Counterfactual Fairness

Pseudo Lab Week.4

Soyeon Kim

왜 선택했는가

- T_T

이 논문은 무엇을 하는 논문인가

Ultimately, what this paper does..

- In US, Law Scholl Admission Council에 의해 163개의 법대와 21,790명 학생들의 입학시험(LSAT), 법대 입학 전 학점(GPA), 1학년대 평점(FYA)를 조사했고,
- FYA 점수에 대한 prediction을 수행

Table 1: Prediction results using logistic regression. Note that we must sacrifice a small amount of accuracy to ensuring counterfactually fair prediction (Fair K, Fair Add), versus the models that use unfair features: GPA, LSAT, race, sex (Full, Unaware).

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

Error니까 작을 수록 좋은 것

Ultimately, what this paper does..

Table 1: Prediction results using logistic regression. Note that we must sacrifice a small amount of accuracy to ensuring counterfactually fair prediction (Fair K, Fair Add), versus the models that use unfair features: GPA, LSAT, race, sex (Full, Unaware).

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

Error니까 작을 수록 좋은 것

Ultimately, what this paper does..

- 일학년 학점 맞추는데, GPA,LSAT 외에도 굉장히 Private한 요소들이 영향을 미칠 수도 있고 안 미칠 수도 있는데
- Bias가 되어 있는 Private한 요소를 포함하는 것은 Fairness 문제에 걸림!
- 그러니까 최종 Prediction에서 그러한 요소들을 사용하지 않고 추론하는 방법이 없을까?

Table 1: Prediction results using logistic regression. Note that we must sacrifice a small amount of accuracy to ensuring counterfactually fair prediction (Fair K, Fair Add), versus the models that use unfair features: GPA, LSAT, race, sex (Full, Unaware).

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

손해는 trade-off.

Motivation & Background

Problem Motivation

- **Unfairly biased** against certain **subpopulations**
  - race, gender, or sexual orientation
  - unfairness for reasons having to do with **historical prejudices or other factors outside an individual's control**.
  - Since **this past data may be biased**, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices.

- Necessary for the modeler **to think beyond the objective of maximizing prediction accuracy and consider the societal impact of their work.** *fair*.

Problem Motivation

- **Unfairly biased** against certain **subpopulations**
  - race, gender, or sexual orientation
  - unfairness for reasons having to do with **historical prejudices or other factors outside an individual's control**.
  - Since **this past data may be biased**, machine learning predictors must account for this to avoid perpetuating or creating discriminatory practices.

- Proposed: a framework for **modeling fairness** using tools from **causal inference.**

Method Motivation

- Causal framework
  - To model the relationship between protected attributes and data.

- *counterfactual fairness*
  - Enforces that a distribution over **possible predictions for an individual should remain unchanged** in a world where an individual's **protected attributes had been different in a causal sense.**

Prediction을 함에 있어서, 어떤 protected attribute 이 변하더라도 prediction이 바뀌지 않도록 하는 것,
(특히 그 protected attribute이 prediction에 bias / unfairness를 줄 수도 있는 요인이라고 보여질 때)

Related Works

- 그냥 sensitive한거 빼..

Definition 1 (Fairness Through Unawareness (FTU)). *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*

Any mapping $\hat{Y} : X \rightarrow Y$ that excludes A satisfies this. Initially proposed as a baseline, the approach has found favor recently with more general approaches such as Grgic-Hlaca et al. [12]. Despite its compelling simplicity, FTU has a clear shortcoming as elements of X can contain discriminatory information analogous to A that may not be obvious at first. The need for expert knowledge in assessing the relationship between A and X was highlighted in the work on individual fairness:

Related Works

- 비슷한 사람한테서 비슷한 결과를 주면..

Definition 2 (Individual Fairness (IF)). *An algorithm is fair if it gives similar predictions to similar individuals. Formally, given a metric* $d(\cdot,\cdot)$*, if individuals* i *and* j *are similar under this metric (i.e.,* $d(i,j)$ *is small) then their predictions should be similar:* $\hat{Y}(X^{(i)}, A^{(i)}) \approx \hat{Y}(X^{(j)}, A^{(j)})$.

As described in [10], the metric $d(\cdot,\cdot)$ must be carefully chosen, requiring an understanding of the domain at hand beyond black-box statistical modeling. This can also be contrasted against population level criteria such as

Related Works

- 비슷한 사람한테서 비슷한 결과를 주면..

Definition 3 (Demographic Parity (DP)). *A predictor* $\hat{Y}$ *satisfies demographic parity if* $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$.

Related Works

- 비슷한 사람한테서 비슷한 결과를 주면..

Definition 4 (Equality of Opportunity (EO)). *A predictor* $\hat{Y}$ *satisfies equality of opportunity if* $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$.

