

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Background

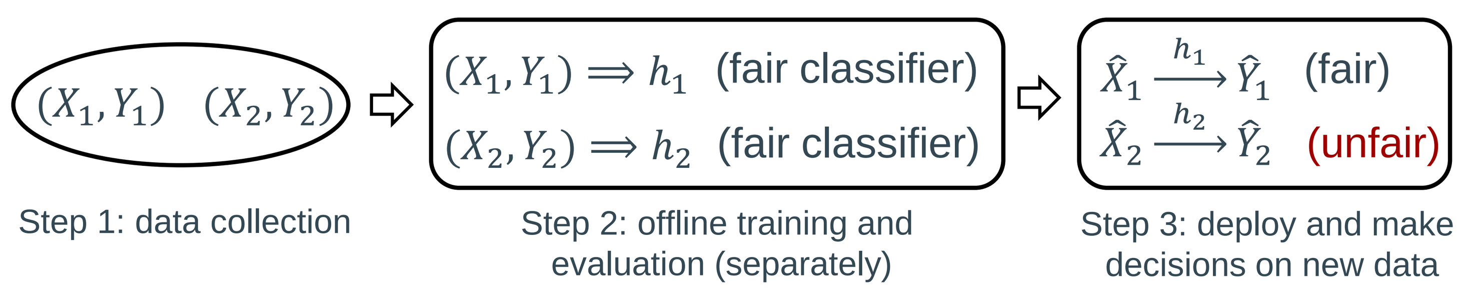
- How to ensure fairness in algorithmic decision making models is an important task in machine learning.
- Most of the previous research focuses on a single decision model, but in reality there may exist multiple decision models.
- All decision models may contain discrimination, either be introduced by themselves or transmitted from upstream models.

Objective & Challenge

- Build fair models for all decision making tasks.
- Difficult even if we know how to build a fair model for each task as data distribution can change as a consequence of deploying new models.

Toy Example

- Consider an intuitive method which builds the fair model for each task independently.



Why unfair?

- Decision $\hat{Y}_1$ will affect values of $\hat{X}_2$
- Distribution X_2 $\neq$ distribution $\hat{X}_2$

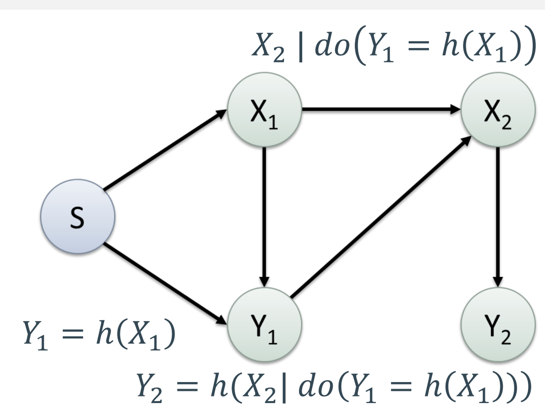
Preliminaries

Structural Causal Model (SEM)

$$S = f_s(U_s)$$

$$X_1 = f_{X_1}(S, U_{X_1}) \quad Y_1 = f_{Y_1}(S, X_1, U_{Y_1})$$

$$X_2 = f_{X_2}(X_1, Y_1, U_{X_2}) \quad Y_2 = f_{Y_2}(X_2, U_{Y_2})$$



- (Hard) intervention: forces variables to take constants.
  - e.g. $do(S = 1)$ or $do(S = 0)$
- Soft intervention: forces variables to take functional relationship in responding to some other variables.
  - e.g. $do\big(Y_1 = h(X_1)\big)$

Causality-based fairness notions

- Various notions are proposed in the literature, including total effect, direct and indirect discrimination, counterfactual fairness, PC-fairness etc.
- In this work, we use total effect for simplicity, but our method is naturally applicable to other notions.

$$T = P(Y = 1 \mid do(S = 1)) - P(Y = 1 \mid do(S = 0))$$

Method

Core idea: leverage Pearl's structural causal model (SCM), treat each decision model as a soft intervention and infer the post-intervention distributions to formulate the loss function as well as the fairness constraints.

