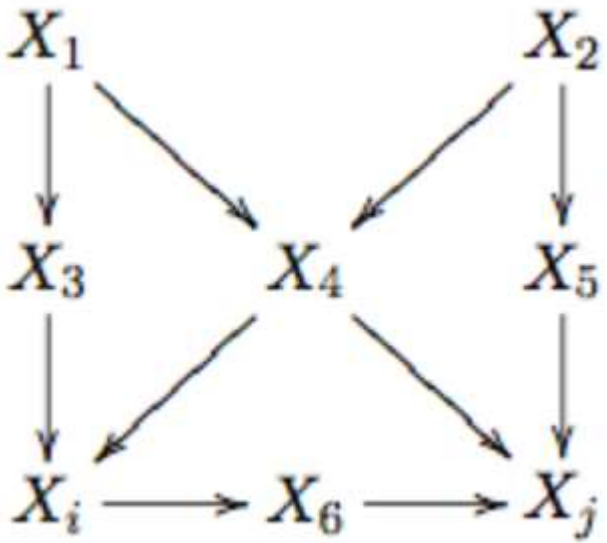
Counterfactual fairness

Shen chuyun

Case：The effect of the treatment

- Zi: i个体是否接受治疗
- {Yi(1), Yi(0)} 表示一组潜在结果，1是接受治疗，0是不接受治疗
- 实际观察结果 Yi = Zi * Yi(1) + (1-Zi) * Yi(0)
- 对Z做随机化
- 平均因果作用 ACE(Z->Y) = E{Yi(1)-Yi(0)}

Causal Diagram



箭头表示存在因果关系
有向无环图DAG

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{p} P(x_i \mid pa_i),$$

Do 算子

- do 的意思可以理解成 “干预” （intervention）
- 在DAG中do(Xi)=x′，表示如下的操作：将DAG中指向 Xi 的所有有向边全部切断，且将Xi的取值固定为常数 xi′

$$P(x_1, \ldots, x_n \mid do(X_i) = x_i') = \frac{P(x_1, \ldots, x_n)}{P(x_i \mid pa_i)} I(x_i = x_i').$$

- ACE(Z->Y) = E{Y|do(Z)=1}-E{Y|do(Z)=0}

Fairness Background

- V: 可观察的特征
- A: 敏感特征，如性别，种族
- X: V中除了A之外的特征
- U: 不可观测的特征

The definition of Fairness

- Fairness Through Unawareness (FTU): 在学习过程中不用A
- Individual Fairness (IF): 相似的人，相似的预测
- Demographic Parity (DP): $P(\hat{Y} | A = 0) = P(\hat{Y} | A = 1)$
- (Equality of Opportunity (EO): $P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1)$

Causal Models and Counterfactuals

- (U, V, F)定义一个因果模型
- 基于假设，不存在W属于U，且W是V中任意M的果
- Vi = fi(pai, Upai) pai是i的父节点集合
- Do操作就是do(Z)=z'，表示把Fz换成z'
- 给定U=u，Do操作记为 $Y_{Z \leftarrow z}(u)$

Counterfactual

- 反事实: $P(Y_{Z \leftarrow z}(U) \mid W = w)$
- 1. Abduction: for a given prior on U, compute the posterior distribution of U given the evidence W = w;
- 2. Action: substitute the equations for Z with the interventional values z, resulting in the modified set of equations Fz;
- 3. Prediction: compute the implied distribution on the remaining elements of V using Fz and the posterior P(U |W = w).