Proposed Method

Several Notations

- A : the set of *protected attributes* of an individual, variables that must not be discriminated against in a formal sense defined differently by each notion of fairness discussed.
- X : the other observable attributes of any particular individual,
- U : the set of relevant latent attributes which are not observed
- Y : the outcome to be predicted
- $\hat{Y}$

Proposed Definition

Given a predictive problem with fairness considerations, where A, X and Y represent the protected attributes, remaining attributes, and output of interest respectively, let us assume that we are given a causal model (U, V, F), where $V \equiv A \cup X$. We postulate the following criterion for predictors of Y.

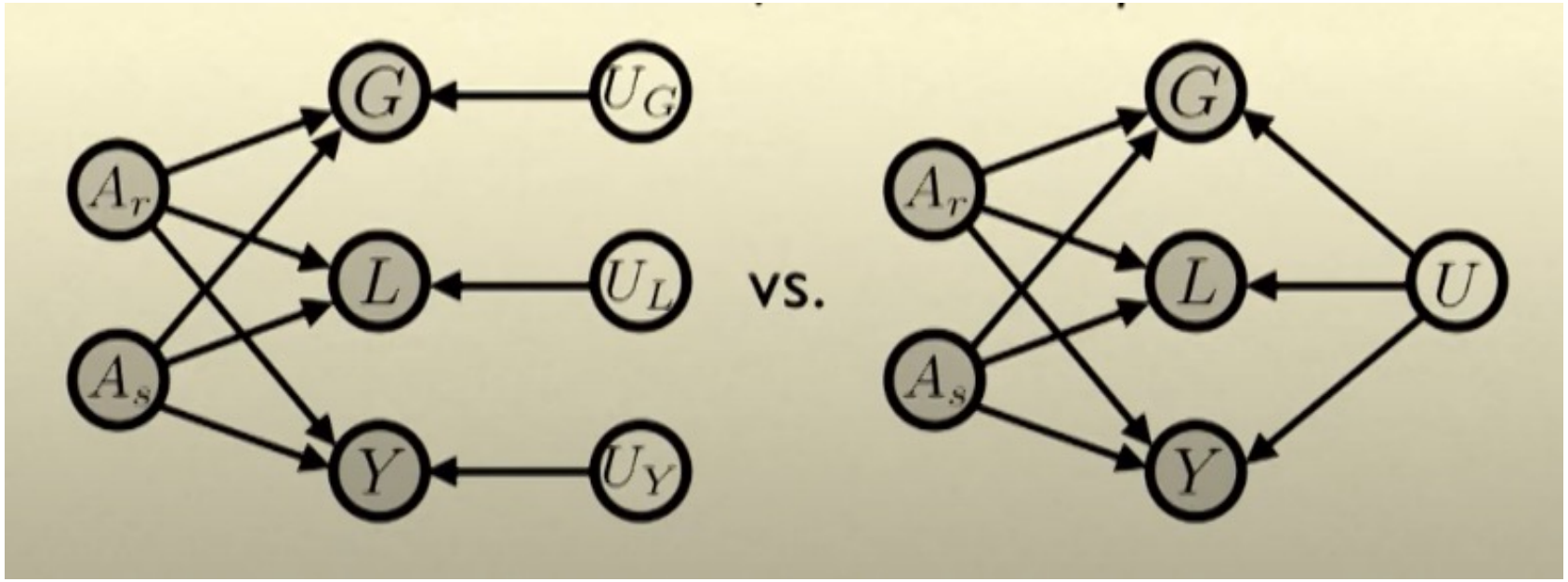
Definition 5 (Counterfactual fairness)**.** *Predictor* $\hat{Y}$ *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a), \tag{1}$$

for all y *and for any value* a' *attainable by* A.

Causal Model?

- (정훈) 논문에서 문제 제기한 social biases를 explicit causal models로 찾았다고 하는데 해당 논문에서 제한 **causal models**가 무엇일까요?



Causal Model

- 구조적 인과 모델(Structural Casual Model)이란 변수들 간의 인과 관계를 구조적인 식으로 나타낸 것입니다. 즉 변수들과 변수들 간 관계를 설명하는 함수로 구성되어 있습니다. **Y라는 변수가 X에 값을 할당하는 어떤 식(함수) 안에 등장할 경우 Y를 X의 직접적 원인(direct cause)**이라고 보겠습니다.
- 즉 X = Y + 2라는 함수가 우리의 모델이라면, 우리 모델은 Y가 X의 원인이라고 주장하는 것이라고 보면 됩니다.

https://pizzathief.oopy.io/causal-graph-for-beginners

Preliminary

https://www.youtube.com/watch?v=_V8eKsto3Ug

Structural Causal Model

Structural equations

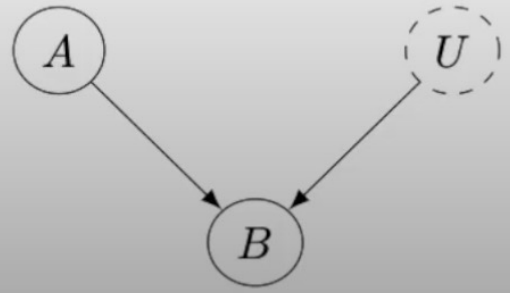
The equals sign does not convey any causal information.

