| U | A | X | Pr |
|---|---|---|---|
| -1 | -1 | 0 | 0.32 |
| 0 | -1 | 0 | 0.24 |
| 1 | -1 | 0 | 0.24 |
| -1 | 1 | 0 | 0.08 |
| 0 | 1 | 1 | 0.06 |
| 1 | 1 | 1 | 0.06 |

Table 1: Joint Distribution

Suppose an causal model consists of $U, X, A$. The prior distribution of $U$ is

$$P(U=-1)=0.4 \quad P(U=0)=0.3 \quad P(U=1)=0.3$$

The distribution of $A$ is

$$P(A=-1)=0.8 \quad P(A=1)=0.2 \tag{1}$$

$X$ is determined by $U$ and $A$ in this way:

$$X=\begin{cases}1, & \text{if } U+A>1\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

Then we can have the joint distribution of $U, A, X$ as the table1. So for the observed data, we only have $A=-1, X=0$, $A=1, X=0$ and $A=1, X=1$.

From the definition of counterfactual fairness, we know that for every given $x$ and $a$,

$$h=\mathbb{E}_{U\sim P(U|X=x,A=a)}[\frac{X+\check{X}}{2}] \tag{3}$$

is a counterfactual fair feature. Now for data $A=-1, X=0$, the posterior distribution of $U$ is

$$P(U=-1)=0.4 \quad P(U=0)=0.3 \quad P(U=1)=0.3$$

When $U=-1$, $\frac{X+\check{x}}{2}=0+0=0$. When $U=0$, $\frac{X+\check{x}}{2}=\frac{0+1}{=}0.5$. When $U=1$, $\frac{X+\check{x}}{2}=\frac{0+1}{=}0.5$. So $h=0.3$.

For $A=1, X=0$, we know that the posterior distribution of $U$ is $P(U=-1)=1$. So $h=0$. And when $A=1, X=1$, the posterior distribution of $U$ is $P(U=0)=P(U=1)=0.5$. So $h=0.5$.

That means when $A=-1$, $P(H=0.3)=1$. When $A=1$, $P(H=0)=0.4, P(H=1)=0.6$. DP not hold.