CausalFairnessInAction: An Open Source Python Library for Causal Fairness Analysis

Kriti Mahajan, Amazon, kritimhj@amazon.com

Forthcoming at: https://github.com/amazon-science/causal-fairness-in-action

Counterfactual Fairness available at:

https://www.pywhy.org/dowhy/main/example_notebooks/counterfactual_fairness_dowhy.html

Motivation & Contribution

The Problem: As machine learning enters high-stakes domains, assessing fairness becomes vital—but the typically used statistical fairness metrics have a key limitation: They are **associations(conditional probabilities) thus, they** can state **what the observed disparity is *but not why* it exists.**

Causal Fairness metrics solve this by using Structural Causal Models (SCMs) to uncover generating mechanisms, but have limited adoption due to **technical** & **computational complexity.**

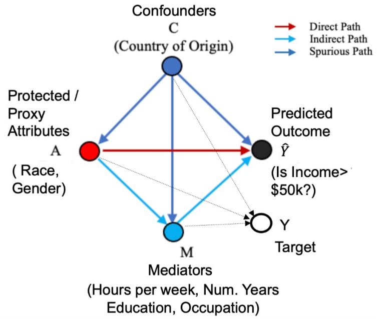
The Solution: CausalFairnessInAction, the first open-source Python package for computing diverse causal fairness metrics, enabling actionable audits by decomposing statistical disparities into causal components.

- **Practical:** applicable across classification and regression tasks ; designed to work with minimal identifiability constraints; doesn't require fully specified SCMs.
- **Comprehensive:** Computes metrics at both **group** & **individual** levels; supports **intersectional analysis**.
- **Efficient:** Optimized for scalability using Gaussian Mixture Models, parallelization to reduce latency

Methodology & Framework

The `CausalFairnessDecomposition` class is built on the standard fairness model [1]:

Figure 1:Standard Fairness Model (Ex: Adult Income Dataset)



Three Implemented Metrics and Methods

`analyse_mean_difference` ➔ **Implements Counterfactual Effects[1]**

Query : What would the disadvantaged (advantaged) **group's acceptance rate** be if they had the identity (A), mediating characteristics (M), or confounding characteristics (C) of the advantaged (disadvantaged) group?

Supported Decompositions: Direct, Indirect, Spurious

`analyse_equalized_odds` ➔ **Implements Counterfactual Equalized Odds[2]**

Query : What would the disadvantaged (advantaged) **group's error rate** be if they had A, M or C of the advantaged (disadvantaged) group?

Supported Decompositions: Direct, Spurious

`analyse_counterfactual_fairness` ➔ **Implements Counterfactual Fairness[3]**

Query : What would the disadvantaged (advantaged) **individual's predicted Y** be if they had the A, M, and C of the advantaged (disadvantaged) group?

Supported Decompositions: N/A

Table 1: Pseudo-Algorithms for Causal Fairness Metrics

`analyse_mean_difference`	`analyse_equalized_odds`	`analyse_counterfactual_fairness`
Inputs: D, A, M, C, a_0, a_1, y	**Inputs:** $D, A, C, a_0, a_1, y, \hat{f}$	**Inputs:** A, M, C, a_0, a_1, DAG
1. For each $(m, c) \in D$: - Compute: $\mathbb{E}(Y = y \mid a_0, m, c)$ - Compute: $\mathbb{E}(Y = y \mid a_1, m, c)$ **2.** Estimate via GMM: $P(m \mid a_0, c), P(m \mid a_1, c)$ $P(c \mid a_0), P(c \mid a_1)$ **3.** Combine expectations and probabilities to compute the counterfactual effects	**1.** For each $c_j \in D$: - Predict: $\hat{f}(c_j, a_0), \hat{f}(c_j, a_1)$ - Obtain: $P(\hat{y}_{a_0,c_j}), P(\hat{y}_{a_1,c_j})$ **2.** Estimate via GMM: $P(c \mid a_0), P(c \mid a_1)$ **3.** Combine predictions and probabilities to compute the Cft-EO	**1.** Fit SCM using DAG and dataset D **2.** For each individual $i \in D$: - Get A_{obs} (observed) and A_{cf} (counterfactual) - Sample from SCM under: $do(A = A_{obs}) \Rightarrow D_{obs}$ $do(A = A_{cf}) \Rightarrow D_{cf}$ - Predict: $\hat{f}(D_{obs}), \hat{f}(D_{cf})$ - Check: $Y_{obs} \neq Y_{cf}$