$B = A$ means the same thing as $A = B$

Structural equation for A as a cause of B:

$$B := f(A)$$

$$B := f(A, U)$$



Preliminary

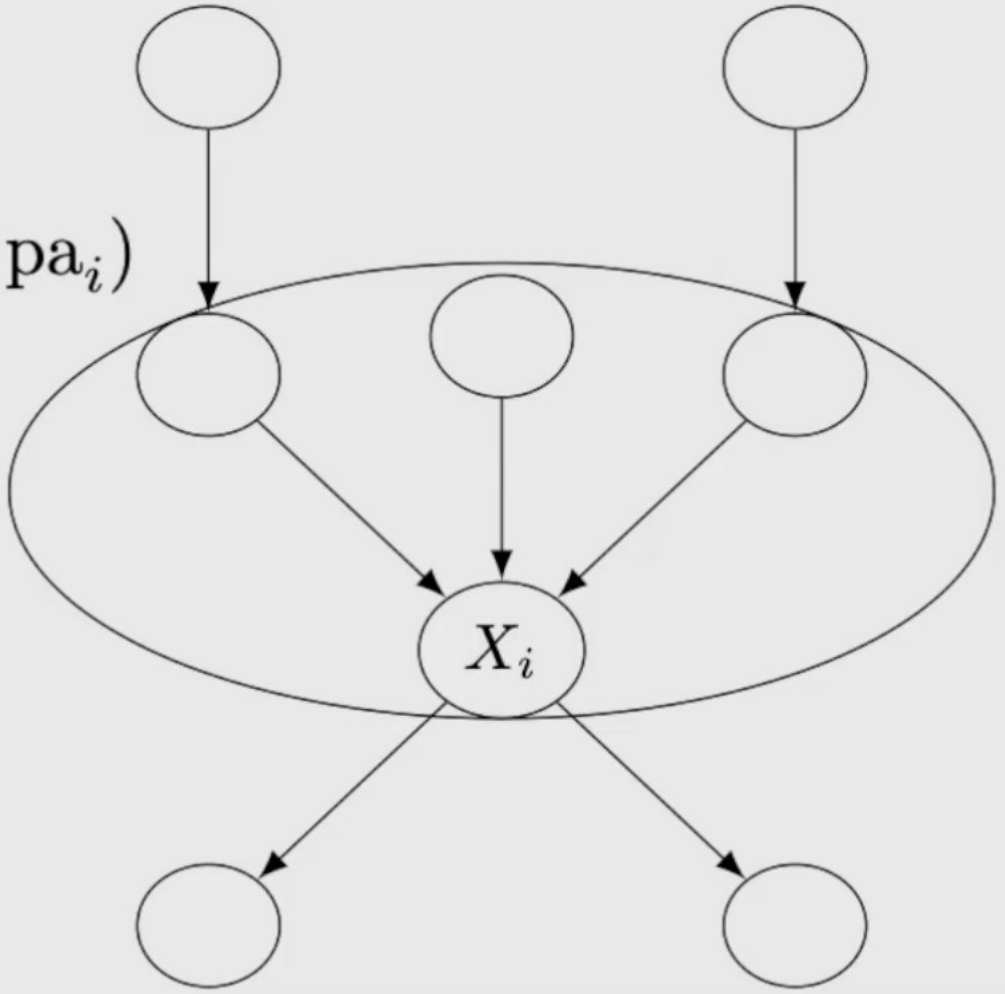
https://www.youtube.com/watch?v=_V8eKsto3Ug

Structural Causal Model

Causal mechanism for X_i

$$X_i := f(A, B, \ldots)$$

Direct causes of X_i



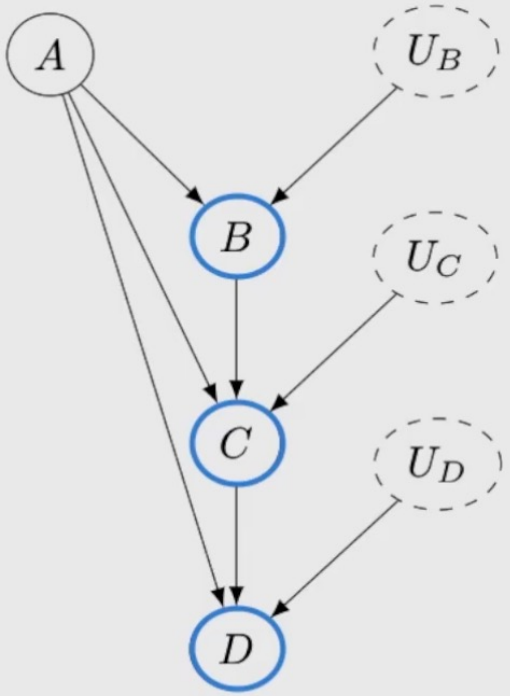
Preliminary

https://www.youtube.com/watch?v=_V8eKsto3Ug

Structural Causal Model

Structural causal models (SCMs)

$$M: \quad \begin{aligned} B &:= f_B(A, U_B) \\ C &:= f_C(A, B, U_C) \\ D &:= f_D(A, C, U_D) \end{aligned}$$



