Fall 2021 Capstone Project

Progress Report 2

Causality-Informed Fairness

November 27, 2021

Student Contributors:

Junzhi Ge
Mohammed Aqid Khatkhatay
Oscar Jasklowski
Xue Gu
Yue Wang

In Collaboration With:

Dr. Sanghamitra Dutta
Dr. Naftali Cohen

*Student contributors listed alphabetically

Table of Contents

1. Progress To Date and Goals of Report #2

2. Data Generation
    2.1. Causal Model
    2.2. Data Generating Process (DGP)
    2.3. Fairness of our DGP
    2.4. Demonstrating Unfairness of our DGP
    2.5. Implications

3. Baseline Unfair Models
    3.1. Brief Summary
    3.2. Model Unfairness Evaluation

4. Fair Inference on Outcomes
    4.1. Brief Summary
    4.2. Implementation
    4.3. Evaluation Metrics
        4.3.1. Path Specific Effect (PSE)
    4.3.2. Results

5. Individually Fair Path-Specific Model
    5.1. Brief Summary
    5.2. Implementation
    5.3. Evaluation

6. Conclusion and Next Steps

7. Contributions

9. References

1. Progress To Date and Goals of Report #2

In our previous report, we sought to develop a research framework that would enable us to efficiently explore the literature on causality-based approaches to fairness. Specifically, we explored the following questions:

1. How does a particular fairness technique perform across a variety of metrics?
2. For a given metric, how do various fairness techniques compare?
3. Do certain fairness techniques lend themselves better to certain real-world problems, datasets, or causal models?

After producing our initial report, it became clear that we needed to generate a synthetic dataset that would satisfy the following conditions:

1. The data generating process (DGP) is fully specified on the basis of a causal graph,
2. And, for ease of interpretation, the DGP would have only one sensitive variable.

For this report, we produce and work with such a synthetic dataset. By using a fully specified DGP and a corresponding dataset (rather than a real world dataset), we have full control over variables and directed paths that govern fairness. As a result:

1. We can guarantee that our DGP is inherently unfair (with respect to only 1 variable),
2. We can subsequently explore the effectiveness of a variety of fair-inference techniques on mitigating unfairness, which is the goal of this project.

In the following sections, outline our progress on implementing an inherently unfair DGP and evaluating a variety of fair inference techniques on the dataset.

2. Data Generation

2.1. Causal Model

Consider a hypothetical situation where candidates are applying for a job in a warehouse that requires, among other qualifications, physical strength and good communication. Suppose our goal is to predict whether each candidate should be hired **(Y)**, and we have the following observable variables at our disposal:

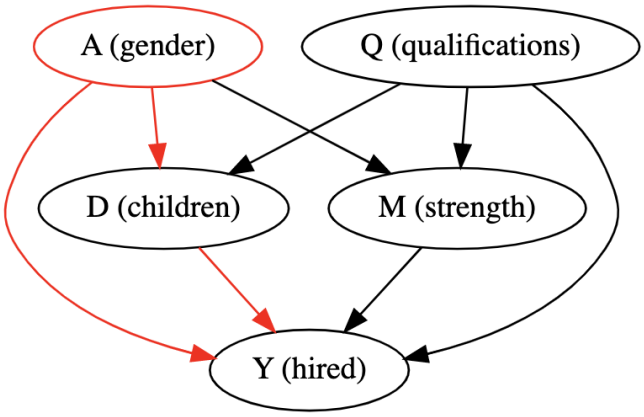


Figure 1: A diagram of the data generating process (DGP) we will use for this report. This corresponds to a fictional hiring decision process, having data for various observable candidate attributes.

- **A:** Gender
  - A *sensitive variable*; it should not influence employment outcome directly
- **Q:** Qualifications
  - A measure of whether they are generally qualified for the job
  - Including, but not limited to, whether they have relevant employment experience
- **D:** Number of Children the candidate has
  - This is *not a sensitive variable on its own*, but it is a descendant of gender (i.e. in this fictional example, we assume female applicants tend to have more children). ***Therefore, more advanced techniques will account for the influence of gender along the A→D→Y path.***
  - This feature can be useful for inference, though, because it could be influenced by unobserved features, like communication skills, which are an asset to the job.
- **M:** Physical Strength
  - A candidate's physical strength is an asset to the job, so despite being a descendant of gender, *this is not a sensitive variable*.
  - Furthermore, *in this context*, not hiring someone on the basis of not meeting physical *would not be considered unfair*, even if that decision was indirectly made based on gender. ***Therefore, A→M→Y is <u>not</u> considered an unfair path.***
- **U_Q, U_D, U_A** (not pictured)**:** Noise Terms
  - To provide stochasticity, Q, D, and A variables descend from noise terms.

