From Counterfactuals to Interventions (Recourse through Minimal Causal Interventions)

Jorge Luiz Franco
jorge.luiz@usp.br

April 1, 2024

Abstract

This GSoC proposal aims to extend the `CounterfactualExplanations.jl` package by incorporating a mathematical foundation for algorithmic recourse through minimal interventions, based on causal reasoning principles. It seeks to enable actionable paths from counterfactual explanations, bridging the gap between theoretical machine learning decision processes and practical user-driven changes.

Contents

1 Introduction

Predictive models significantly impact high-stake decisions, requiring not only interpretability but actionable pathways for altering undesirable outcomes. Counterfactual explanations highlight possible changes to alter model decisions. This project suggests enhancing `CounterfactualExplanations.jl` with a module for generating minimal interventions grounded in causal models.

1.1 Mathematical Background

We will build upon the groundwork of structural causal models (SCMs) and counterfactual reasoning to establish our intervention strategies.

1.1.1 Structural Causal Models (SCMs)

An SCM is defined as follows, capturing the relations between observed variables X and unobserved variables U:

$$X = f(U, X_{pa}), \tag{1}$$
$$U \sim P(U), \tag{2}$$

where X_{pa} represents the parent variables of X, and $P(U)$ is the probability distribution of U.

1.1.2 Counterfactual Reasoning

Counterfactual reasoning assesses what the outcome Y would have been under a different circumstance $X = x'$:

$$Y_{x'} = f_Y(x', U_X = u_{X|Y=y}), \tag{3}$$

where $u_{X|Y=y}$ is the instantiation of U_X leading to $Y = y$ when $X = x'$.

2 Objective

The objective is to implement mathematical algorithms translating counterfactual insights into actionable minimal interventions, thus providing explicit strategies for users to influence predictive outcomes.

3 Methodology

We aim to find a vector of interventions δ^* that minimally perturbs the features to achieve a desired outcome, adhering to:

$$\delta^* = \arg\min_{\delta} C(\delta, X) \text{ s.t. } h(X + \delta) = y', \tag{4}$$

where $C(\delta, X)$ encapsulates the cost of altering X by δ to realize y'.

4 Julia Implementation

Consider a pseudo-Julia implementation sketch provided for the minimal intervention algorithm; in actual LaTeX with basic formatting, it would look like this:

```
function minimal_intervention(model, X, y_desired)
    # Define the objective function
    objective = delta -> intervention_cost(delta, X)
    # Define the constraint for achieving the desired outcome
    constraint = delta -> model(X + delta) == y_desired
    # Solve for minimal delta
    delta_star = optimize(objective, constraint, initial_guess(delta))
    return delta_star
end
```

5 Project Plan and Timeline

- **Weeks 1-4:** Mathematical formulation and Julia prototyping.
- **Weeks 5-8:** Development and testing of the minimal intervention algorithm.
- **Weeks 9-12:** Refinement, benchmarking, and comprehensive documentation.

6 Conclusion

Enriching `CounterfactualExplanations.jl` with capabilities for causal interventions through minimal adjustments promises vast improvements in machine learning model interpretability and navigability, handing users a practically applicable toolset for modulating and understanding model predictions.

References

[1] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera, "Algorithmic Recourse: from Counterfactual Explanations to Interventions," *arXiv preprint arXiv:2002.06278*, 2020.