Endogenous variables

Endogenous variables:
B,C,D에 대한 cause를 모델링하는 것. 즉 모델링의 타겟

exogeneous variables:
B,C,D에 대한 cause를 모델링하는데 있어서, 얼만큼의 cost (영향)을 주는 요인들
Parents X

Preliminary

https://www.youtube.com/watch?v=_V8eKsto3Ug

Structural Causal Model

$$
\begin{aligned}
& B := f_B(A, U_B) \\
M: \quad & C := f_C(A, B, U_C) \\
& D := f_D(A, C, U_D)
\end{aligned}
$$

SCM Definition

A tuple of the following sets:

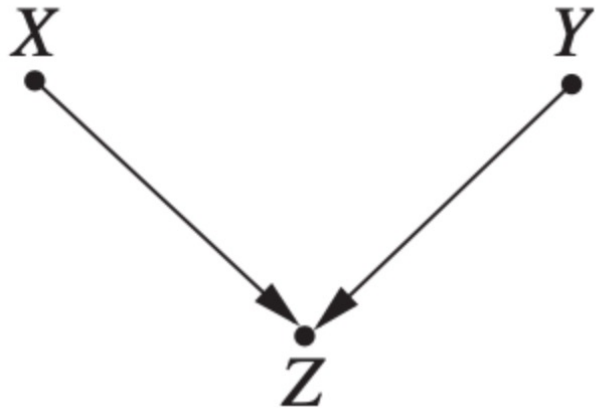
1. A set of endogenous variables
2. A set of exogenous variables
3. A set of functions, one to generate each endogenous variable as a function of the other variables

Structural Causal Model

모든 SCM은 그래프로 표현될 수 있습니다. U와 V에 속한 변수들을 노드로 하고 f의 함수들을 엣지라고 생각하는 것입니다. Y가 X의 직접적 원일일 경우 `Y -> X` 이렇게 X를 Y의 자식 노드로 그리게 되면, 익숙한 **방향성 있는 비순환 그래프(DAG)**의 모습으로 그려질 것입니다. 외생 변수의 경우 모델 내에서 설명하지 않는 변수이므로, 어느 노드의 자식도 아닌 루트 노드여야 할 것입니다.

예를 들면 직원들의 월급(Z), 교육 수준(X), 경력(Y)이 있을 때,

- SCM으로 표현하기 :
  - $U = \{X, Y\}, V = \{Z\}, F = \{f_z\}$
  - $f_z : Z = 2X + 3Y$
- 그래프로 표현하기 :



https://pizzathief.oopy.io/causal-graph-for-beginners

Why Causal Model?

- 인과관계에 대해 사실 수치적으로 규명할 수 있는 경우는 드물고 대부분 질적 관계만 발견해낼 수 있기 때문에
- 직관적인 이해에 도움이 되기 때문에
- 변수들 간 결합분포(joint distribution)를 매우 효율적으로 표현할 수 있게 해 주기 때문에
  - **고차원의 추정 문제를 저차원의 확률 분포 문제로 바꿀 수 있다(독립/조건부 독립 등 파악 가능)**

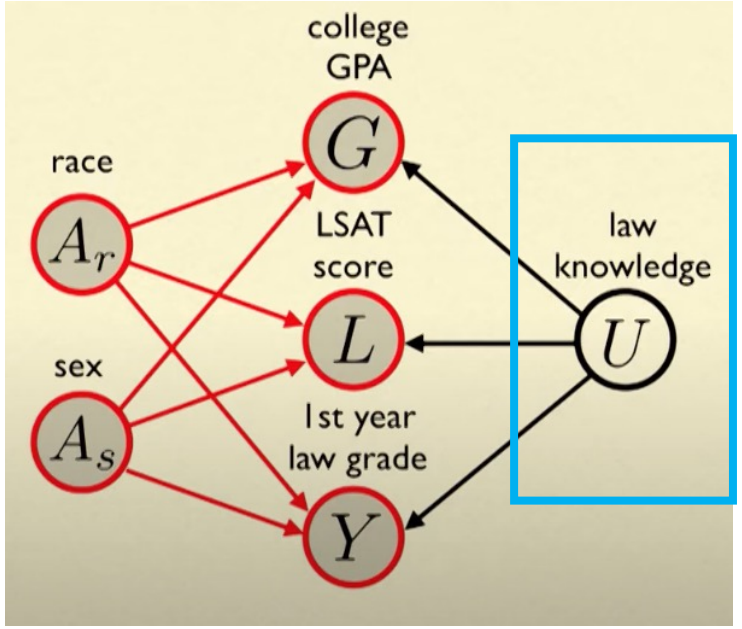
https://pizzathief.oopy.io/causal-graph-for-beginners

Proposed Definition

Definition 5 (Counterfactual fairness). *Predictor* Y *is* **counterfactually fair** *if under any context* $X = x$ *and* $A = a$,

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a), \quad (1)$$

for all y *and for any value* a' *attainable by* A.



