Introduction

In section Section

Methodological background

Counterfactual search happens in the feature space: we are interested in understanding how we need to change individual attributes in order to change the model output to a desired value or label [**?**]. Typically the underlying methodology is presented in the context of binary classification: $M : \mathcal{X} \mapsto y$ where and $y \in \{0, 1\}$. Let $t = 1$ be the target class and let $\overline{x}$ denote the factual feature vector of some individual outside of the target class, so $\overline{y} = M(\overline{x}) = 0$. Then the counterfactual search objective originally proposed by [**?**] is as follows

$$\min_{\underline{x} \in \mathcal{X}} h(\underline{x}) \quad \text{s. t.} \quad M(\underline{x}) = t \tag{1}$$

where $h(\cdot)$ quantifies how complex or costly it is to go from the factual $\overline{x}$ to the counterfactual $\underline{x}$. To simplify things we can restate this constrained objective (Equation 1) as the following unconstrained and differentiable problem:

$$\underline{x} = \arg\min_{\underline{x}} \ell(M(\underline{x}), t) + \lambda h(\underline{x}) \tag{2}$$

Here ℓ denotes some loss function targeting the deviation between the target label and the predicted label and λ governs the stength of the complexity penalty. Provided we have gradient access for the black-box model M the solution to this problem (Equation 2) can be found through gradient descent. This generic framework lays the foundation for most state-of-the-art approaches to counterfactual search and is also used as the baseline approach in `CounterfactualExplanations`.

That being said, numerous extensions of this simple approach have been developed since counterfactual explanations were first proposed in 2017 (see [**?**] and [**?**] for surveys). The various approaches largely differ in how they define the complexity penalty. In [**?**], for example, $h(\cdot)$ is defined in terms of the Manhattan distance between factual and counterfactual feature values. While this is an intuitive choice, it is too simple to address many of the desirable properties of effective counterfactual explanations that have been set out. These desiderata include:

Closeness: the average distance between factual and counterfactual features should be small ([**?**]).

Actionability: the proposed feature perturbation should actually be actionable ([**?**], [**?**]).

Plausibility: the counterfactual explanation should be plausible to a human ([**?**]).

Unambiguity: a human should have no trouble assigning a label to the counterfactual ([**?**]).

Sparsity: the counterfactual explanation should involve as few individual feature changes as possible.

Robustness: the counterfactual explanation should be robust to domain and model shifts ([**?**]).

Diversity: ideally multiple diverse counterfactual explanations should be provided ([**?**]).

Causality: counterfactual explanations reflect the structual causal model underlying the data generating process ([**?**],[**?**]).

Using `CounterfactualExplanations`

Counterfactual generators

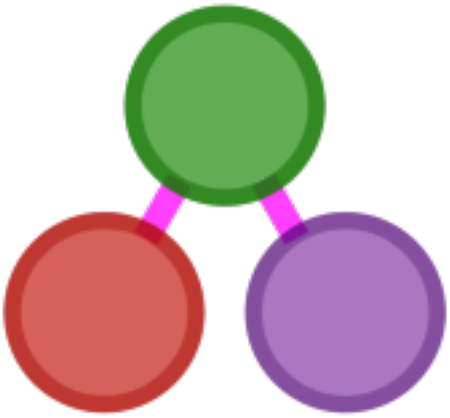

Figure 1: Figure

Empirical example

Related and future work