2.2. Data Generating Process (DGP)

The data generating process is specified by the following probability distributions and parameters:

$$A = U_A,\ U_A \sim Bernoulli\,(0.6)$$

$$Q = \lfloor U_Q \rfloor,\ U_Q \sim N\,(2,\ 5^2)$$

$$D = A + \lfloor 0.5 * Q * U_D \rfloor,\ U_D \sim TrN(2, 1^2,\ 0.1,\ 3)$$

$$M = 3A + 0.4 * Q * U_M,\ U_M \sim TrN(3, 2^2,\ 0.1,\ 3)$$

$$Y = h(A, Q, D, M)$$

$$P(Y = 1 | A,\ Q,\ M) = Bernoulli(\varsigma(-10 + 5A + Q + D + M))$$

Where h(), called the "hiring rule", is a Bernoulli random variables on the result of the sigmoid function ς. ς is the standard sigmoid function, $\varsigma(x) = 1/(1 + exp(-x))$ and TrN are truncated normal distributions. We are also implicitly assuming that all noise terms U_A, U_Q, U_D, U_M are non-negative. Namely, any negative noise terms will be set to 0.

Some important relationships worth highlighting:

- On average, 60% of the applicants in this dataset will be male
- Number of children is positively influenced by gender AND qualifications
- Physical strength is positively influenced by gender AND qualifications

Finally, hiring is *positively influenced* by the following observed variables via *direct relationships*:

- Gender (A)
- Qualifications (Q)
- Children (D)
- Physical strength (M)

2.3. Fairness of our DGP

As per Lemma 1 in Kusner et al., the DGP above is unfair because the protected attribute gender has both direct impact on the final hiring decision and indirect impact through its descendents D and M. However, if we only use Q to build prediction models, that would eliminate dependence on the unfair factor, resulting in a counterfactually fair classifier under the following definition:

$$P(\widehat{Y}_{A=a'}(Q) \mid X = x,\ A = a) = P(\widehat{Y}_{A=a}(Q) \mid X = x,\ A = a)$$

Such a classifier is learned and explored in section 3.

2.4. Demonstrating Unfairness of our DGP

We created the following algorithm to demonstrate unfairness:

1. First, create FACTUAL dataset:
    a. Generate noise terms (U_Q, U_D, U_M)
    b. Generate A
    c. Generate observable variables (Q, D, M)

    d. Generate Y using observed variables

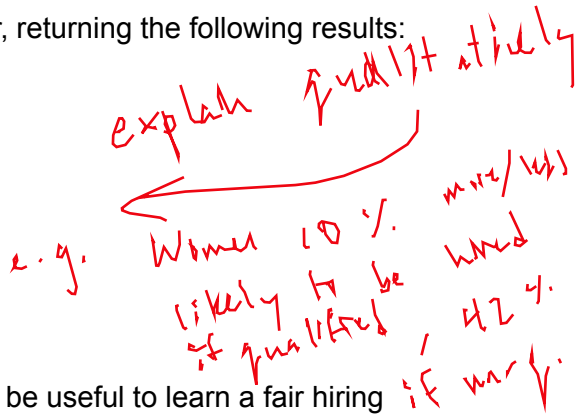
2. Then, create COUNTERFACTUAL dataset:
    a. Use the same error terms as before
    b. Fix A (i.e. convert all male to female)
    c. Again, generate observable variables (Q, D, M)
    d. Again, generate Y using observed variables

3. Evaluate counterfactual fairness:
    a. Merge the two datasets on index
    b. Keep on the examples where male was changed to female
    c. For each level of Q (qualified, unqualified), compute the equation in section 4.

As you might expect, our dataset was NOT counterfactually fair, returning the following results:

Table 1: Proof of unfairness for synthetic dataset.

Counterfactual Fairness for each stratum of qualifications:	
Qualified individuals	0.102
Un-qualified individuals	0.427



2.5. Implications

Having proven that our data generating process is unfair, it can be useful to learn a fair hiring rule, and based on that fair hiring rule, generate an inherently fair dataset (we could trivially prove its fairness using the algorithm described above). This will be a worthwhile exercise for our final report, as we'll want to evaluate our fairness techniques on fair (not just unfair datasets). Ideally, we should not see deterioration in prediction accuracy when using our fair techniques on fair datasets.

3. Baseline Unfair Models

3.1. Brief Summary

For the synthetic dataset, we performed evaluation on their fairness level. The full model involves all attributes (A, D, Q, and M) where A is the protected attribute that could represent sex. D and M are descendants of A, attributes affected by gender. Q is totally not influenced by A. The full model means to emulate the situation where potential bias is not treated.

