Causal inference: Counterfactual fairness

Yunhao CHEN, Hugo PELTIER

March 30, 2023

Plan

Introduction

- ML models are trained on data that might be biased
- ML prediction might be discriminatory (crime prediction, credit scoring...)
- Fairness as been studied as a mathematical framework (Berk et al.)
- The article introduces a causal notion of fairness

What is fairness? A mathematical approach

- X: set of observable attributes
- U: set of unobservable attributes
- Y: outcome to predict
- A: protected attributes that should not be discriminated against
- $\hat{Y}$: predicted outcome

Usual definitions of fairness

- Fairness Through Unawareness (FTU): A is not used in the decision making process
- Individual Fairness (IF): $(A^{(i)}, X^{(i)}) \approx (A^{(j)}, X^{(j)}) \Rightarrow \hat{Y}(A^{(i)}, X^{(i)}) \approx \hat{Y}(A^{(j)}, X^{(j)})$
- Demographic Parity (DP): $P(\hat{Y}|A=0) = P(\hat{Y}|A=1)$
- Equality of Opportunity (EO): $P(\hat{Y}|A=0, Y=1) = P(\hat{Y}|A=1, Y=1)$

Pearl's causal model

- U latent background variables, not caused by any variable in V
- F is a set of functions $\{f_1, ..., f_n\}$, one for each $Vi \in V$, such that $Vi = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \setminus \{Vi\}$ and $U_{pa_i} \subseteq U$
- These are the structural equations (DAG)

Counterfactual fairness: definition

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

Examples

- Red cars in insurance
- High Crime Regions

Major theoretical result

- Lemma: 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.
- Counterintuitive result: ancestors of members of A can be used to create a counterfactually fair estimator

Fair learning part 1

- $X_{\nsucc A} \subseteq X$ is the set of non-descendants of A
- $\hat{Y} = g_\theta(U, X_{\nsucc A})$
- $L(\theta) = \sum_{i=1}^{n} E[l(y^{(i)}, g_\theta(U^{(i)}, x_{\nsucc A}^{(i)}))|x^{(i)}, a^{(i)}]/n$, an empirical loss to minimize
- $U^{(i)} \backsim P_{\mathcal{M}}(U|x^{(i)}, a^{(i)})$, which is either given or estimated by MCMC

Fair learning part 2

- For i in the dataset $\mathcal{D}$ sample m MCMC samples $U_1^{(i)}, ..., U_m^{(i)} \backsim P_{\mathcal{M}}(U|x^{(i)}, a^{(i)})$
- replace $\mathcal{D}$ by $\mathcal{D}'$ where the points $(a^{(i)}, x^{(i)}, y^{(i)})$ are replaced by the sets $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}$
- $\hat{\theta} \leftarrow argmin_\theta \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_\theta(u^{(i')}, x_{\nsucc A}^{(i')}))$

Input causal model

- Level 1: use only the observables non-descendant from A
- Level 2: use $P(U|X, A)$, to take the unobservables into account
- Level 3: use a deterministic model to determine unobservables

Dataset

- Law school success study
- Collected by Wightman et al. in 1998 in the paper Lsac national longitudinal bar passage study, available on Kaggle

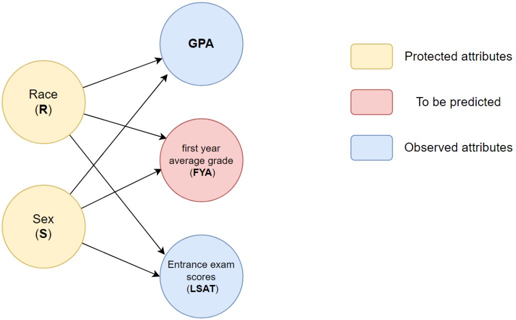

Figure: General causal model to predict law school success

Experimental protocol

- Data: 21406 valid samples, split it 80/20 into train/test set, preserving race and sex balance. C.f Page 22 in Appendix
- Framework: **Pyro**, a probabilistic programming language built on Python and PyTorch
- Four methods:
  - **Full** model
  - **Unaware** model
  - **Fair K** model
  - **Fair Add** model
- Linear regression for all above methods
- Code: Find the code on Github

