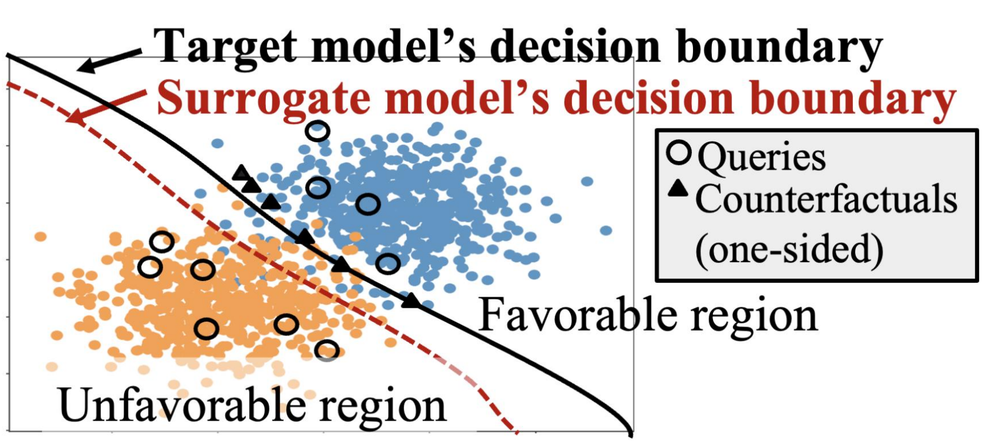
Classifier Reconstruction Through Counterfactual-Aware Wasserstein Prototypes

Xuan Zhao[1], Zhuo Cao[1], Arya Bangun[1], Hanno Scharr[1], Ira Assent[1,2]

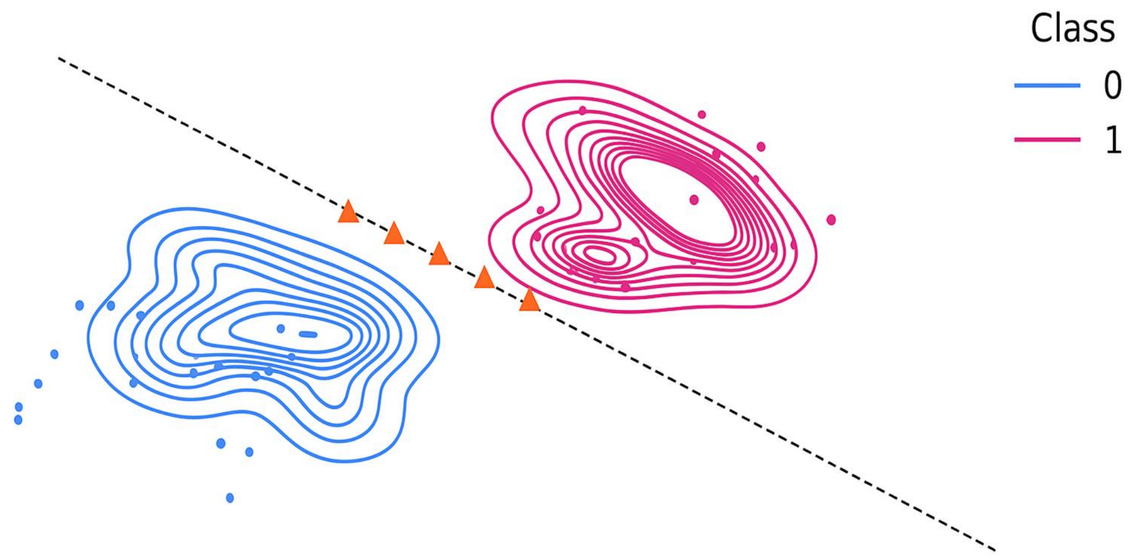
[1]Forschungszentrum Jülich [2]Aarhus University

MODEL RECONSTRUCTION WITH ORIGINAL SAMPLES AND COUNTERFACTUALS

- Counterfactual explanations [1,2] aim to identify minimal, semantically meaningful changes to an input that lead to a different, desired prediction outcome. On one hand, they provide users with intuitive insights into model behavior. On the other hand, they raise significant privacy concerns.
- In this work, we investigate the extent to which counterfactual samples can leak information about the underlying model. Specifically, we approximate an unknown model by querying it to generate counterfactuals, which are near the decision boundary. These counterfactuals enrich the dataset with informative yet atypical examples, potentially exposing sensitive characteristics of the model.



Counterfactuals should not be treated as normal data samples. The figure is taken from [3].



INTUITION BEHIND OUR APPROACH

- Original class samples and represent "pure" classes; counterfactual samples blend features of both classes near the decision boundary.
- We treat counterfactuals as soft or ambiguous samples that influence how we represent each class in feature space.
- To approximate the model decision boundary, we propose to use Wasserstein barycenter as the prototypes.

KEY FINDINGS

- Our method achieves higher fidelity than state-of-the-art baselines.
- Especially effective in low-query regimes (300–400 queries).
- Performance benefits from counterfactuals that are realistic and actionable.