Counterfactual Fairness

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a),$$

为什么不用这个 因为下面这个算的是一个平均

$$P(\hat{Y} = 1 \mid do(A = a)) = P(\hat{Y} = 1 \mid do(A = a'))$$

Lemma

Lemma 1. *Let $\mathcal{G}$ be the causal graph of the given model (U, V, F). Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.*

Proof. Let W be any non-descendant of A in $\mathcal{G}$. Then $W_{A\leftarrow a}(U)$ and $W_{A\leftarrow a'}(U)$ have the same distribution by the three inferential steps in Section 2.2. Hence, the distribution of any function $\hat{Y}$ of the non-descendants of A is invariant with respect to the counterfactual values of A. □

4.1 Algorithm

Let $\hat{Y} \equiv g_\theta(U, X_{\not\succ A})$ be a predictor parameterized by θ, such as a logistic regression or a neural network, and where $X_{\not\succ A} \subseteq X$ are non-descendants of A. Given a loss function $l(\cdot, \cdot)$ such as squared loss or log-likelihood, and training data $\mathcal{D} \equiv \{(A^{(i)}, X^{(i)}, Y^{(i)})\}$ for $i = 1, 2, \ldots, n$, we define $L(\theta) \equiv \sum_{i=1}^{n} \mathbb{E}[l(y^{(i)}, g_\theta(U^{(i)}, x^{(i)}_{\not\succ A})) \mid x^{(i)}, a^{(i)}]/n$ as the empirical loss to be minimized with respect to θ. Each expectation is with respect to random variable $U^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$ where $P_{\mathcal{M}}(U \mid x, a)$ is the conditional distribution of the background variables as given by a causal model $\mathcal{M}$ that is available by assumption. If this expectation cannot be calculated analytically, Markov chain Monte Carlo (MCMC) can be used to approximate it as in the following algorithm.

1: **procedure** FAIRLEARNING($\mathcal{D}, \mathcal{M}$) ▷ Learned parameters $\hat{\theta}$
2: For each data point $i \in \mathcal{D}$, sample m MCMC samples $U^{(i)}_1, \ldots, U^{(i)}_m \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
3: Let $\mathcal{D}'$ be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in $\mathcal{D}$ is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u^{(i)}_j)\}$.
4: $\hat{\theta} \leftarrow \operatorname{argmin}_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(U^{(i')}, x^{(i')}_{\not\succ A}))$.
5: **end procedure**

At prediction time, we report $\tilde{Y} \equiv \mathbb{E}[\hat{Y}(U^\star, x^\star_{\not\succ A}) \mid x^\star, a^\star]$ for a new data point $(a^\star, x^\star)$.

Three level

Level 1. Build $\hat{Y}$ using only the observable non-descendants of A. This only requires partial causal ordering and no further causal assumptions, but in many problems there will be few, if any, observables which are not descendants of protected demographic factors.

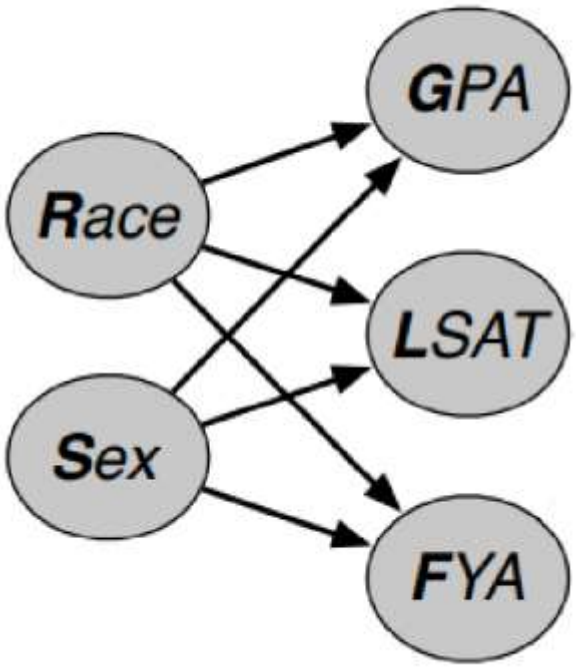
Level 2. Postulate background latent variables that act as non-deterministic causes of observable variables, based on explicit domain knowledge and learning algorithms[5]. Information about X is passed to $\hat{Y}$ via $P(U \mid x, a)$.

