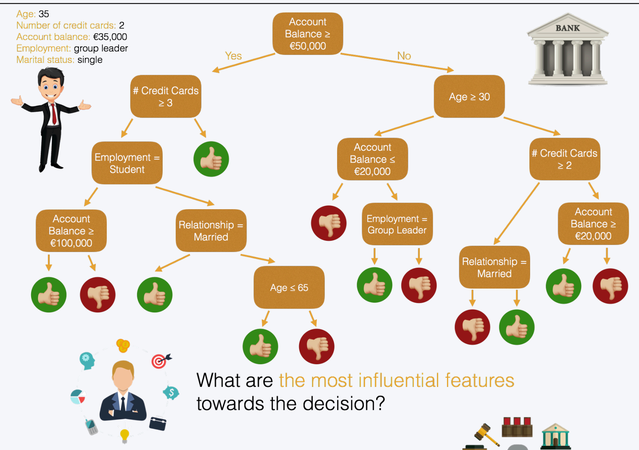
Scaling Guarantees for Nearest Counterfactual Explanations

Kiarash Mohammadi[1,2] Amir-Hossein Karimi[1,3], Gilles Barthe[4], Isabel Valera[5]
[1]MPI for Intelligent Systems, Tübingen, Germany; [2]Ferdowsi University of Mashhad, Mashhad, Iran;
[3]ETH Zürich, Zürich, Switzerland; [4]MPI for Security and Privacy, Bochum, Germany;
[5]Department of Computer Science, Saarland University, Saarbrücken, Germany.

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

Summary



Method	Opt. Distance	100% Coverage	Efficiency	Neural Models	Qualitative Features	Complex Constraints
Our approach	✓	✓	✓	✓	✓	✓
MACE [1]	✓	✓		✓	✓	✓
DiCE [2]			✓	✓	✓	
Efficient Search [3]	✓	✓	✓		✓	✓

Notation & Background

- Classifier: $h : \mathcal{X} \to \mathbb{R}$
- Label (e.g., loan approval): $h(\mathbf{x}) \geq 0 \Rightarrow y = +1, \quad h(\mathbf{x}) < 0 \Rightarrow y = -1$
- Individual/factual observation: $\boldsymbol{x}^{\mathbf{F}}$ s.t. $h(\boldsymbol{x}^{\mathbf{F}}) < 0$
- Two ways to formulate CFE generation:

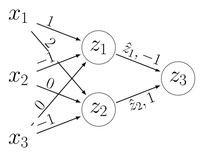
Optimization Formulation

$$\mathbf{x}^{CFE} \in \underset{\mathbf{x}\in\mathcal{X}}{argmin} \; dist(\mathbf{x}, \mathbf{x}^F) \quad s.t. \; h(\mathbf{x}) \geq 0, \; \mathbf{x} \in Plausible, \; \mathbf{x} \in Actionable \tag{1}$$

Verification Formulation

$$\exists \mathbf{x}. dist(\mathbf{x}, \mathbf{x}^F) \leq \delta, \; h(\mathbf{x}) \geq 0, \; \mathbf{x} \in Plausible, \; \mathbf{x} \in Actionable \tag{2}$$

- Encoding a **ReLU Neural Network (NN)** as an SMT formula:



$$\begin{aligned}&\phi_f(x) = \\ &(z_1 = x_1 - x_2) \wedge (z_2 = 2x_1 - x_3) \\ &\wedge((\hat{z}_1 = z_1 \wedge z_1 \geq 0) \vee (\hat{z}_1 = 0 \wedge z_1 < 0)) \\ &\wedge((\hat{z}_2 = z_2 \wedge z_2 \geq 0) \vee (\hat{z}_2 = 0 \wedge z_2 < 0)) \\ &\wedge(z_3 = -\hat{z}_1 + \hat{z}_2)\end{aligned}$$

MIP Encodings

For NNs, we adopt a *bounded* encoding by by Tjeng and Tedrake [4], i.e., for $i \in \{1, ..., n\}$ ($\mathbf{l}_i$ and $\mathbf{u}_i$ indicate the lower/upper bounds):

$$\mathbf{z}_i = \mathbf{W}_i \hat{\mathbf{z}}_{i-1} + \mathbf{b}_i \tag{3a}$$

$$\boldsymbol{\delta}_i \in \{0,1\}^{k_i}, \quad \hat{\mathbf{z}}_i \geqslant 0, \quad \hat{\mathbf{z}}_i \leqslant \mathbf{u}_i \cdot \boldsymbol{\delta}_i, \quad \hat{\mathbf{z}}_i \geqslant \mathbf{z}_i, \quad \hat{\mathbf{z}}_i \leqslant \mathbf{z}_i - \mathbf{l}_i \cdot (1 - \boldsymbol{\delta}_i) \tag{3b}$$

As a preliminary step, a linear approximation of ReLUs from Ehlers [5] replaces (3b) to compute the lower/upper bounds:

$$\hat{\mathbf{z}}_i \geqslant \mathbf{z}_i, \quad \hat{\mathbf{z}}_i \geqslant 0, \quad \hat{z}_{i,j} \leqslant u_{i,j} \frac{z_{i,j} - l_{i,j}}{u_{i,j} - l_{i,j}} \tag{4}$$

These bounds are then placed in (3) to complete an *exact* MIP encoding.