Code block 1: Example API Call Applied To The Adult Income Dataset

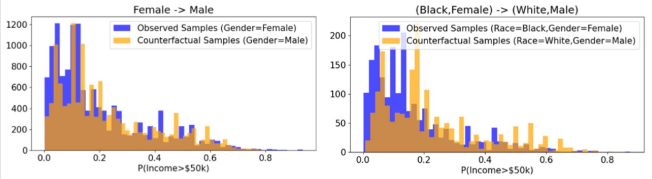
```
cfd = CausalFairnessDecomposition(**{"X": X_train,
                                     "y_true" : y_train.values,
                                     "y_pred": X_train["prediction"],
                                     "model": trained_logistic_regression_classifier,
                                     "protected_attr": ["Sex_Female"],
                                     "mediators": ([x for x in X_train.columns if "Occupation_" in x]
                                                   + ["EducationNum"]
                                                   + ["HoursPerWeek"]),
                                     "confounders": [x for x in X_train.columns if "Country_" in x],
                                     "yi": 1,
                                     "advantage_group": 0,
                                     "disadvantage_group": 1,
                                     "continuous": False})
mean_diff_decomposition = cfd.analyse_mean_difference(how='decompose')
fig_md, ax_md = plot_mean_diff_waterfall(mean_diff_decomposition,mean_difference)
```

Application To Benchmark Datasets

We benchmarked the library on 3 datasets : **Adult Income**, **COMPAS**, and **LSAC**

- **Direct discrimination is the primary contributor** to mean difference and equalized odds across all 3 datasets
- The classifier for Adult Income, COMPAS is not counterfactually fair but is counterfactually fair for LSAC i.e. **group fairness can differ from individual fairness**
- **Intersectional Analysis (Race x Sex) worsens direct discrimination** across all three datasets

Figure 2: Counterfactual Fairness Plots



A. Adult Income: Gender

B. Adult Income: Gender x Race

Limitations

- **Lack of identifiability can limit analysis:** Ex - in the Adult Income dataset, identifiability issues prevent the causal decomposition of equalized odds

Conclusion & Future Work

- **Actionable:** Provides specific targets for bias mitigation (e.g., fixing the 16.5% direct effect in Adult Income)
- **Future:** Extending the package to include remediation algorithms and sensitivity analysis.

References

1 Plecko, D. &; Bareinboim, E.,2024 . "Causal fairness analysis.". In: Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 1–238

2 Zhang, J. &; Bareinboim, E, "Equality of opportunity in classification: A Causal approach. ." In: Advances in Neural Information Processing Systems.

3. Kusner, M.J. et al., 2017." Counterfactual fairness" In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems

Table 2: CausalFairnessInAction Benchmarking

Dataset	Protected Attribute	Mean Difference	FNR	FPR	$\mathrm{DE}_a^{\mathrm{sym}}(y\|a)$	$\mathrm{IE}_a^{\mathrm{sym}}(y\|a)$	$\mathrm{SE}_{a_0,a_1}(y\|a)$	ER^d	ER^i	ER^s	Counterfactual Fairness
Adult Income	Gender	0.203	0.410	-0.104	0.165	0.039	0.000	0.000	0.000	0.000	-0.031
Adult Income	Intersectional	0.221	0.445	-0.115	0.152	0.069	0.000	0.000	0.000	0.000	-0.068
COMPAS	Race (Black)	0.326	-0.310 (-42)	-0.253 (-0.41)	0.154	0.071	0.101	FPR: -0.297, FNR: -0.265	0	FPR: 0.113, FNR: 0.162	0.055
COMPAS	Intersectional	0.620	-0.620	-0.518	0.513	0.081	0.027	-	-	-	0.640
LSAC	Race (Black)	0.978	-	-	0.554	0.429	0.000	-	-	-	0.001
LSAC	Intersectional	0.990	-	-	0.531	0.458	0.000	-	-	-	-0.007