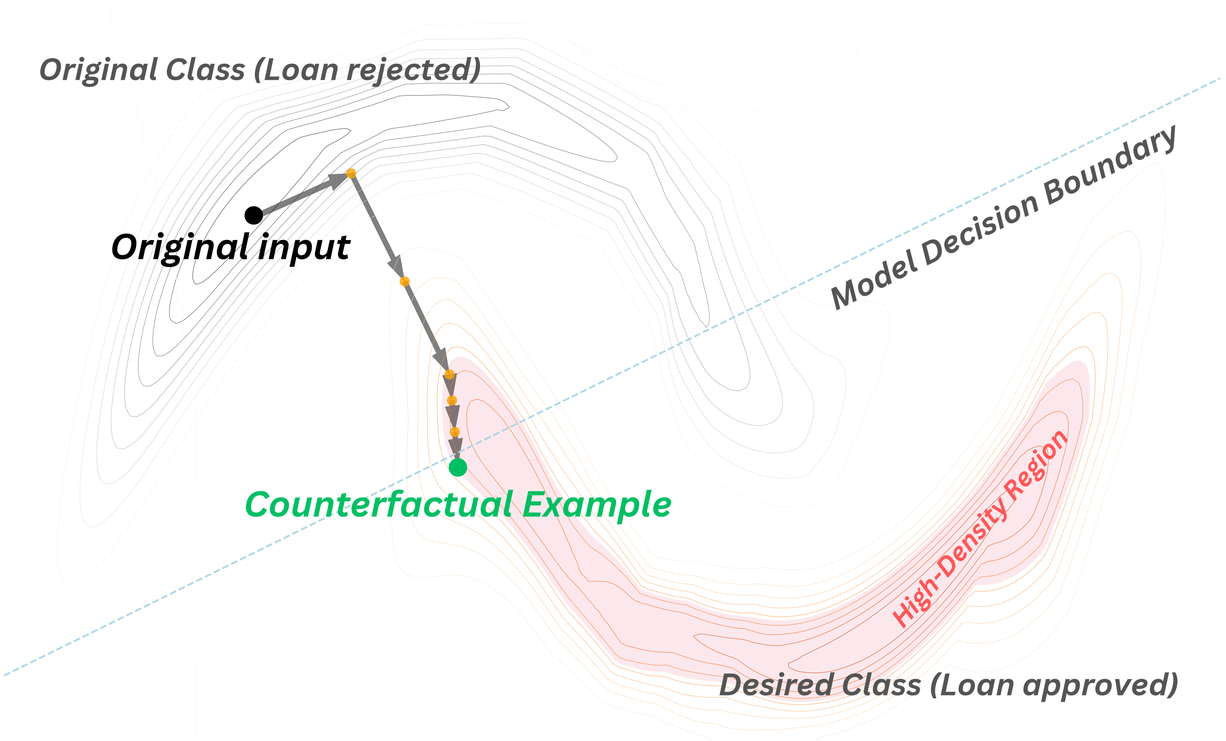
Probabilistically Plausible Counterfactual Explanations with Normalizing Flows

Patryk Wielopolski [1], Oleksii Furman [1], Jerzy Stefanowski [2], Maciej Zięba [1,3]

[1]Wrocław University of Science and Technology [2]Poznań University of Technology [3]Tooploox

Ever Wondered "What If?"

What changes would turn a loan rejection into approval?



Motivation and Problem Statement

- Ensuring **validity**: change the model's prediction to the desired class.
- Maintaining **plausibility**: reside in a high-density region of the target class distribution.
- Balancing **closeness**: achieving minimal changes from the original input.

Method Overview

- **Unconstrained Optimization**: balances validity, plausibility, and proximity.
- **Direct Density Estimation with Normalizing Flows**: models complex data distributions for realistic counterfactuals.
- **Batch Processing Capability**: efficient generation for multiple data points simultaneously.