Results: Full model

Use all attributes including the protected ones (race and sex) to predict FYA

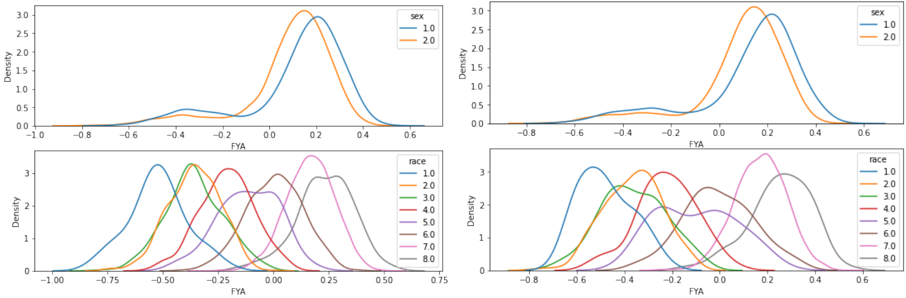

Figure: Full model: Dist. of FYA separated out by Race and Sex: train set (left), test set (right)

Results: Unaware model

Use only non-protected observed attributes (GPA and LSAT) to predict FYA

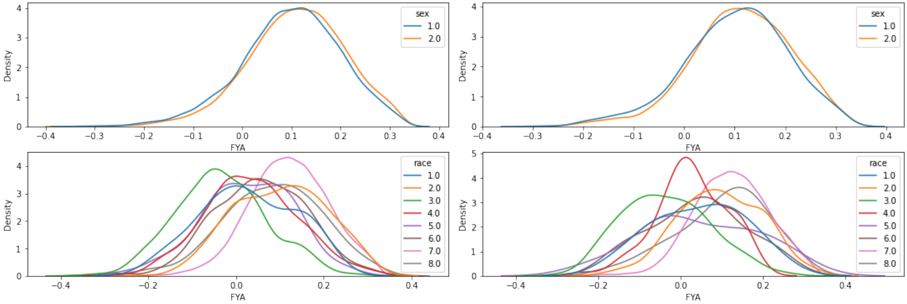

Figure: Unaware model: Dist. of FYA separated out by Race and Sex: train set (left), test set (right)

Results: Fair K model

Infer the latent variable knowledge(K) from observations, to predict FYA

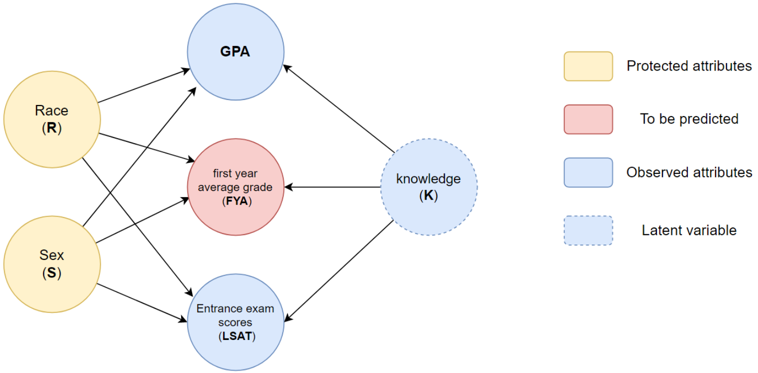

Figure: Causal model with latent ‘fair’ variable K which is parent of GPA and LSAT

Results: Fair K model

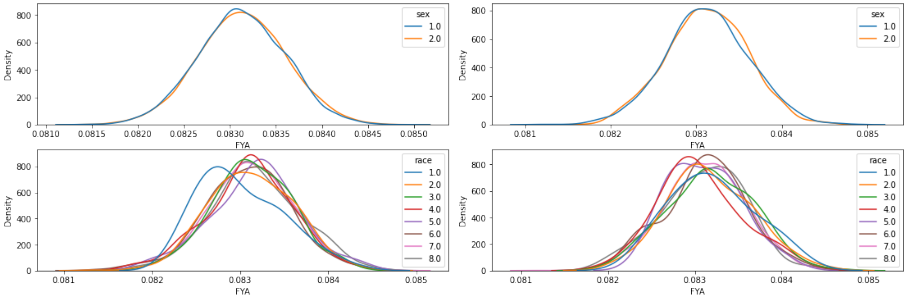

Figure: Fair K model: Dist. of FYA separated out by Race and Sex: train set (left), test set (right)