Level 3. Postulate a fully deterministic model with latent variables. For instance, the distribution $P(V_i \mid pa_i)$ can be treated as an additive error model, $V_i = f_i(pa_i) + e_i$ [31]. The error term e_i then becomes an input to $\hat{Y}$ as calculated from the observed variables. This maximizes the information extracted by the fair predictor $\hat{Y}$.

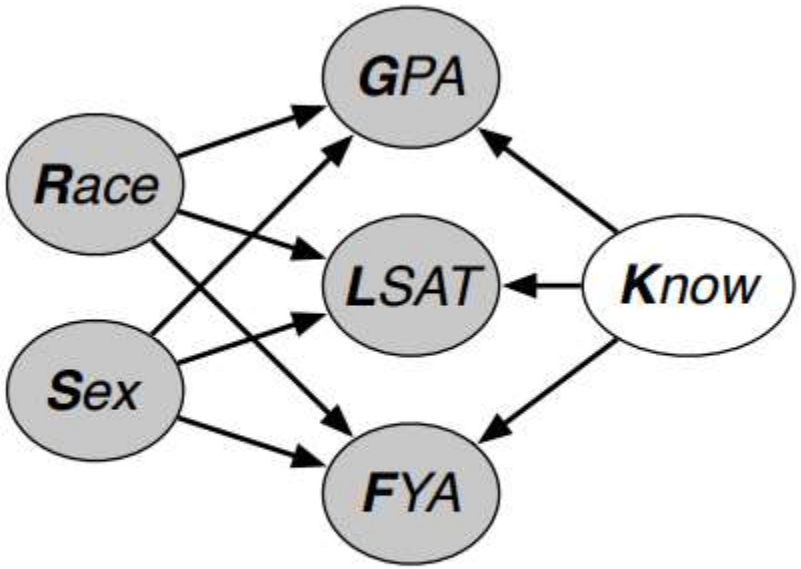
Law School

- entrance exam scores (LSAT)
- grade-point average (GPA) collected prior to law school
- their first year average grade (FYA)
- Given this data, a school may wish to predict if an applicant will have a high FYA.

Law School



Law School

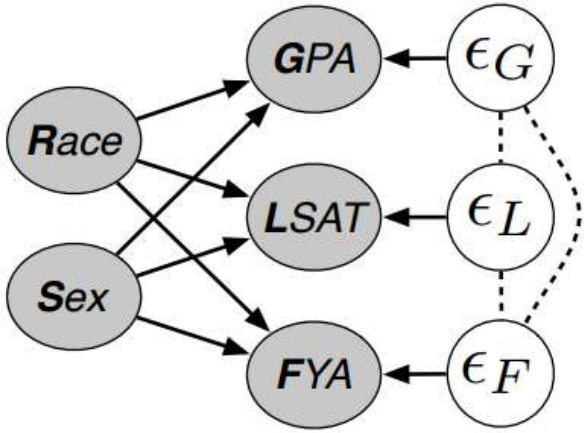


Level 2

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G),$$
$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)),$$

$$\text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$
$$\text{K} \sim \mathcal{N}(0, 1)$$