Experimental Results

Dataset	Method	Validity ↑	Prob. Plaus. ↑	L1 ↓	L2 ↓	Time (s) ↓
Moons	CBCE	**1.00**	0.10	0.62	0.48	**0.07**
	CEGP	**1.00**	0.09	0.36	**0.28**	904.11
	CEM	**1.00**	0.14	0.55	0.50	211.56
	WACH	**1.00**	0.11	0.49	0.36	198.29
	ARTELT	**1.00**	0.08	**0.32**	0.32	4.15
	PPCEF	**1.00**	**1.00**	0.45	0.36	1.85
Law	CBCE	**1.00**	0.49	0.61	0.40	**0.23**
	CEGP	**1.00**	0.49	0.23	0.18	1973.76
	CEM	**1.00**	0.26	0.33	0.31	368.10
	WACH	**1.00**	0.39	0.45	0.35	359.00
	ARTELT	**1.00**	0.40	**0.20**	**0.20**	4.02
	PPCEF	**1.00**	**1.00**	0.37	0.23	2.42
Audit	CBCE	**1.00**	0.79	2.55	1.24	**0.04**
	CEGP	**1.00**	0.02	1.56	0.57	561.04
	CEM	**1.00**	0.00	1.20	**0.37**	105.92
	WACH	**1.00**	0.02	1.78	0.80	101.27
	ARTELT	0.97	0.00	**0.90**	0.88	43.84
	PPCEF	**0.99**	**0.99**	2.04	0.79	7.01
Heloc	CBCE	**1.00**	0.54	2.84	0.82	**5.71**
	CEGP	**1.00**	0.29	**0.26**	**0.10**	9654.60
	CEM	**1.00**	0.07	0.35	0.20	1639.16
	WACH	**1.00**	0.00	0.74	0.37	1600.28
	ARTELT	-	-	-	-	-
	PPCEF	**1.00**	**1.00**	0.90	0.23	12.44

Mathematical Formulation

Given an input x_0 and target class y', **PPCEF** solves:

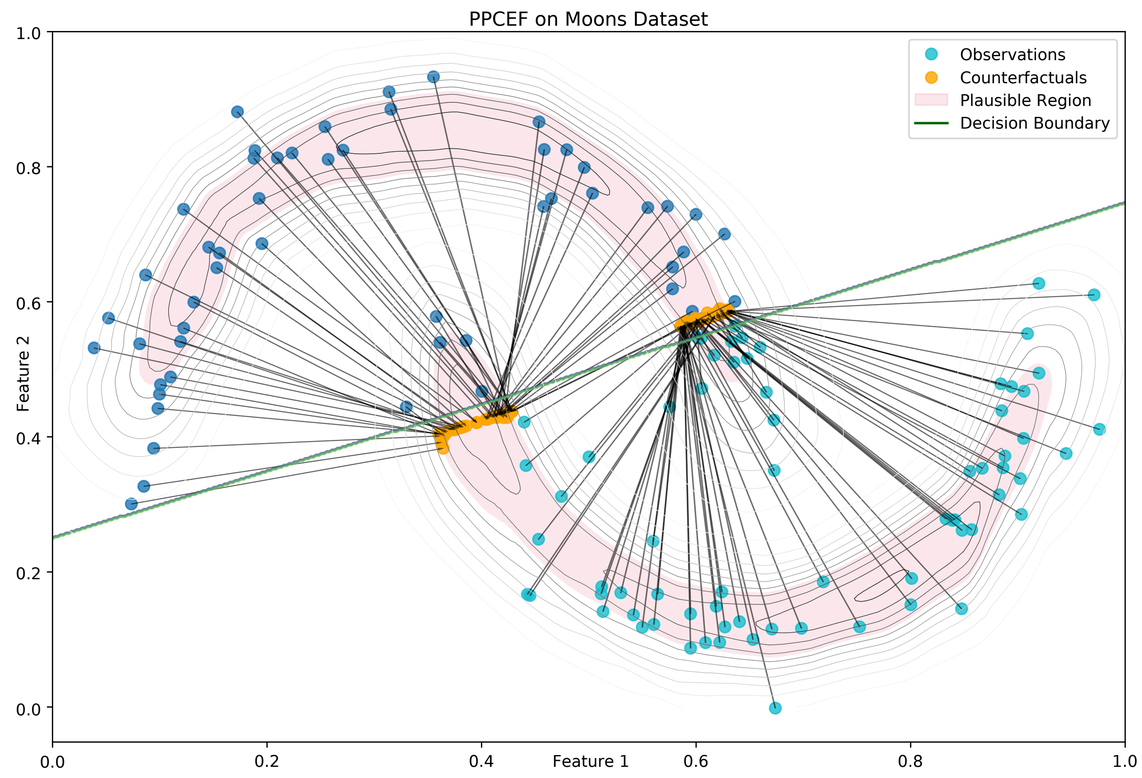
$$\arg\min_{x' \in \mathbb{R}^d} d(x_0, x') + \lambda(\ell_v(x', y') + \ell_p(x', y'))$$

