Algorithmic Recourse under incomplete causal graph

Algorithmic Recourse: from Counterfactual Explanations to Interventions

ACM 2021

Algorithmic Recourse

The systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios.

Problem Formulation

$$x^{*\mathrm{CFE}} \in \underset{x}{argmin} \quad dist(x, x^F) \quad s.t.\ h(x) \neq h(x^F), x \in \mathcal{P}$$

Contd.

Ustun et al.

$$\delta^* \in \underset{\delta}{argmin} \quad cost(\delta; x^F) \quad s.t.\ h(x^{\mathrm{CFE}}) \neq h(x^{\mathrm{F}}),$$
$$x^{\mathrm{CFE}} = x^{\mathrm{F}} + \delta,$$
$$x^{\mathrm{CFE}} \in \mathcal{P},\ \delta \in \mathcal{F}$$

Why does the above formulation fail?

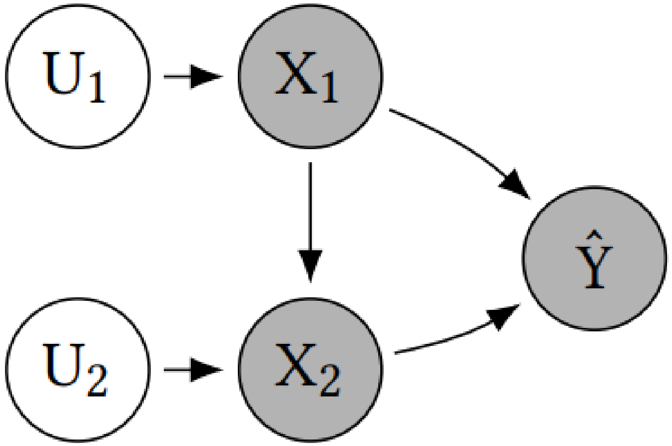
- Example: Consider an example where an individual has an annual salary of \$75,000 and an account balance \$25,000. Say the model is

$$h = sgn(X_1 + 5X_2 - \$225,000)$$

- The counterfactual explanation could be
    i. Annual salary: \$100,000 or
    ii. Bank Balance: \$30,000

In a world where home-seekers save %30 of their salary, a salary increase of 14% would result in positive decision of loan-granting algorithm.

Structural Causal Model



$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= f_2(X_1) + U_2 \end{aligned}\right\} \mathcal{M}$$

$$\hat{Y} = h(X_1, X_2)$$

Proposition-1

A CFE-based action, $\mathbf{A}^{CFE}$, where $I = \{i | \delta_i^* \neq 0\}$, performed by individual x^F, in general results in the structural counterfactual, $x^{SCF} = x^{*CFE} := x^F + \delta^*$, if and only if, the set of descendants of the acted upon variables, determined by I, is the empty set.

Corollary-1

If the true world $\mathcal{M}$ is independent, i.e. all the observed features are root-nodes, then CFE-based actions always guarantee recourse.

Causal Persepective

$$
\begin{aligned}
\mathbf{A}^* \in \underset{A}{argmin} \quad & cost(\mathbf{A}; x^F) \\
s.t. \quad & h(x^{\mathrm{SCF}}) \neq h(x^{\mathrm{F}}), \\
& x^{\mathrm{CFE}} = \mathbb{F}_{\mathbf{A}}(\mathbb{F}^{-\mathbb{1}}(\mathbb{x}^{\mathrm{F}})) \qquad x^{\mathrm{CFE}} \in \mathcal{P},\ \mathbf{A} \in \mathcal{F}
\end{aligned}
$$

Proposition-2

Given an individual x^F observed in world $\mathcal{M}$, a family of feasible actions $\mathcal{F}$. Assume that there exists CFE-based actions $A^{CFE} \in \mathcal{F}$ that achieves recourse, i.e., $h(x^F) \neq h(x^{*CFE})$. Then, $cost(A^*; x^F) \leq cost(A^{CFE}; x^F)$

Algorithm

1. **Abduction:** uniquely determines the values of all exogenous variables.
2. **Action:** modify the SCM according to the hypothetical interventions. ($\mathbb{F}_A$)
3. **Prediction:** Determine the values of of all endogenous variables.

The assignment of structural counterfactual values can generally be written as

$$\begin{aligned} x_i^{\mathrm{SCF}} =& [i \in I] \cdot (x_i^{\mathrm{F}} + \delta_i) \\ &+ [i \notin I] \cdot (x_i^{\mathrm{F}} + f_i(pa_i^{\mathrm{SCF}}) + f_i(pa_i^{\mathrm{F}})) \end{aligned}$$