Assignment 2 - Fairness & Causality

Arathy Bastin (03754963)
Varsha Raveendran(03745012)
Sandhanakrishnan Ravichandran (03755546)

June 9, 2022

Exercise 1

R2 satisfies equal opportunity because both the groups have equal true positive rate. P(R2 = 1|A=0,Y=1) = P(R2 = 1|A=1,Y=1) where A=0 represents Hispanic and A=1 represents White

Exercise 2

The four weights are: w = $\frac{P(expected)}{P(observed)}$
Male loan = $\frac{0.5*0.3}{0.25}$ = 0.6

Male no loan = $\frac{0.5*0.7}{0.3}$ = 1.66

Female loan = $\frac{0.5*0.3}{0.2}$ = 0.75

Female no loan = $\frac{0.5*0.7}{0.25}$ = 1.4

Exercise 3.1

Lifestyle (e.g., smoking) influences whether a woman takes oral contraceptives (OC) and, independently, her risk of breast cancer. Select the causal DAG that represents this scenario

Option a

Exercise 3.2

You believe that aspirin can only reduce the risk of stroke through the reduction of platelet aggregation. Select the causal DAG that represents your belief.

Option b

Exercise 3.3

Which DAG is consistent with the following conclusion? To determine the total effect of aspirin on the risk of stroke, we should adjust for platelet aggregation in our statistical analysis.

Option c. Platelet aggregation is a mediation variable. To understand total effect of aspirin, we need to look at both the ”direct effect” of aspirin on risk of stroke and the indirect effect caused by platelet aggregation.

Exercise 3.4

Given this causal DAG, if we do not adjust for history of heart disease then surgery is expected to be marginally associated with death. a. True b. False

Option b. False

Exercise 3.5

Choose the causal DAGs that show a backdoor path between the variables poverty and tuberculosis.

Option a - closed backdoor path
Option c - open backdoor path

Exercise 3.6

For blocking the back-door path and enable computing P(Y | do(T)) , what needs to be done in a) and b) ?

(a) In this case, X is a collider and blocks the back-door path. We can directly compute P(Y | do(T)).

(b) Here, X is a confounding variable. To enable computing P(Y | do(T)), we need to break independence between T and X by conditioning on X. The experiments are randomized in order to assign T independent of any confound X.

Exercise 4.1

Mitigation by adjusting for race:

Fairness metrics with respect to sex

strategy	Test accuracy	Statistical parity difference	Equal opportunity difference
LR (no mitigation)	66.11%	-0.248	-0.162
LR Unaware	66.41%	-0.115	-0.064
Reweighing	66.64%	-0.224	-0.135
Adversarial Debiasing	65.58%	-0.198	-0.157
Calibrated Equalized Odds	63.99%	-0.221	-0.120

The above table shows that mitigation for race does not help unfairness with respect to sex. The metric values are not close to zero.

Exercise 4.2

Mitigation by adjusting for sex:

Fairness metrics with respect to sex

strategy	Test accuracy	Statistical parity difference	Equal opportunity difference
LR	66.11%	-0.248	-0.162
LR Unaware	66.41%	-0.115	-0.064
Reweighing	66.34%	-0.009	0.041
Adversarial Debiasing	65.13%	-0.258	-0.199
Calibrated Equalized Odds	65.50%	-0.476	-0.314

Fairness metrics with respect to race

strategy	Test accuracy	Statistical parity difference	Equal opportunity difference
LR	66.11%	-0.269	-0.181
LR Unaware	66.41%	-0.115	-0.064
Reweighing	66.34%	-0.259	-0.188
Adversarial Debiasing	65.13%	-0.235	-0.160
Calibrated Equalized Odds	65.50%	-0.257	-0.164

Exercise 4.3

Mitigation by adjusting for sex and race:

Evaluating with respect to sex

strategy	Test accuracy	Statistical parity difference	Equal opportunity difference
Reweighing	66.03%	-0.033	0.017
Adversarial Debiasing	65.58%	-0.179	-0.133

Evaluating with respect to race

strategy	Test accuracy	Statistical parity difference	Equal opportunity difference
Reweighing	66.03%	-0.065	-0.002
Adversarial Debiasing	65.58%	-0.157	-0.092

Exercise 4.4

(Assumed typo in question - results shown for sex and race instead of age since the fetch_compas api returns only sex and race as protected attributes in index)

Summarizing the values obtained from above solutions,

Strategy: Reweighing

Adjusted attribute	Test accuracy	SPD wrt sex	EOD wrt sex	SPD wrt race	EOD wrt race
race (only)	66.64%	-0.224	0.135	-0.070	-0.003
sex (only)	66.34%	-0.009	0.041	-0.259	-0.188
sex and race	66.03%	-0.033	0.017	-0.065	-0.002

Strategy: Adversarial Debiasing

Adjusted attribute	Test accuracy	SPD wrt sex	EOD wrt sex	SPD wrt race	EOD wrt race
race (only)	65.58%	-0.198	-0.157	-0.132	-0.068
sex (only)	65.13%	-0.258	-0.199	-0.235	-0.160
sex and race	65.58%	-0.179	-0.133	-0.157	-0.092

Observations (next page)

1. With the reweighing strategy, the statistical parity and equal opportunity difference adjusting for sex is significantly less than adjusting both. Only a slight difference is seen with adjusting only race in comparison to both. Hence, adjusting attributes separately works for reweighing.

2. With the adversarial debiasing strategy, adjusting for both results in smaller difference in parity compared to adjusting separately. Therefore, it is better to adjust both simultaneously while using this strategy. However, the model is still unfair.