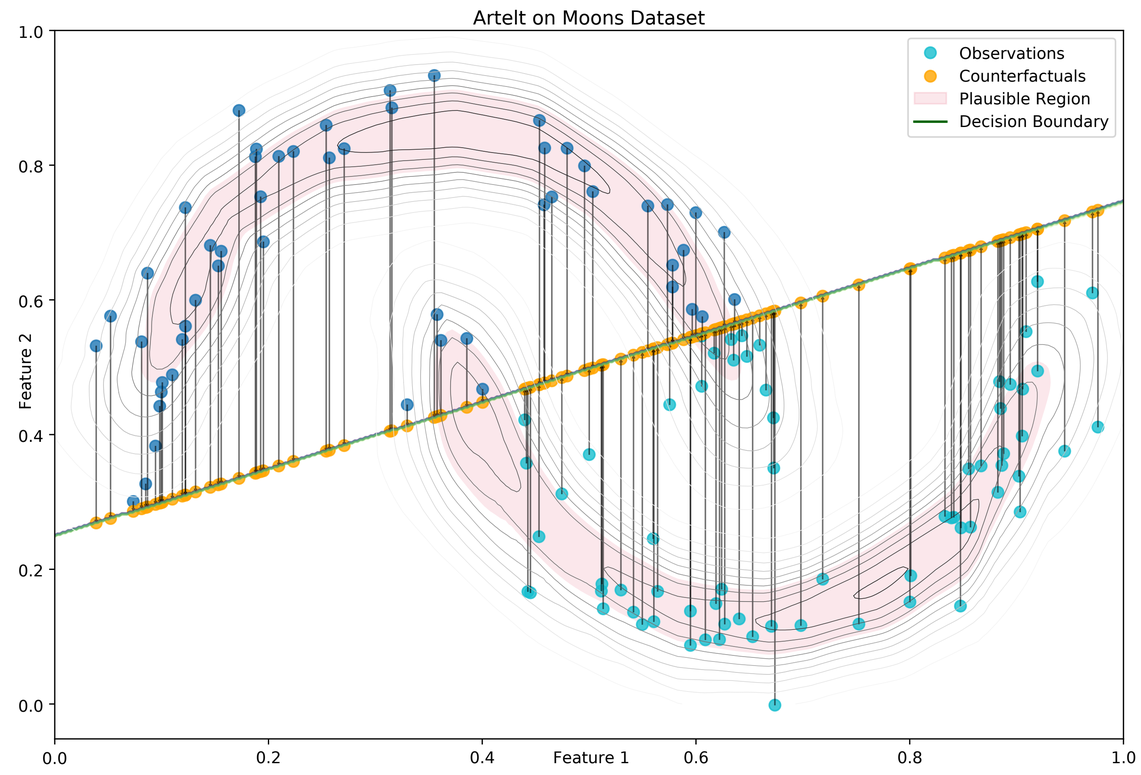
where x' is the counterfactual, $d(x_0, x')$ is the proximity term, λ balances the trade-off.

Validity Loss (ℓ_v): $\ell_v(x', y') = \max(0.5 + \epsilon - p_d(y'|x'), 0)$

Plausibility Loss (ℓ_p): $\ell_p(x', y') = \max(\delta - p(x'|y'), 0)$

Visual Example





Contributions

- **Unified Framework**: Combines validity, proximity, and plausibility in counterfactual generation.
- **Normalizing Flows**: Estimates complex data distributions for realistic outcomes.
- **Efficient Generation**: Achieves fast counterfactuals via gradient-based optimization and batch processing.
- **Wide Applicability**: Outperforms prior methods across diverse datasets and models.

Contact Information

Paper

GitHub Repository

Contact