Model-agnostic and Scalable Counterfactual Explanations via Reinforcement Learning

Robert-Florian Samoilescu[1,2], Arnaud Van Looveren[1], Janis Klaise[1]

[1]Seldon Technologies Ltd., [2]University Politehnica of Bucharest

Motivation

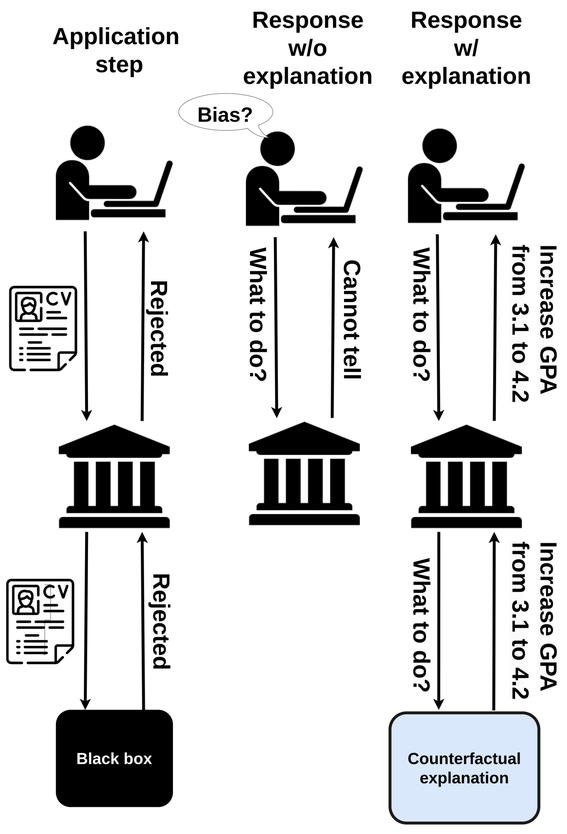

Figure 1:Rejection response to an university application.

What are counterfactuals?

Counterfactual:

- minimal necessary change in the input space to alter the prediction.

To be practical:

- sparse - close to the input instance.
- in-distribution - indistinguishable from real instances.
- diverse - multiple options.

Desirable properties

- immutable features (e.g., gender, race).
- feature constraints.

Current limitations

Iterative procedures:

- separate, computationally expensive optimization per instance.

Access to gradients:

- operate only in the white-box regime.
- not practical for tabular data (SoTA - Random Forest, XGBoost).

Do not allow feature conditioning:

- sensitive immutable features are changed.
- lead to unactionable recourse (e.g., decrease age).

Example

	IN	CF	Condition
Age	40	40	[40, 45]
Workclass	Private	Private	{Private, Federal-gov, Self-emp-inc}
Education	High School grad	Masters	{High School grad, Bachelors, Masters}
Marital Status	Married	Married	{Married}
Occupation	Sales	White-Collar	{Sales, White-Collar, Admin}
Relationship	Husband	Husband	{Husband}
Race	White	White	{White}
Sex	Male	Male	{Male}
Capital Gain	0	0	[0, 0]
Capital Loss	0	0	[0, 0]
Hours per week	60	60	[60, 60]
Country	Latin-America	Latin-America	{Latin-America}
Prediction	≤ $50k/y	> $50k/y	

Figure 2:Conditional counterfactual instance on Adult dataset.

Our contributions

- Model-agnostic, target-conditional framework primarily focused on heterogeneous tabular datasets.
- Flexible feature range constraints for numerical and categorical features.
- Fast counterfactual generation process.
- Easily extendable framework to other data modalities (e.g. images).