Predictions **using non-descendants of A** are counterfactually fair

Proposed 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 $U^{(i)}_1, \ldots, U^{(i)}_m \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.

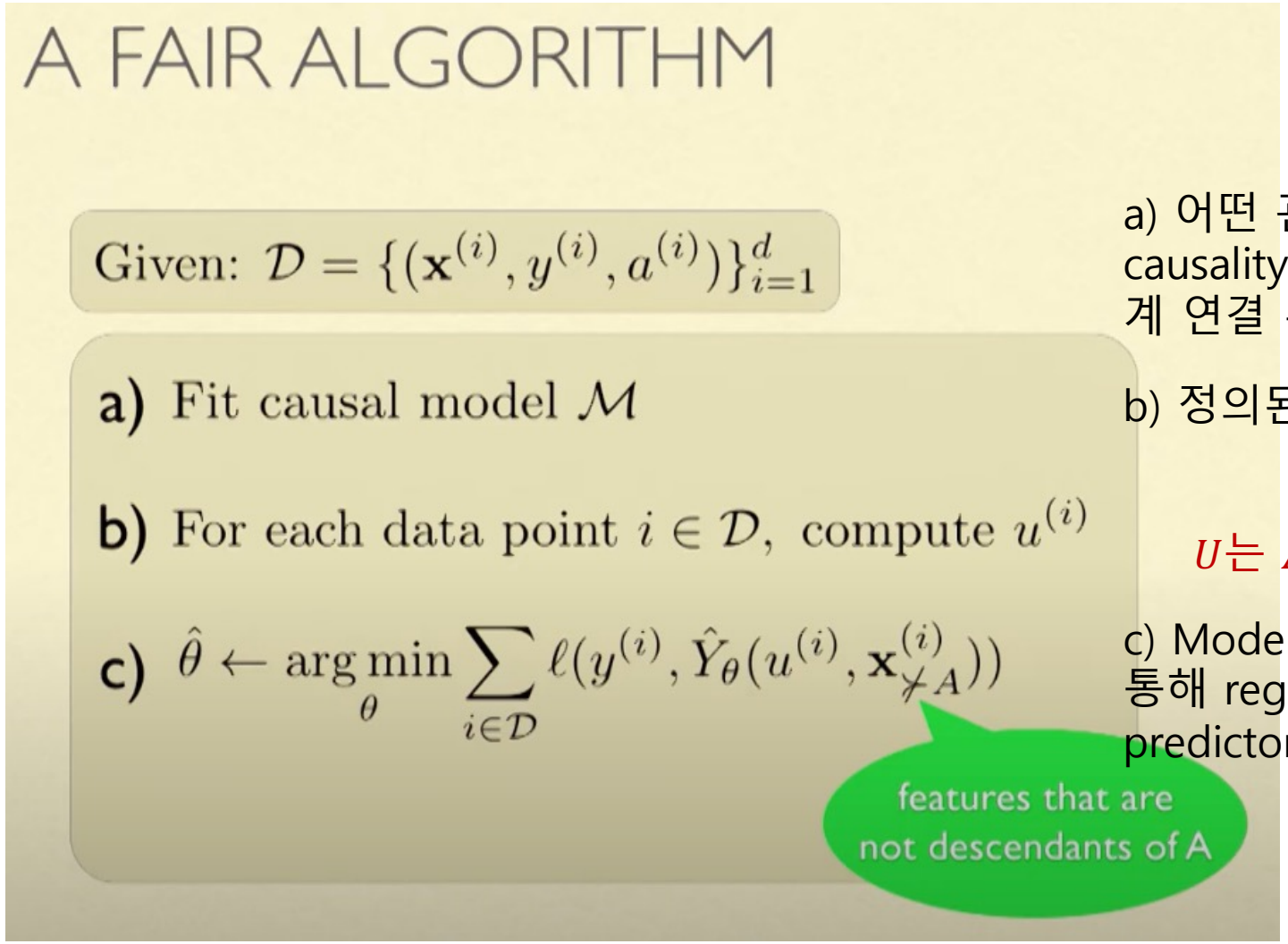
3: Let $\mathcal{D}'$ be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in $\mathcal{D}$ is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u^{(i)}_j)\}$.

4: $\hat{\theta} \leftarrow \text{argmin}_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(U^{(i')}, x^{(i')}_{\not\succ A}))$.

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 $(a^\star, x^\star)$.

Proposed Algorithm



a) 어떤 관계가 있을거라고 보여지는 모형을 causality를 고려해서 modeling하고(modeling이 관계 연결 + U에 대한 distribution도 정의하는듯!)