Results: Fair Add model

Consider GPA and LSAT as continuous variables with additive error terms independent of protected attributes (race and sex)
Use these two error terms to predict FYA

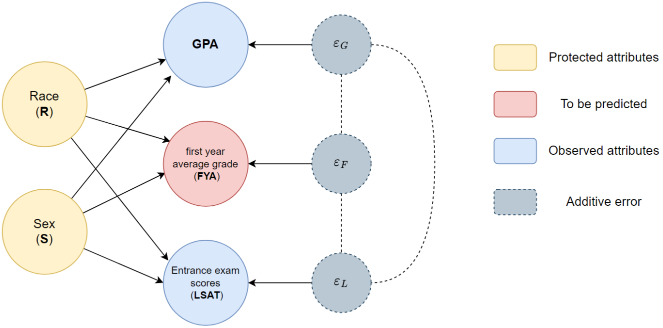

Figure: Causal model with additive error terms independent of protected attributes (may be correlated with one-another)

Results: Fair Add model

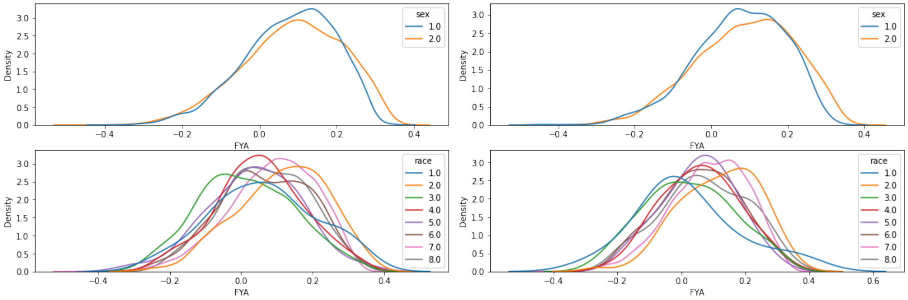

Figure: Fair Add model: Dist. of FYA separated out by Race and Sex: train set (left), test set (right)

Results: RMSE

	Full	Unaware	Fair K	Fair Add
RMSE	0.615	0.671	0.746	0.723

Table: RMSE on test set

- The **Full** model achieves the lowest RMSE as it uses race and sex to more accurately predict FYA.
- The (also unfair) **Unaware** model does not use race and sex therefore it cannot match the RMSE of the Full model.
- The **Fair K** model and **Fair Add** model are counterfactually fair, whose RMSE are slightly higher.

Data split balance

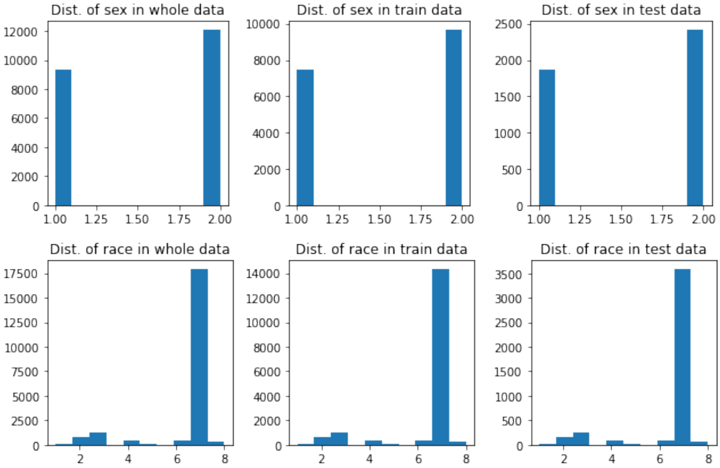

Figure: Dist. of Race and Sex in whole/train/test set

References

- M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. Advances in neural information processing systems, 30, 2017.
- E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep Universal Probabilistic Programming. Journal of Machine Learning Research, 2018
- D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. arXiv preprint arXiv:1912.11554, 2019
- https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage
- https://github.com/mkusner/counterfactual-fairness
- https://github.com/Kaaii/CS7290_Fairness_Eval_Project