Metrics to Ascertain the Plausibility and Faithfulness of Counterfactual Explanations

Ali Faruk Yücel (A.F.Yucel@student.tudelft.nl), **Responsible Professor:** Cynthia Liem, **Supervisor:** Patrick Altmeyer

EEMCS Faculty, Delft University of Technology, The Netherlands

1. Introduction

Counterfactual explanations (CEs) are classified as a *post-hoc, locally interpretable, model-agnostic* explainability method within the Explainable AI (XAI) ontology. These explanations play a vital role in *preparative* and *affective* functions, helping individuals learn from past mistakes and course correct via algorithmic recourse, alongside aiding them to be more content about their current situation.

Despite growing body of research aimed at producing CE generators, there is a lack of advancement on evaluating the properties of generated counterfactuals, which has guided the focus of our research. Specifically the **plausibility** and **faithfulness** of CEs are of great importance as these largely indicate the *feasibility* and *trustworthiness* of CEs.

Definition 1. (Plausibility): A counterfactual (x′) is considered plausible if it's distributed by the true conditional distribution of samples (x) in the target class ($X \mid y^+$).

$$\mathcal{X} \mid \mathbf{y}^+ = p(\mathbf{x} \mid \mathbf{y}^+) \qquad \mathbf{x}' \sim \mathcal{X} \mid \mathbf{y}^+$$

Definition 2. (Faithfulness): A counterfactual (x′) is considered faithful if it's distributed by the learned conditional distribution of samples (x) in the target class. The symbol θ represents the parameters of the trained model.

$$\mathcal{X}_\theta \mid \mathbf{y}^+ = p_\theta(\mathbf{x} \mid \mathbf{y}^+) \qquad \mathbf{x}' \sim \mathcal{X}_\theta \mid \mathbf{y}^+$$

Research Question (RQ): **How to evaluate the plausibility and faithfulness of counterfactual explanations?**

A holistic exploration of the RQ is performed through three sub-questions:

Sub-question 1: **What are the *shortcomings* of *methods to quantify the data manifold?***
Sub-question 2: **Which *metrics* are used to *quantify the degree of closeness* of a counterfactual to the data manifold?**
Sub-question 3: **What *novel metrics* could be used as *proxies to estimate the degree of closeness* of a counterfactual to the data manifold?**

4. Limitations & Future Work

- Lack of existence of packages for multivariate K-S test calculation limited empirical support for its proposal.
- As experiments are conducted with deep learning models, time constraints limited an in-depth reliability analysis.
- IM1 metric could be probed further to be used to compute faithfulness scores.

5. Conclusions

- LOF score could be used as a proxy metric for quantifying the degree of closeness to the data manifold.
- Utilizing Gower distance does not alter the outcome of the LOF score significantly enough for it to be considered beneficial

2. Methods & Metrics

Methods to Quantify the Data Manifold

- FACE (Feasible and Actionable Counterfactual Explanations) by Poyiadzi et. al. (2020)
  - Constructs a graph over input data with edge-weights determined by one of three methods: Kernel Density Estimation (KDE), k-Nearest Neighbour (k-NN), ε-Graph
- Diffusion Distance & Directional Coherence by Domnich et. al. (2024)
- DiCE (Diverse Counterfactual Explanations) by Mothilal et. al. (2019)
- Predictive Uncertainty by Schut et. al. (2021)
- Algorithmic Recourse Under Imperfect Causal Knowledge by Karimi et. al. (2020)
- Diffeomorphic Counterfactuals with Generative Models
  - Utilizes normalizing flows to perform a diffeomorphic coordinate transformation. This method faces scalability issues due to the large memory footprint required by normalizing flows.
- ECCo (Energy-Constrained Conformal Counterfactuals) by Altmeyer et. al. (2023)
  - Quantifies the learned conditional distribution of samples in the target class by utilizing Stochastic Gradient Langevin Dynamics (SGLD).
- Variational Autoencoders by Joshi et. al. (2019)
- NAE by Pawelczyk et. al. (2021)

Metrics to Quantify the Degree of Closeness

IM1 metric was introduced by Arnaud et. al. (2020) to measure *realism*, which has the same definition as *plausibility*. AE_t stands for an autoencoder trained on t.

$$\text{IM1} = \frac{\|x' - \text{AE}_{y'}(x')\|_2^2}{\|x' - \text{AE}_y(x')\|_2^2 + \epsilon}$$

Sufficiency and **Comprehensiveness** metrics based on the conception by DeYoung et. al. (2020) and may be used to measure *faithfulness*.

$$\text{Comp.} = \frac{1}{N}\sum_{i=1}^{N}\left(p(y_i \mid x_i) - p(y_i \mid x_i \setminus e_i)\right)$$

$$\text{Suff.} = \frac{1}{N}\sum_{i=1}^{N}\left(p(y_i \mid x_i) - p(y_i \mid e_i)\right)$$

Diffusion distance metric is derived from diffusion maps and it measures the *connectivity* between points in a dataset. It is robust to noise, and can handle non-linear manifolds. It can be leveraged to measure both *plausibility* and *faithfulness*.