Advantages

- Learn multiple fair classifiers simultaneously and only require static training data.
- Can employ off-the-shelf classification models and optimization algorithms.
- Achieve causal-aware fairness.

Using Soft Interventions to Simulate Decision Model Deployments

- In general, we have l decisions $\{Y_1, \cdots, Y_l\}$.
- For each decision Y_k, we build a classifier h_k $(\mathbf{z}_k)$.
- The soft intervention for deploying all these models is $do(h_1, \cdots, h_l)$.

Loss Function and Fair Constraints

- Traditionally, classification error of classifier $h: \mathbf{Z} \to Y$ is:

$$R(h_k) = \mathbb{E}_{\mathbf{Z}}\left[P(y^+|\mathbf{z})\mathbf{1}_{h(\mathbf{z})<0} + P(y^-|\mathbf{z})\mathbf{1}_{h(\mathbf{z})\geq 0}\right]$$

- Under soft intervention of deploying all models, for classifier h_k

$$R(h_k) = \mathop{\mathbb{E}}_{\mathbf{Z}_k|do(h_1,\cdots,h_l)}\left[P(y_k^+|\mathbf{z}_k)\mathbf{1}_{h_k(\mathbf{z}_k)<0} + P(y_k^-|\mathbf{z}_k)\mathbf{1}_{h_k(\mathbf{z}_k)\geq 0}\right]$$

- Similarly, fairness constraints is given by total effect

$$T(h_k) = P(y_k^+|do(s^+, h_1, \cdots, h_l)) - P(y_k^+|do(s^-, h_1, \cdots, h_l))$$

Deriving Loss Function and Fair Constraints with Observed Data

$$R_\phi(h_k) = \mathop{\mathbb{E}}_{S,\mathbf{X}'_{Y_k}}\Bigg[P(y_k^+|\mathbf{z}_k)\phi(h_k(\mathbf{z}_k))\sum_{\mathbf{Y}'_{Y_k}}\prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+}\phi(-h_i(\mathbf{z}_i))\prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-}\phi(h_i(\mathbf{z}_i))\prod_{X_i\in\mathbf{X}'_{Y_k}}\frac{P(\mathbf{y}'_{X_i}|s,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s,\mathbf{x}'_{X_i})}$$

$$+ P(y_k^-|\mathbf{z}_k)\phi(-h_k(\mathbf{z}_k))\sum_{\mathbf{Y}'_{Y_k}}\prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^+}\phi(-h_i(\mathbf{z}_i))\prod_{Y_i\in\mathbf{Y}'_{Y_k},y_i^-}\phi(h_i(\mathbf{z}_i))\prod_{X_i\in\mathbf{X}'_{Y_k}}\frac{P(\mathbf{y}'_{X_i}|s,x_i,\mathbf{x}'_{X_i})}{P(\mathbf{y}'_{X_i}|s,\mathbf{x}'_{X_i})}\Bigg].$$

- Similarly derive $T_\phi(h_k)$

Problem Formulation for Fair Multiple Decision Making

Problem Formulation The problem of fair multiple decision making for $\boldsymbol{Y} = \{Y_1, \cdots, Y_l\}$ is formulated as the following constrained optimization problem:

$$\min_{h_1,\cdots,h_l\in\mathcal{H}} \sum_{k=1}^{l} R_\phi(h_k) \quad s.t. \quad \forall k, -\tau_k \leq T_\phi(h_k) \leq \tau_k$$

where $R_\phi(h_k)$ and $T_\phi(h_k)$ are smoothed loss function and fair constraint.

