Fairness CI in ML

阅读流程

- Abstract + Intro + Conclusion
- Problems
- Solutions
- Eval

Abstract + Intro + Conclusion

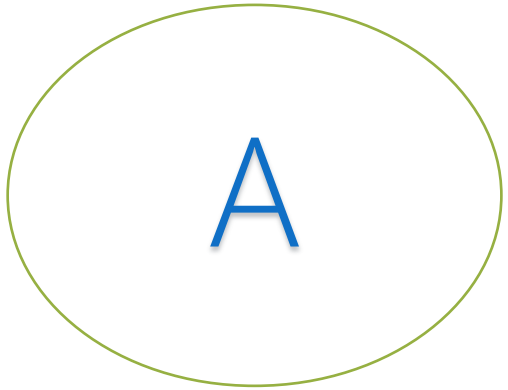
- Abstract
  - 观测数据
  - 使用观测数据中基于protected attributes所带来一些潜在的歧视
  - 也明确了自己研究目标

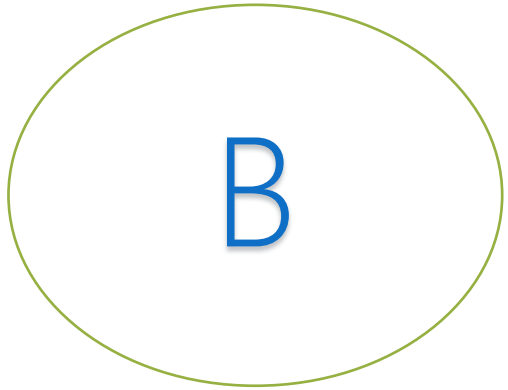
Intro

- Protected attribute
- 观测数据
  - 现实世界中观察、记录和收集的数据。研究者无法直接控制或者操控数据的形成。
- Contribution points
  - Revisiting the two scenarios proposed .
  - algorithms

Related work

- Demographic parity
  - 通俗来说对不同人群的统计数据。





Conclusion

- 框架
- 目标
- resolving variables & proxy variables
- solutions
- limitations to solutions.

Unresolved discrimination

1973年，44%男性学生vs only 35%女性学生被录取。（整体的录取率）

Table 1: Data From Six Largest Departments of 1973 Berkeley Discrimination Case

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

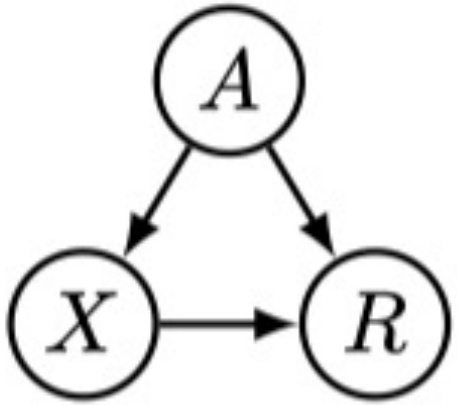


Figure 1: The admission decision R does not only directly depend on gender A, but also on department choice X, which in turn is also affected by gender A.

- A是protected attribute, 代表性别；
- X表示录取的department院系；
- R表示被录取.

unresolved discrimination

- A variable V in a causal graph exhibits unresolved discrimination if there exists a directed path from A to V that is not blocked by a resolving variable and V itself is non-resolving.
- All paths from the protected attribute A to R are problematic, unless they are justified by a resolving variable.

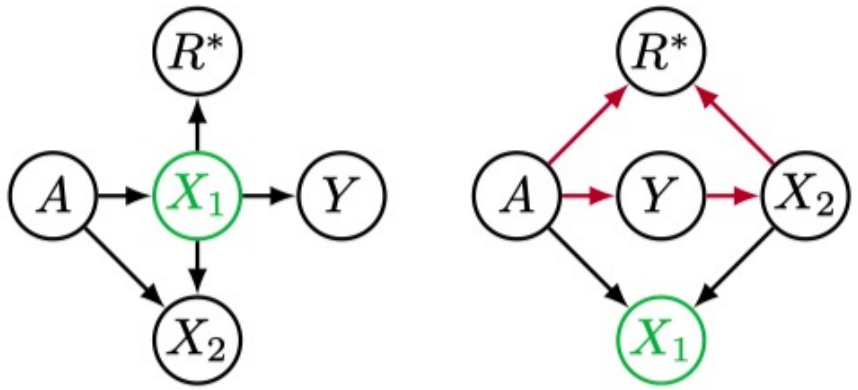
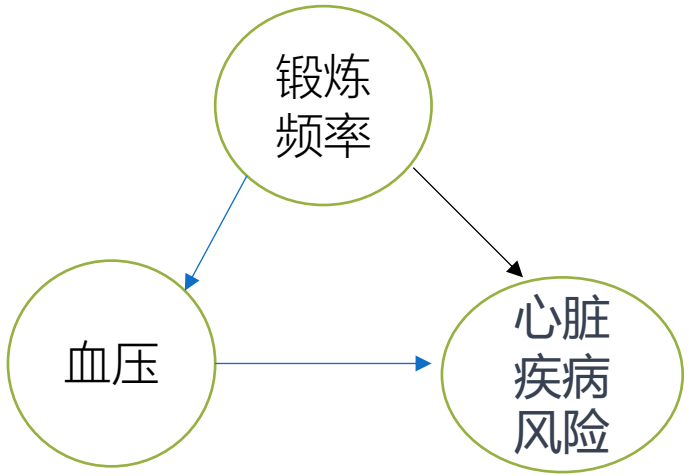