Proposal Metrics to Quantify the Degree of Closeness

Kolmogorov-Smirnov (K-S) Test

One-Sample K-S Test:

$$D = \sup_x |F_n(x) - F(x)|$$

$F(x)$: Cumulative Distribution Function of reference distribution
$F_n(x)$: Empirical Distribution Function of a sample of size n

Two-Sample K-S Test:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

$F_n(x)$: Empirical Distribution Function of a sample of size n
$G_m(x)$: Empirical Distribution Function of a sample of size m

Local Outlier Factor with Gower Distance

LOF metric of a given point p:

$$\text{LOF}(p) = \frac{\sum_{o \in N_k(p)} \text{LRD}(o)}{|N_k(p)|\,\text{LRD}(p)}$$

$N_k(p)$: k nearest neighbours of p

$$\text{LRD}(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} \text{RD}(p,o)}$$

$$\text{RD}(p,o) = \max(k\text{-distance}(o), d(p,o))$$

Gower distance is used for $d(p,o)$

3. Results

For every dataset and every basis model, we have trained 10 different models for 10 epochs on the respective datasets. Counterfactuals were generated over 5 identical samples for each dataset.

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	6.185 ± 0.290	3.274 ± 0.017	5.054 ± 0.025
	DiCE ($\lambda_2 = 0.5$)	6.181 ± 0.268	3.268 ± 0.024	5.059 ± 0.287
	DiCE ($\lambda_2 = 1$)	6.210 ± 0.038	3.264 ± 0.018	5.018 ± 0.035
	ClaPROAR	6.117 ± 0.017	3.274 ± 0.038	4.945 ± 0.044
Dropout	Generic	6.181 ± 0.442	3.278 ± 0.288	4.933 ± 0.184
	DiCE ($\lambda_2 = 0.5$)	6.181 ± 0.413	3.270 ± 0.237	4.921 ± 0.316
	DiCE ($\lambda_2 = 1$)	6.179 ± 0.042	3.248 ± 0.016	5.053 ± 0.017
	ClaPROAR	6.103 ± 0.015	3.177 ± 0.047	5.020 ± 0.057

Table 1: Average implausibility score per model and dataset for different generators.

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	0.972 ± 0.010	0.465 ± 0.010	0.702 ± 0.024
	DiCE ($\lambda_2 = 0.5$)	0.971 ± 0.010	0.462 ± 0.010	0.701 ± 0.027
	DiCE ($\lambda_2 = 1$)	0.971 ± 0.010	0.548 ± 0.023	0.722 ± 0.015
	ClaPROAR	0.991 ± 0.002	0.461 ± 0.107	0.796 ± 0.144
Dropout	Generic	0.975 ± 0.010	0.484 ± 0.019	0.694 ± 0.009
	DiCE ($\lambda_2 = 0.5$)	0.975 ± 0.010	0.480 ± 0.017	0.694 ± 0.010
	DiCE ($\lambda_2 = 1$)	0.974 ± 0.010	0.486 ± 0.166	0.714 ± 0.008
	ClaPROAR	0.992 ± 0.002	0.639 ± 0.194	0.783 ± 0.223

Table 2: Average LOF score (using L2 Norm as distance) per model and dataset for various generators.

Upon examination of Table 1 and 2, it is evident that LOF score approximates the degree of closeness to the data manifold well since *implausibility* and *LOF* scores follow the same general trend for identical CEs. As LOF score approaches 1, the CE is understood to be less plausible or faithful.

Model	Generator	German Credit	California Housing	Adult Income
Neural Network	Generic	0.972 ± 0.010	0.465 ± 0.009	0.709 ± 0.018
	DiCE ($\lambda_2 = 0.5$)	0.972 ± 0.010	0.464 ± 0.011	0.710 ± 0.017
	DiCE ($\lambda_2 = 1$)	0.972 ± 0.010	0.480 ± 0.009	0.702 ± 0.019
	ClaPROAR	0.991 ± 0.002	0.546 ± 0.022	0.770 ± 0.018
Dropout	Generic	0.975 ± 0.010	0.482 ± 0.020	0.703 ± 0.01
	DiCE ($\lambda_2 = 0.5$)	0.974 ± 0.010	0.483 ± 0.022	0.701 ± 0.009
	DiCE ($\lambda_2 = 1$)	0.974 ± 0.010	0.492 ± 0.021	0.700 ± 0.010
	ClaPROAR	0.992 ± 0.001	0.638 ± 0.016	0.765 ± 0.023

Table 3: Average LOF score (using Gower distance) per model and dataset for various generators.