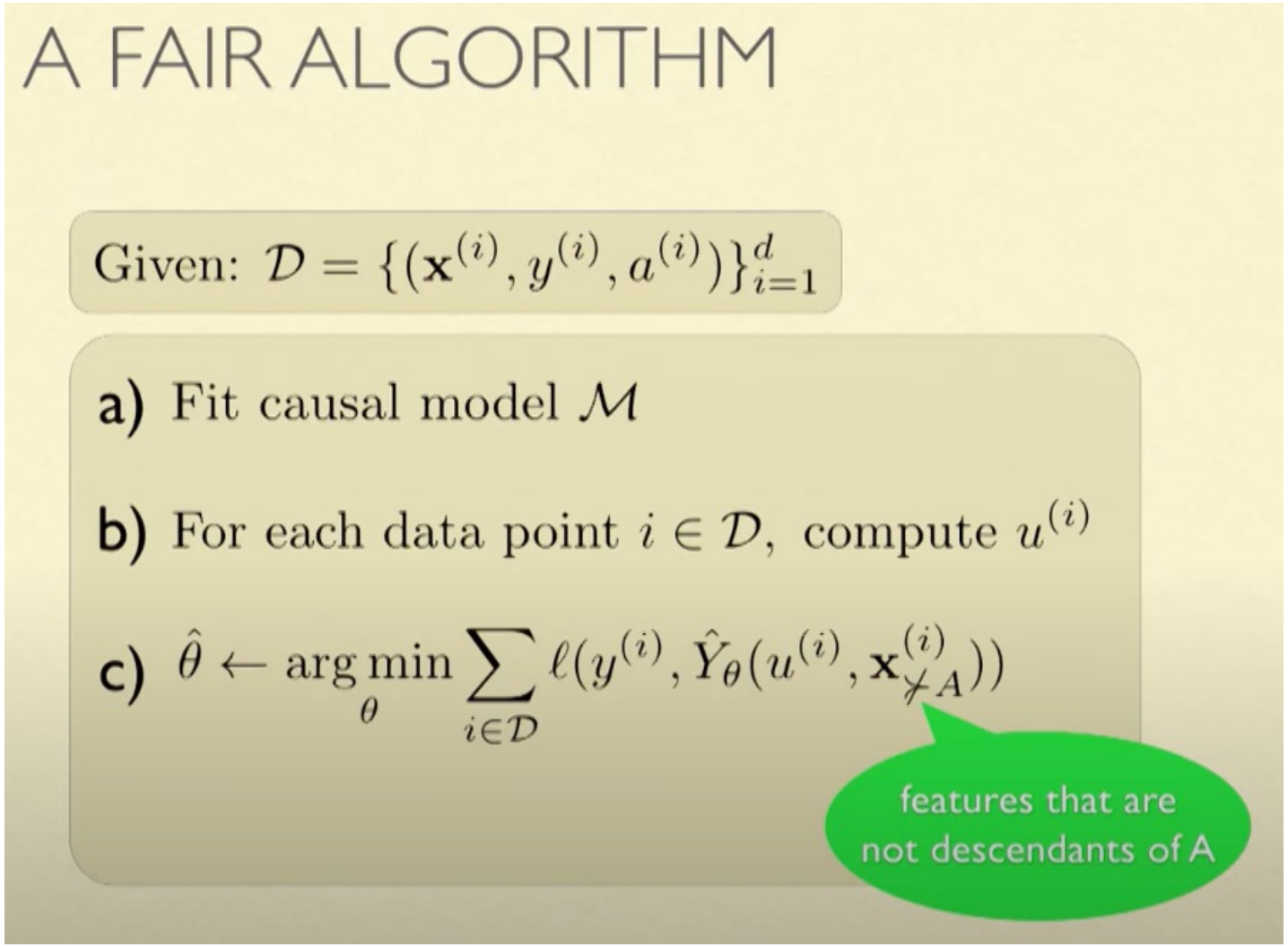
b) 정의된 distribution에 대해서 u를 sampling하고

$$U_1^{(i)}, \ldots, U_m^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$$

U는 A 의 descendent가 아닌 요소로 modeling

c) Modeling, 그리고 input으로 들어가는 value 등을 통해 regression/ neural network등을 이용해서 predictor를 optimize

Proposed Algorithm

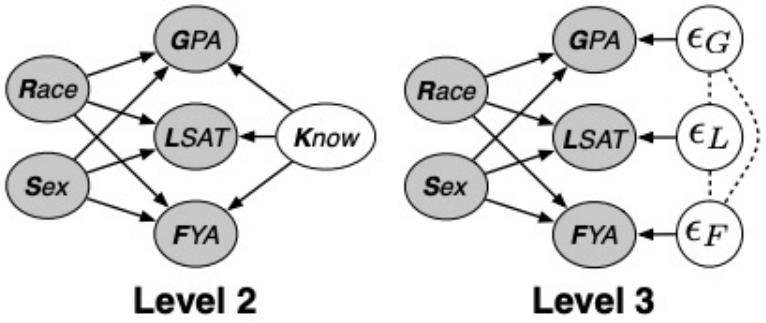


- 즉, data들 중 sensitive할 수 있는 것들을 포함하지 않고도 그래도 좋은 성능의 predictor를 만들고 싶음
  - Counterfactual fairness 이라는 definition을 정의하고
  - 이 정의에 적합한 fairness를 guarantee하는 classifier를 학습하는 알고리즘.
- 그러한 sensitive attribute을 포함 하지도 않고, 그것들의 descendent도 아닌,
- 어떤 latent 요인에 대해서 영향을 받는다는 생각 하에
- Causal modeling을 하고
- 그 modeling에 기반하여 predictor(e.,g NN) 를 학습시킴

Experiments

Fair K, Fair ADD

Level Description



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 $\hat{Y}$.

Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

Level 1: 관찰 가능한(데이터에서 주어지는) Protected Attribute A의 descendent가 아닌 feature 사용

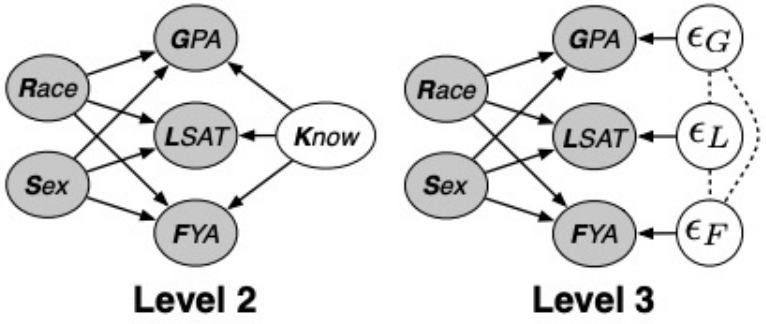
Level 2: observable variable에 대해 latent variable이 non-deterministic하게 있을 것이라는 가정, 이때의 latent variable은 도메인 지식이나 학습 알고리즘에 의해 근거

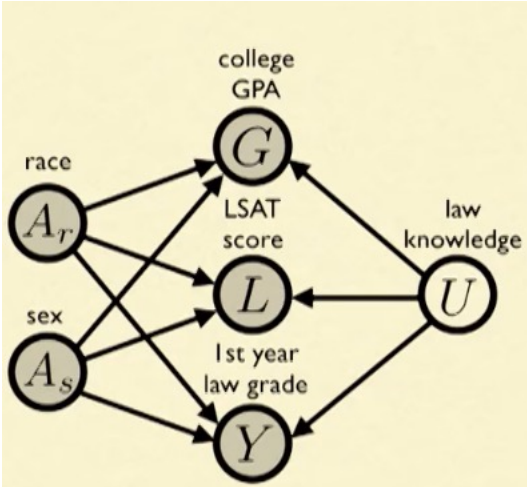
Level 3: 각각의 항목에 대해 latent variable이 deterministic하게 존재할 것.

Fair K, Fair ADD

Causal model 1

Parameterizing, and fit W,b on the given data





structural causal model
(non-deterministic)

$$G \sim \text{Normal}(b_G + [U, A_r, A_s]\mathbf{w}_G, \sigma_G)$$
$$L \sim \text{Poisson}(\exp(b_L + [U, A_r, A_s]\mathbf{w}_L))$$
$$Y \sim \text{Normal}([U, A_r, A_s]\mathbf{w}_Y, 1)$$
$$U \sim \text{Normal}(0, 1)$$

Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

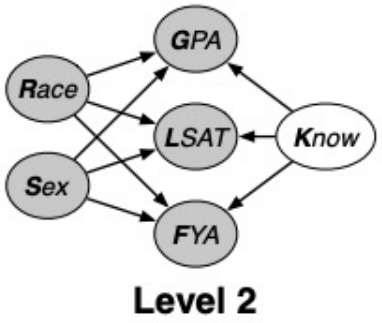
In Level 2, we postulate that a latent variable: a student's **knowledge** (K), affects GPA, LSAT, and FYA scores. The causal graph corresponding to this model is shown in Figure 2, (**Level 2**). This is a short-hand for the distributions:

$$\text{GPA} \sim \mathcal{N}(b_G + w_G^K K + w_G^R R + w_G^S S, \sigma_G), \qquad \text{FYA} \sim \mathcal{N}(w_F^K K + w_F^R R + w_F^S S, 1),$$
$$\text{LSAT} \sim \text{Poisson}(\exp(b_L + w_L^K K + w_L^R R + w_L^S S)), \qquad \text{K} \sim \mathcal{N}(0, 1)$$

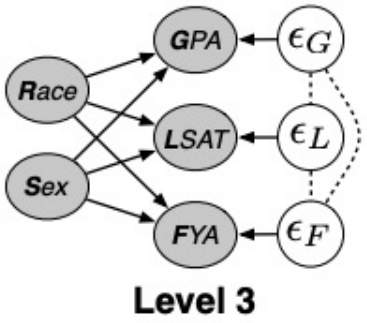
Fair K, Fair ADD

Causal model 2

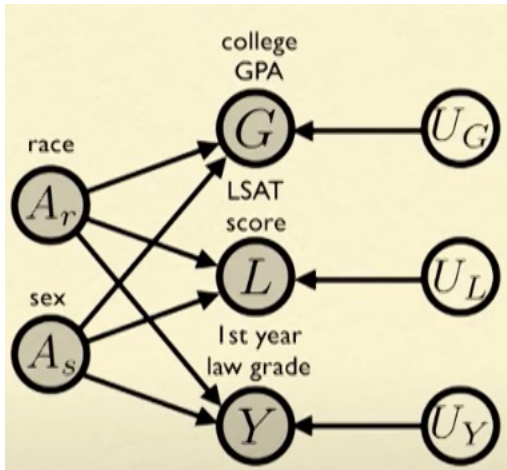
Parameterizing, and fit W,b on the given data