Law School

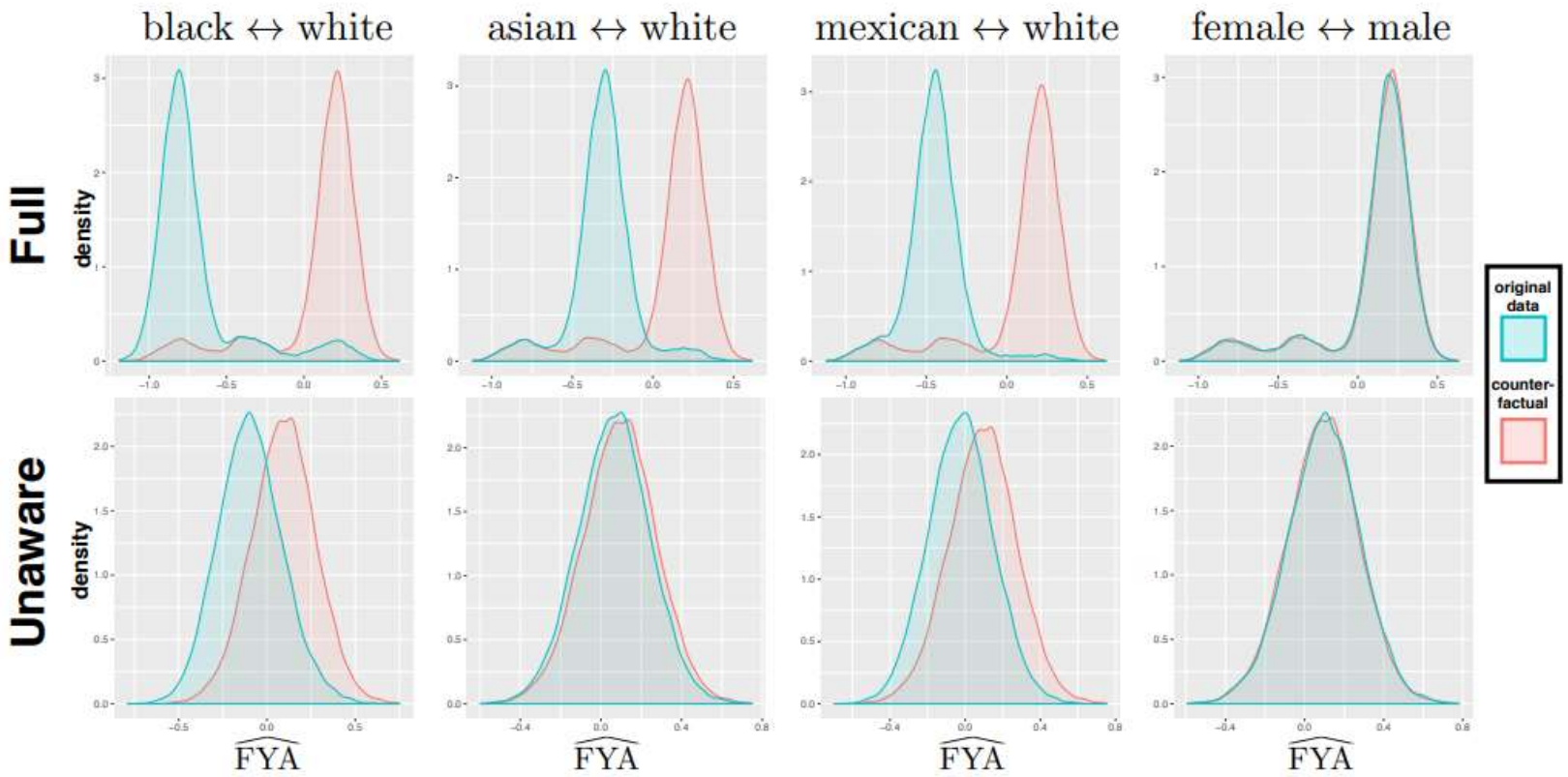


Level 3

$$\text{GPA} = b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G)$$
$$\text{LSAT} = b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L)$$
$$\text{FYA} = b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)$$

We estimate the error terms ϵ_G, ϵ_L by first fitting two models that each use race and sex to individually predict GPA and LSAT. We then compute the residuals of each model (e.g., $\epsilon_G = \text{GPA} - \hat{Y}_{\text{GPA}}(R, S)$). We use these residual estimates of ϵ_G, ϵ_L to predict FYA. We call this *Fair Add*.

Law School


black ↔ white
asian ↔ white
mexican ↔ white
female ↔ male
Full
Unaware
density
density
FYA
FYA
FYA
FYA
original data
counter-factual

Law School

	Full	**Unaware**	**Fair K**	**Fair Add**
RMSE	0.873	0.894	0.929	0.918