Figure 2: Two graphs that may generate the same joint distribution for the Bayes optimal unconstrained predictor R^*. If X_1 is a resolving variable, R^* exhibits unresolved discrimination in the right graph (along the red paths), but not in the left one.

Proxy Discrimination



HOW TO SOLVE?

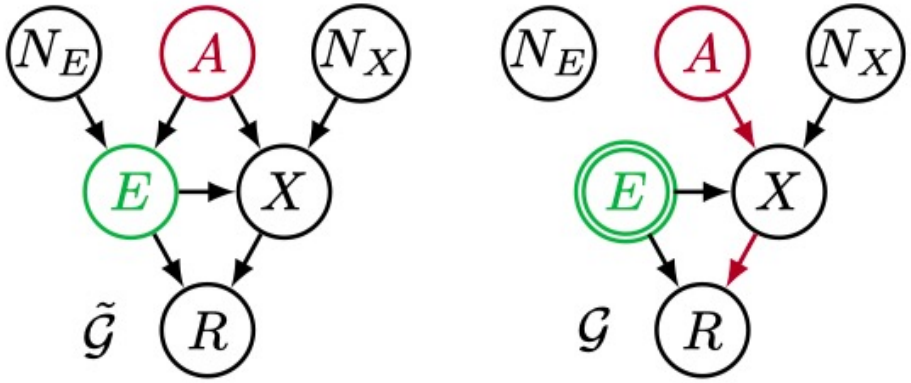


Figure 4: A template graph $\tilde{\mathcal{G}}$ for unresolved discrimination (left) with its intervened version $\mathcal{G}$ (right). While from the skeptical viewpoint we generically do not want A to influence R, we first intervene on E interrupting all paths through E and only cancel the remaining influence on A to R.

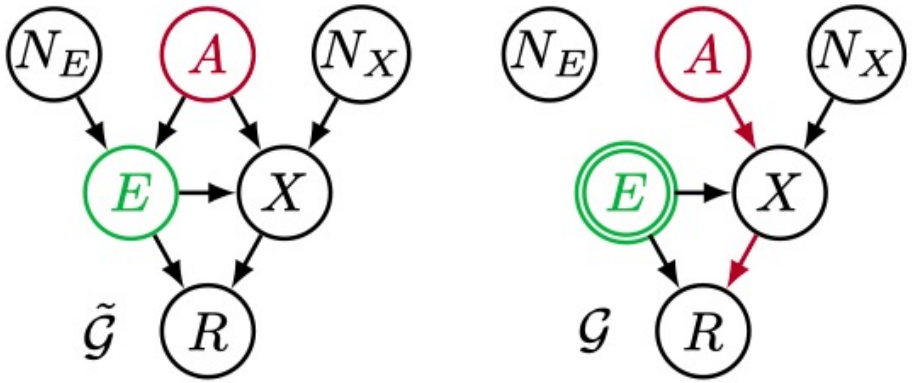


Figure 4: A template graph $\tilde{\mathcal{G}}$ for unresolved discrimination (left) with its intervened version $\mathcal{G}$ (right). While from the skeptical viewpoint we generically do not want A to influence R, we first intervene on E interrupting all paths through E and only cancel the remaining influence on A to R.

1. Intervene on E
   - $E = \alpha_E A + N_E$, => $E = \eta$
   - $X = \alpha_X A + \beta E + N_X$
   - $R_\theta = \lambda_E E + \lambda_X X$
2. By iterative substitution
   - $E = \eta$
   - $X = \alpha_X A + \beta E + N_X$
   - $R_\theta = \lambda_E E + \lambda_X X$...
   - ....
   - $R_\theta = (\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X A + \lambda_X N_X$
3. We now demand the distribution be invariant

$$\mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a + \lambda_X N_X)) = \mathbb{P}((\lambda_E + \lambda_X \beta)\eta + \lambda_X \alpha_X a' + \lambda_X N_X)) \,.$$

Evaluation

- 梳理了机器学习中变量之间带来的一些潜在的歧视
  - 不容易应用
- Future work
  - 对于目前我们所接触到数据，利用这个思路，看能不能develop新的算法去避免数据之间的歧视
  - 这篇论文主要关注了变量之间的线性关系，我们能否尝试应用于非线性关系中给的变量