On the other hand, the unaware model, pointed out in the "Counterfactual Fairness"[2] paper, meant to achieve a certain degree of fairness by not involving the protected attribute in the model explicitly. However, the descendants of A are still included to make predictions.

3.2. Model Unfairness Evaluation

To evaluate the fairness of different models, we regenerated the counterfactual data (counterfactual A, D and M, and original Q) and compared the probabilities of being predicted to get a job or not to the value achieved by factual data. The unfairness is defined as counterfactual probability - factual probability:

$$P(\widehat{Y}_{A=a'} \mid X = x, A = a) - P(\widehat{Y}_{A=a} \mid X = x, A = a)$$

We built different classification models and below is the summary of results:

Table 2: Baseline models fairness comparison

Full Model			Models	Unaware Model		
Unfairness Level	Var A	Accuracy		Accuracy	Var A	Unfairness Level
0.264524	0	**0.9970**	Logistic Regression	**0.9745**	0	0.239802
-0.277917	1				1	-0.237616
0.271941	0	**0.9990**	Decision Tree	**0.9940**	0	0.265761
-0.281276	1				1	-0.272041
0.254636	0	**0.9900**	SVM	**0.9825**	0	0.242274
-0.265323	1				1	-0.244333

From the table 1, above we can see, by giving up the sensitive variable, the unaware models lose a small degree of accuracy on the test set for all logistic regression, Decision Tree, and SVM. The unfairness level is the difference in probability, therefore, ideally lower the better. Comparing the full model and unaware model, we confirmed that under our causal relationship, unaware models can bring up fairness. Moreover, we can also observe that different machine learning algorithms could result in different fairness, with the full model being the fairest under SVM while the unaware model being the fairest under logistic regression.

Table 3: SVM Fairness based on Q=1 (Qualified Individuals)

Full Model			Models	Unaware Model		
Unfairness Level	Var A	Accuracy		Accuracy	Var A	Unfairness Level
0.343137	0	**0.9900**	SVM	**0.9825**	0	0.323529
-0.371972	1				1	-0.33737

Note: Modified Equation $P(\widehat{Y}_{A=a'} \mid X = x, A = a, Q = 1) - P(\widehat{Y}_{A=a} \mid X = x, A = a, Q = 1)$

Taking the SVM model as an example, if we condition on qualified individuals, we can observe that the unaware model also achieves some degree of fairness by shrinking the difference between counterfactual and factual probability of getting fired.

4. Fair Inference on Outcomes

4.1. Brief Summary

In the paper [1], by Razieh Nabi, Ilya Shpitser, they use a reverse causal approach to move from an unfair world to a fair world. They train on two models, one is constrained, and the other is not constrained. The PSE in the unconstrained model was calculated on the adult data set. The PSE value implied that, the odds of having a high income were more than thrice for a female with the same marital status and sex of her male counterpart in a counterfactual world.

The authors solved the constrained problem by restricting the PSE, to lie between 0.95 and 1.05. They were able to boost the accuracy of the constrained model. We will try to implement this technique for our synthetically generated data set.

4.2. Implementation

We modify the R code provided by Razieh Nabi and Ilya Shpitser in their appendix, to suit our synthetically generated dataset. We also modify the ACE and PCE evaluation metrics in order to suit our causal model. Furthermore, we tabulate the contingency table with both constraints and no constraints and also compute the accuracies for both using linear and logistic regression classifiers.

4.3. Evaluation Metrics

4.3.1. Path Specific Effect (PSE)

Path Specific Effects need to be considered, since Discrimination is basically the presence of effect along unfair causal pathways. Along a given path in the graph, all nodes behave as if A = a. Similarly, along all other paths, nodes behave as if A = a'.
It can be formulated as follows:

$$PSE = \Sigma_{C,M,W} E[Y|\, a',\, W,\, M,\, C] * p(W \mid a,\, M,\, C) * p(M|a',\, C) * p(C)$$

4.3.2. Results

The following tables summarize our results using PSE.

Table 4: Fair Inference on outcomes evaluation

Metric	Model	Value
PSE (Non Constrained)	Linear Regression	189.707
PSE (Constrained)	Logistic Regression	0.952
Accuracy (Non Constrained)	Linear Regression	0.887
Accuracy (Constrained)	Logistic Regression	0.929

Table 5: Contingency Table Fair Inference on outcomes

Prediction	True Label = 0	True Label = 1
Prediction = 0	434	67
Prediction = 1	46	453

From Table 1. For the linear no constrained model we get a PSE Value of 189 which means there is unfairness present in the model. As the table shows, the PSE in our model is 0.952, which means that the odds of being hired is a little lower than 1 for a female, if the number of children and strength are the same as if she had the sex value of a male in a counterfactual world. Also, the accuracy for the model is 87.6%.

