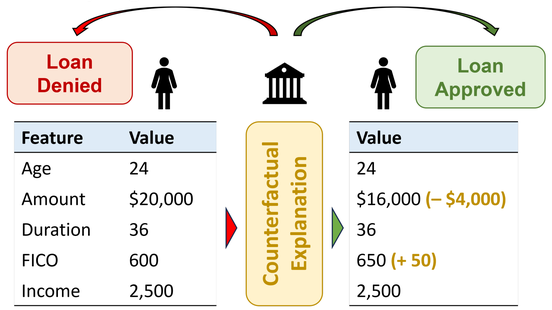
CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations

Hangzhi Guo[1], Thanh Hong Nguyen[2], Amulya Yadav[1]

[1]Pennsylvania State University, [2]University of Oregon

Background

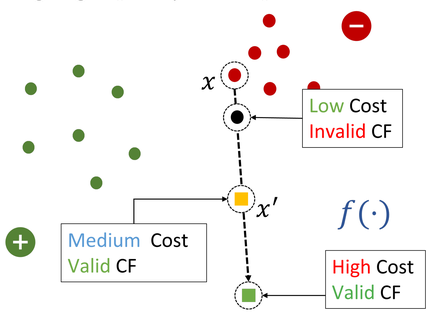
Counterfactual (CF) explanations offer contrast cases for individuals who are *adversely impacted* by algorithmic decisions.



Feature	Value
Age	24
Amount	$20,000
Duration	36
FICO	600
Income	2,500

Value
24
$16,000 (– $4,000)
36
650 (+ 50)
2,500

Balancing the **cost-invalidity trade-off** is key to generating high-quality CF explanations.



Post-hoc Explainability. Existing methods are post-hoc (*explain after prediction*)

- by assuming underlying models as *black boxes*.

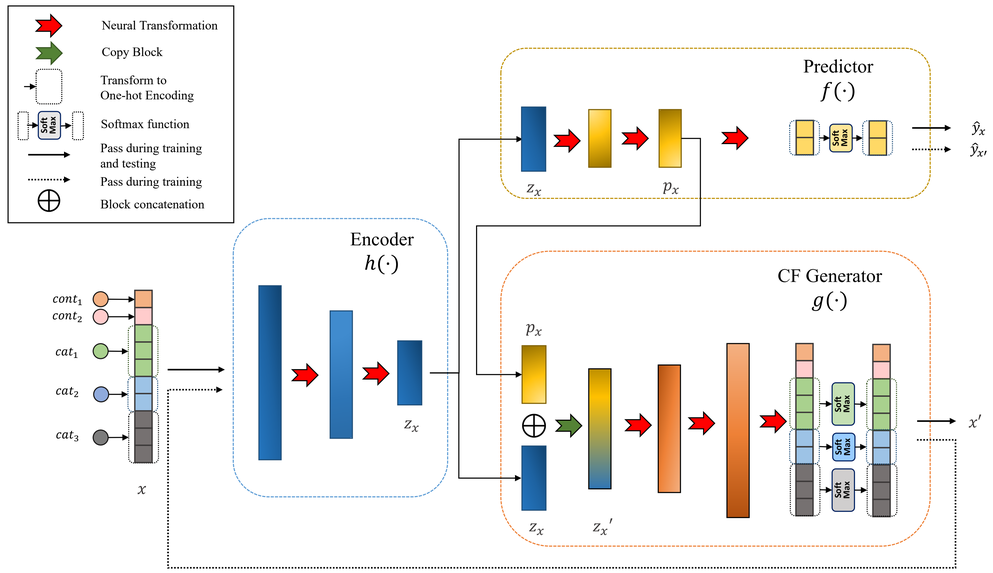
Motivations

1. EU-GDPR enforces the "*Right to Explanation*".
   - ML developers want to pair explanations with predictions (i.e., the black-box assumption is overly limiting).
2. The post-hoc paradigm is *sub-optimal*.
   - It fails to properly balance the cost-invalidity trade-off.
   - Most of the post-hoc explanation methods run slow.

Research Objective

Can we depart from the dominant post-hoc paradigm in CF explanations by integrating predictive model training and CF explanation generation within ***an end-to-end pipeline****?*

CounterNet Architecture



CounterNet's Objective Function

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left[\lambda_1 \cdot \underbrace{(y_i - \hat{y}_{x_i})^2}_{\text{Prediction Loss } (\mathcal{L}_1)} + \lambda_2 \cdot \underbrace{\left(\hat{y}_{x_i} - (1 - \hat{y}_{x'_i})\right)^2}_{\text{Validity Loss } (\mathcal{L}_2)} + \lambda_3 \cdot \underbrace{(x_i - x'_i)^2}_{\text{Change Loss } (\mathcal{L}_3)} \right] \quad (1)$$

Training Issues

Two issues occur when directly optimizing E.q. 1:

- **Issue I: Poor Convergence.**

LEMMA I*. $\nabla\mathcal{L}_1 \cdot \nabla\mathcal{L}_2 < 0$.