Reinforcement learning training pipeline

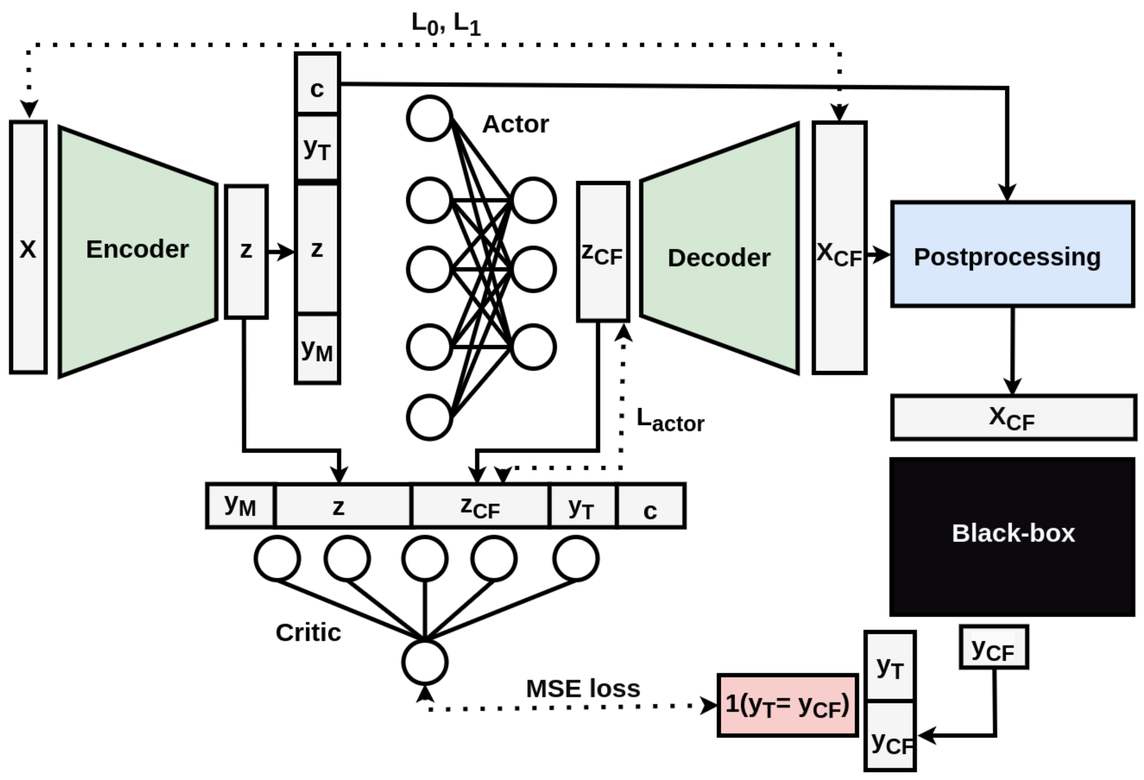

Figure 3:Generative model training pipeline using RL for counterfactual generation.

Diversity

	IN	CF(1)	CF(2)	CF(3)
Age	40	40	40	40
Workclass	Private	Private	Private	Private
Education	Masters	Dropout	Masters	High School grad
Marital Status	Married	Married	Married	Married
Occupation	White-Collar	Service	Professional	Sales
Relationship	Husband	Husband	Husband	Husband
Race	White	White	White	White
Sex	Male	Male	Male	Male
Capital Gain	0	0	0	0
Capital Loss	0	0	0	0
Hours per week	40	40	28	32
Country	United-States	United-States	United-States	United-States
Prediction	> $50k/y	≤ $50k/y	≤ $50k/y	≤ $50k/y

Figure 4:Diverse counterfactual instances via feature conditioning subsampling.

Validity

Table 1:Percentage of generated counterfactuals of the desired target label - higher is better.

	Validity (%)			
Method	**Adult**	**Cancer**	**Portug.**	**Spam.**
LORE	18.08	25.95	19.07	9.53
MO	91.00	100.00	100.00	100.00
DiCE(r)	99.93	100.00	99.98	99.58
DiCE(g)	33.94	60.86	90.97	40.93
Ours	98.59	99.24	98.27	99.18

Sparsity

Table 2:$\mathcal{L}_0$ and $\mathcal{L}_1$ distance - lower is better.

	Adult		Cancer	Portug.		Spam.
Method	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_0$	$\mathcal{L}_0$	$\mathcal{L}_1$	$\mathcal{L}_1$
LORE	0.09	0.11	0.11	0.04	0.12	0.09
MO	0.20	0.39	0.53	0.22	0.31	0.30
DiCE(r)	0.05	1.76	0.29	0.07	1.39	0.28
DiCE(g)	0.18	0.09	0.56	0.23	0.60	0.31
Ours	0.11	0.19	0.44	0.23	0.15	0.17

In-distributioness

Table 3:Negative class conditional MMD - lower is better.

	MMD_0^2 (10^{-1})			
Method	**Adult**	**Cancer**	**Portug.**	**Spam.**
LORE	0.31	1.09	0.08	0.26
MO	0.45	0.50	0.16	0.61
DiCE(r)	0.56	1.03	1.75	0.80
DiCE(g)	0.28	0.85	0.68	0.25
Ours	0.36	0.36	0.10	0.32

Other data modalities

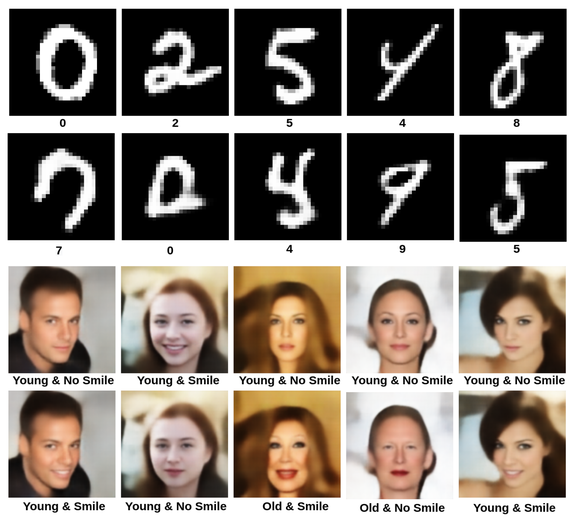

Figure 5:MNIST (top half) and CelebA (bottom half) counterfactual instances.

ALIBI EXPLAIN