Level 2



Level 3



structural causal model
(deterministic)

$$G = b_G + [A_r, A_s]\mathbf{w}_G + U_G$$
$$L = b_L + [A_r, A_s]\mathbf{w}_L + U_L$$
$$Y = b_Y + [A_r, A_s]\mathbf{w}_Y + U_Y$$
$$U_G \sim \mathbb{P}(U_G)$$
$$U_L \sim \mathbb{P}(U_L)$$
$$U_Y \sim \mathbb{P}(U_Y)$$

Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

in Figure 2, (**Level 3**), and is expressed by:

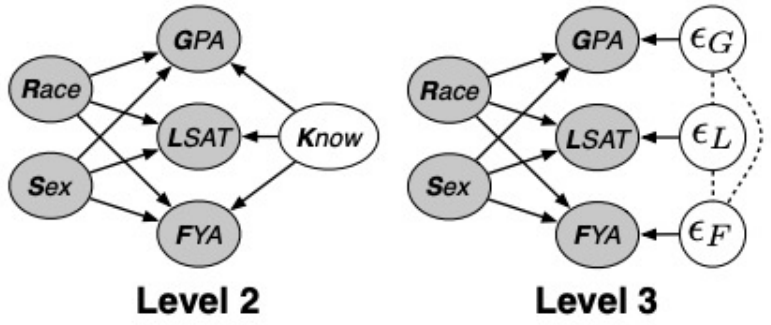
$$\text{GPA} = b_G + w_G^R R + w_G^S S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G)$$
$$\text{LSAT} = b_L + w_L^R R + w_L^S S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L)$$
$$\text{FYA} = b_F + w_F^R R + w_F^S S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)$$

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*.

Fair K, Fair ADD

Full, Unaware 모드는 각 요소들을 바꿔서 inference했을 때의 Posterior distribution이 너무 상이하다 → unfair!

Inference on posterior distribution



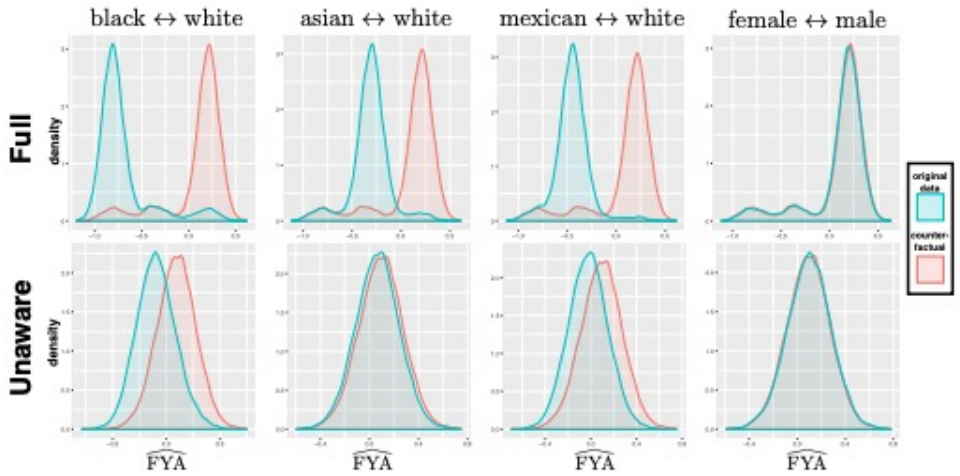
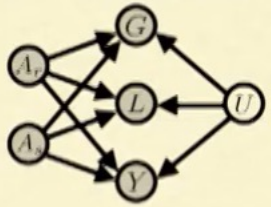


Figure 2: **Left:** A causal model for the problem of predicting law school success fairly. **Right:** Density plots of predicted FYA_a and $\text{FYA}_{a'}$.

THIS TALK

Part I: Counterfactual Fairness

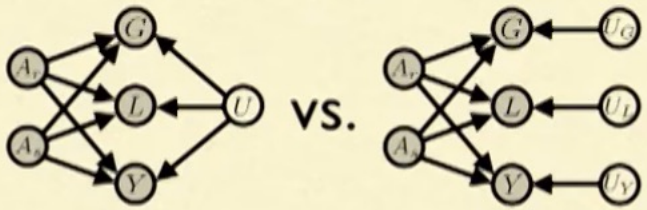
a solution via **causal models**



fair law school predictions

Part II: Fairness in Multiple Worlds

fair predictions in **multiple possible causal models**



optimization problem with fairness constraints

$$\min_{\hat{Y}} \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{Y}(\mathbf{x}_i, a_i), y_i) + \lambda \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \sum_{a' \neq a_i} \text{UNF}_j(\hat{Y}, \mathbf{x}_i, a_i, a')$$

Discussion Points