- **Issue II: Adversarial Vulnerability.**

LEMMA II*. *Minimizing $\mathcal{L}_2$ w.r.t. θ_f increases the Lipschitz constant of f.*

*Refer to our paper for details of Lemma I & II.

Block-Wise Coordinate Descent

The Solution to Issue I & II: a block-wise coordinate descent procedure.

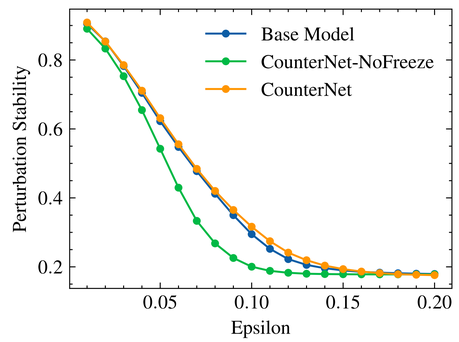
For each batch of m data points $\{x^{(i)}, y^{(i)}\}^m$,

- $\theta^{(1)} = \theta^{(0)} - \nabla_{\theta}^{(0)}(\lambda_1 \cdot \mathcal{L}_1)$
- $\theta_g^{(2)} = \theta_g^{(1)} - \nabla_{\theta_g^{(1)}}(\lambda_2 \cdot L_2 + \lambda_3 \cdot \mathcal{L}_3)$

Ablation Study

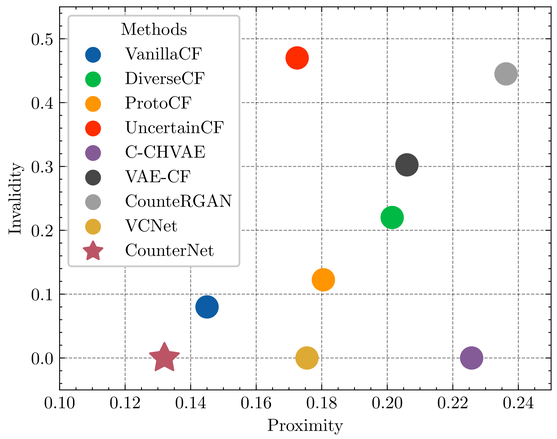
Ablation	Adult		Credit		HELOC		OULAD	
	Val.	Prox.	Val.	Prox.	Val.	Prox.	Val.	Prox.
CounterNet-SingleBP	0.64	.248	0.92	.251	0.93	.206	0.94	.110
CounterNet-Separate	0.96	.257	0.99	.265	0.91	.161	0.94	.097
CounterNet-No-p_x	0.97	.256	0.99	.339	0.98	.147	0.98	.101
CounterNet-Posthoc	1.00	.276	1.00	.247	1.00	.153	0.99	.099
CounterNet	**1.00**	**.196**	**1.00**	**.132**	**1.00**	**.125**	**1.00**	**.075**

Adversarial Robustness



Evaluation

- CounterNet is the **best-performing** method by achieving the **lowest** invalidity and proximity.



- CounterNet runs **faster** than all of the baselines.

Method	Adult	Credit	HELOC	OULAD
VanillaCF	1432.09	1358.26	1340.42	1705.93
DiverseCF	4685.39	3898.43	3921.72	5478.17
ProtoCF	2348.21	2056.01	1956.71	2823.29
UncertainCF	379.95	60.80	7.91	6.81
C-CHVAE	3.28	568.28	2.68	4.79
VAE-CF	1.72	1.28	1.48	1.84
CounteRGAN	1.96	1.77	1.59	2.40
VCNet	1.39	1.23	1.13	1.81
CounterNet	**0.64**	**0.39**	**0.44**	**0.79**

- CounterNet **matches** predictive accuracy.

Model	Adult	Credit	HELOC	OULAD
Base Model	0.831	0.813	0.717	0.934
CounterNet	0.828	0.819	0.716	0.929

Key Insights

- Post-hoc explainability can be sub-optimal and overly limiting in counterfactual explanations.
- CounterNet represents a first step towards developing end-to-end CF explanation systems.

Code: `https://github.com/BirkhoffG/counternet`

PennState

Oregon State University