5. Individually Fair Path-Specific Model

5.1. Brief Summary

The path-specific fairness technique, which optimizes the probability of individual unfairness (PIU), has been introduced in the previous report, based on [3]. It guarantees individual level counterfactual fairness without imposing impractical assumptions. While it is impossible to directly minimize PIU as we cannot observe $A = 0$ and $A = 1$ at the same time, the paper in [3] defines a feasible upper bound, estimated from the data, that can be easily minimized. In our case, this technique allows us to learn a more fair classifier, but at a cost of accuracy.

explain notation.

5.2. Implementation

With the fully specified, synthetic data set, we are able to generate potential outcomes $Y_{A\leftarrow 0} = h_\theta(0, Q, D(0), M(0))$ and $Y_{A\leftarrow 1\,|\pi} = h_\theta(1, Q, D(1), M)$. Specifically, $M(0) = 0.4 * Q * U_M$, $D(0) = [0.5 * Q * U_D]$, $D(1) = 1 + [0.5 * Q * U_D]$ with U_M, U_D, Q fixed, and $\pi = \{A \to Y, A \to D \to Y\}$ is the set of unfair pathways.

In addition, estimator of marginal distributions $P(Y_{A\leftarrow 0} = 1)$ and $P(Y_{A\leftarrow 1\,|\pi} = 1)$ can be used to estimate the upper bound

$$P^I(Y_{A\leftarrow 0} \neq Y_{A\leftarrow 1\,|\pi}) = P(Y_{A\leftarrow 1\,|\pi} = 1)(1 - P(Y_{A\leftarrow 0} = 1)) + P(Y_{A\leftarrow 0} = 1)(1 - P(Y_{A\leftarrow 1\,|\pi} = 1))$$

where P^I is the independent joint distribution by the two marginal distributions.

With the full synthetic data set specified and the estimated upper bounds, we are able to feed the complete synthetic data into the algorithm, which is implemented officially by the authors in [3].

5.3. Evaluation

The model achieves a PIU of 0.17 by minimizing the upper bound $2P^{I}(Y_{A\leftarrow 0} \neq Y_{A\leftarrow 1\,|\pi})$. The prediction accuracy during test time is 89.6% with a more detailed classification report below:

Table 6: Classification outcomes using path-specific fairness with PIU in test time

Prediction	True Label = 0	True Label = 1
Prediction = 0	418	19
Prediction = 1	85	478

We can see that the accuracy is close to the one achieved in the FIO method, lower than both of the unfair baseline models.

6. Conclusion and Next Steps

So far, we have implemented and evaluated two baseline unfair models, namely full and unaware models, and two “fair” models: FIO and individual level path-specific model. The performance in classification accuracies during test time and the counterfactual fairness level are summarized below in table 6. We expect to see that the “fair” models perform slightly worse in prediction power, but have better fairness scores. Our experiments, so far, have yet to show that the results are consistent with our expectations.

Table 7: Comparisons between full, unawareness FIO and Individual level path-specific models

	Test Accuracy (percentage)	Counterfactual fairness level
Full model	99.9 (best)	0.553
Unaware model	99.4 (best)	0.537
FIO	88.7	N/A
Individual level path-specific	89.6	N/A

Some of our next steps are:

1. For the synthetic dataset, we will also conduct the latent variable inference to get an estimated distribution of Q and compare the result with the true distribution. This is to prove our inference technique from report 1 is correct.
2. We need to further improve our implementation of the counterfactual fairness and ETT evaluation metrics for our two fair models.
3. Implement our fairness estimation techniques to a real world dataset based on an assumed causal model.

7. Contributions

Index	Task	Owner
1.	Progress To Date and Goals of Report #2	mk4427, ovj2101
2.	Data Generation	xg2353, ovj2101
3.	Baseline Unfair Models	yw3576
4.	Fair Inference on Outcomes	mk4427, jg4281
5.	Path-Specific Fairness	xg2353
6.	Conclusion and Next Steps	Entire Team
7.	Contributions	Entire Team
8.	References	Entire Team

9. References

[1] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In AAAI, pages 1931–1940, 2018. https: //github.com/raziehna/fair-inference-on-outcomes.

[2] Matt J Kusner et al. “Counterfactual fairness”. In: arXiv preprint arXiv:1703.06856 (2017).

[3] Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. arXiv preprint arXiv:2002.06746, 2020.