CFE Generation

Within each iteration of an exponential search that gradually increases the distance interval, tight bounds of the hidden units are computed.

- **MIP-SAT**: Relies on SMT solving. Removes fixed-state ReLUs from the SMT formula given their bounds. Verifies the new formula.
- **MIP-EXP**: Relies on MIP solving. Uses the bounds to implement (3) and optimizes it toward and until flipping the output logit sign.
- **MIP-OBJ**: No exponential search. Uses the distance MIP as the objective, with a constraint of the output logit being flipped, directly minimizes distance.

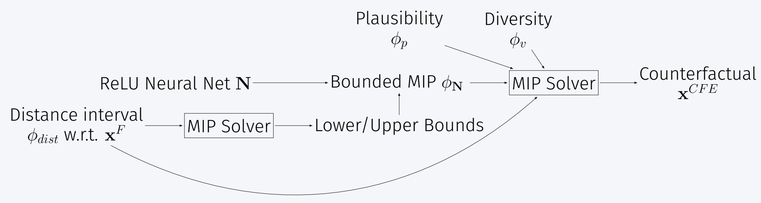
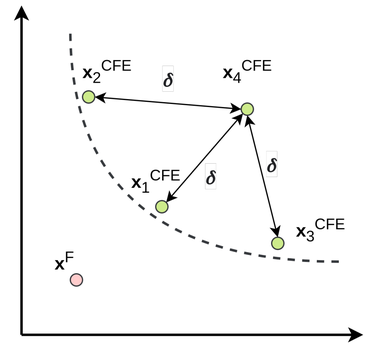


Figure 2. An iteration of the search



Different qualitative features, such as diversity, may also be encoded. For diversity, the following constraints are added for the $i-th$ CFE to be generated:

$$dist(\mathbf{x}_1^{CFE}, \mathbf{x}_i^{CFE}) \geq \delta \quad \dots \quad dist(\mathbf{x}_{i-1}^{CFE}, \mathbf{x}_i^{CFE}) \geq \delta \tag{5}$$

Experiments

Setup.

We compare our approaches against MACE [1] and DiCE [2] in various settings on runtime, distance, and coverage. We employ three widely used real-world datasets from the literature: Adult ($d = 51$), COMPAS ($d = 7$), and Credit ($d = 20$). We use a two-layer ReLU-activated NN with 10 neurons for most experiments. While NNs of this scale can sufficiently discriminate between the classes of the supervised learning task, we also include experiments exploring the scalablity (deepening or widening of the NN) of our approach against the opponents. Finally, we show that qualitative features, diversity in this case, can be encoded within the framework to efficiently generate *sets* of CFEs with guarantees

Results.

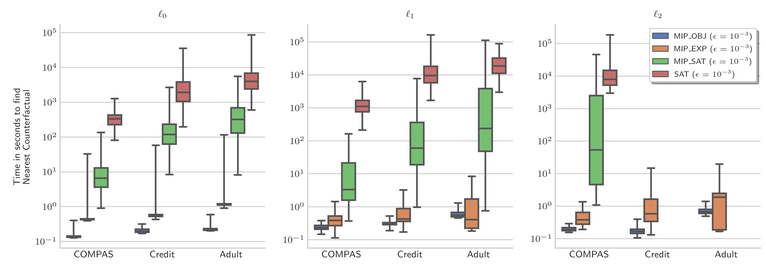


Figure 3. Full-setting runtime comparison of two-layer ReLU-activated NN with 10 neurons in each layer among our approach and MACE (SAT) [1]. Coverage is perfect by design. Each setting has been evaluated on 500 instances, however, SAT and MIP-SAT timed out on some samples. For such cases, only the samples for which all approaches have successfully finished running are included.

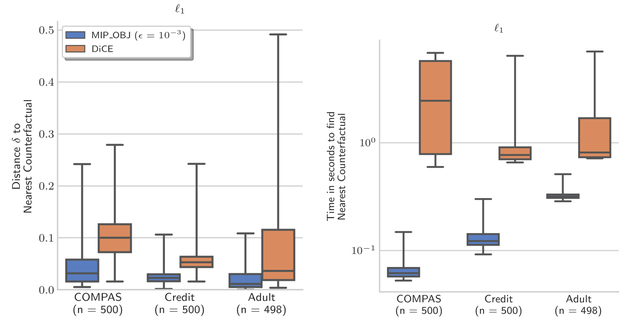


Figure 4. Distance and time comparison against DiCE as a gradient-based optimization approach. The NN model is same as above. MIP-OBJ coverage is perfect by design and DiCE coverage is also perfect but slightly dips for Adult dataset (99.6%).

References

[1] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, 2020.

[2] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[3] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 20–28, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. *CoRR*, abs/1711.07356, 2017.

[5] Rüdiger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. *CoRR*, abs/1705.01320, 2017.