Excess Risk Bound

Theorem 1. For any classification-calibrated surrogate function satisfying $\phi(0) = 1$ and $\inf_{\alpha\in\mathbb{R}}\phi(\alpha) = 0$, any measurable function h_k for predicting Y_k, we have

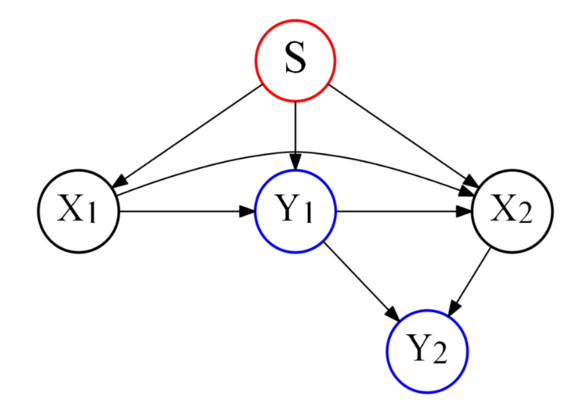
$$\psi(R(h_k) - R^*) \leq R_\phi(h_k) - R^*_\phi,$$

where $\psi(\delta)$ is a non-decreasing function mapping from $[0, \infty)$ to $[0, 1]$.

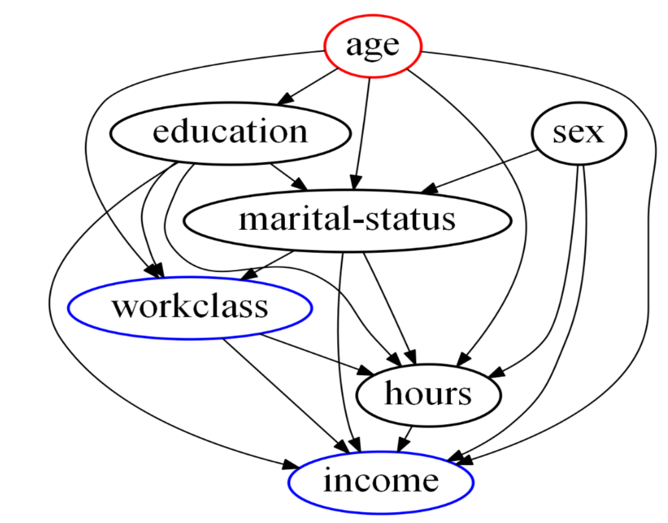
Corollary 1. $R_\phi(h_k) \to R^*_\phi$ indicates $R(h_k) \to R^*$.

Experiments

Datasets



Causal graph for synthetic data



Causal graph for Adult data

Baselines

- Separate method: Each classifier is learned separately on training data.
- Serial method: Classifiers are learned sequentially following the topological order of the causal graph.

Accuracy and unfairness from Unconstrained, Separate, Serial and Joint methods on synthetic and Adult data (bold values indicate violation of fairness).

Phase			Synthetic				Adult			
			Uncons.	Separate	Serial	Joint	Uncons.	Separate	Serial	Joint
Train	h_1	Acc. (%)	80.32	75.35	75.35	75.35	55.71	55.64	55.63	55.63
		Unfairness	**0.15**	0.01	0.01	0.01	**0.15**	0.05	0.05	0.05
	h_2	Acc. (%)	90.13	75.79	84.02	82.77	76.75	71.17	68.90	69.31
		Unfairness	**0.23**	0.04	0.03	0.04	**0.24**	0.10	0.10	0.10
Test	h_1	Acc. (%)	80.70	75.54	75.54	75.54	55.63	55.56	55.57	55.57
		Unfairness	**0.15**	0.01	0.01	0.01	**0.15**	0.05	0.05	0.05
	h_2	Acc. (%)	89.95	77.06	84.16	82.09	77.07	73.33	68.91	69.40
		Unfairness	**0.13**	**0.09**	0.03	0.03	**0.23**	**0.17**	0.10	0.10

Acknowledgement

This work was supported in part by NSF 646654, 1920920, and 1946391.

UNIVERSITY OF ARKANSAS

CLEMSON UNIVERSITY