WASSERSTEIN DISTANCE AND BARYCENTERS

The Wasserstein distance [4], also known as Earth Mover's Distance, measures the minimal cost of transforming one probability distribution into another, reflecting both the amount and distance of probability mass that must be moved.

$$W_2^2(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \in_{M\times M} d(x,y)^2 d\gamma(x,y)$$

The Wasserstein barycenter [5] is a geometric notion of the average or prototype of multiple probability distributions under the optimal transport framework. This property makes Wasserstein barycenters particularly suitable for combining distributions with complex structures, such as class distributions and counterfactuals in our setting, resulting in meaningful prototypes that capture both data variation and class characteristics.

$$\nu^* = \arg\min_{\nu\in\mathcal{P}_2(M)} \sum_{i=1}^{N} \lambda_i W_2^2(\nu,\mu_i)$$

EXPERIMENTS

Metric: Fidelity between the predictions of the target model and surrogate model over a reference dataset

$$\mathrm{Fid}_{m,\mathcal{D}_{\mathrm{ref}}}(\hat{m}) = \frac{1}{|\mathcal{D}_{\mathrm{ref}}|} \sum_{x\in\mathcal{D}_{\mathrm{ref}}} \mathbb{I}_{[0,1]}[\hat{y}_m(x) = \hat{y}_{\hat{m}}(x)]$$

Table 1: Average fidelity with 500, 400, 300 queries on the datasets. Comparison across Adult Income, Compas, DCCC and HELOC. Our method produces the results with the highest fidelity across all settings.

	Query Size (500)			Query Size (400)			Query Size (300)		
	Baseline 1	Baseline 2	Ours	Baseline 1	Baseline 2	Ours	Baseline 1	Baseline 2	Ours
Adult In.	91 ± 3.2	94 ± 3.2	96 ± 2.5	89 ± 3.5	92 ± 3.5	94 ± 2.8	87 ± 3.8	90 ± 3.8	93 ± 3.2
COMPAS	92 ± 3.2	94 ± 2.0	96 ± 2.3	90 ± 3.5	92 ± 2.3	94 ± 2.6	88 ± 3.8	90 ± 2.6	94 ± 3.0
DCCC	89 ± 8.9	91 ± 0.9	97 ± 1.5	87 ± 9.2	89 ± 1.2	95 ± 1.8	85 ± 9.5	87 ± 1.5	93 ± 2.1
HELOC	91 ± 4.7	93 ± 2.2	95 ± 2.0	89 ± 5.0	91 ± 2.5	93 ± 2.3	87 ± 5.3	89 ± 2.8	93 ± 2.6

PROPOSED METHOD: MODEL RECONSTRUCTION VIA WASSERSTEIN BARYCENTERS

Step 1: Computing Barycenters between the data and counterfactuals

For each class $c \in \{0,1\}$, compute a barycenter distribution that balances the original class distribution and the counterfactual distribution with weight λ_c.

$$\min_{\mathbb{Q}_0,\mathbb{Q}_1} \sum_{c\in 0,1} (W_2^2(\mathbb{Q}_c,\mathbb{P}_c) + \lambda_c W_2^2(\mathbb{Q}_c,\mathbb{P}_{cf})) + \gamma\mathcal{R}(\mathbb{Q}_0,\mathbb{Q}_1)$$

where $\mathbb{Q}_c = \arg\min_{\mathbb{Q}\in\mathbb{P}(\mathcal{X})} (W_2^2(\mathbb{Q},\mathbb{P}_c) + \lambda_c W_2^2(\mathbb{Q},\mathbb{P}_{cf}))$

$$\lambda_c = \frac{W_2^2(\mathbb{P}_{cf},\mathbb{P}_{1-c})}{W_2^2(\mathbb{P}_{cf},\mathbb{P}_0)+W_2^2(\mathbb{P}_{cf},\mathbb{P}_1)}$$

This barycenter acts as a soft prototype incorporating different samples. We introduce a symmetry regularization ensuring counterfactuals lie approximately equidistant between barycenters to reflect the decision boundary structure.

$$\mathcal{R}(\mathbb{Q}_0,\mathbb{Q}_1) = (W_2(\mathbb{Q}_0,\mathbb{P}_{cf}) - W_2(\mathbb{Q}_1,\mathbb{P}_{cf}))$$

Step 2: Classification Using Learned Prototypes – classify new inputs by comparing Wasserstein distances

$$\hat{y}_{\hat{m}}(x) = \begin{cases} 0 \text{ if} W_2(\delta_x,\mathbb{Q}_0) < W_2(\delta_x,\mathbb{Q}_1) - \tau \\ 1 \text{ if} W_2(\delta_x,\mathbb{Q}_1) < W_2(\delta_x,\mathbb{Q}_0) - \tau \end{cases}$$

Limitations and Future work: this work demonstrates that Wasserstein barycenters provide a robust framework for classifier reconstruction, particularly in scenarios with limited data and the presence of counterfactual examples. In low-data regimes—where overfitting and poor generalization are common—our approach exhibits superior stability and flexibility compared to baselines. Future research should quantitatively investigate how the size of the available dataset affects the fidelity of model reconstruction. Additionally, exploring alternative prototype representations—beyond the Wasserstein barycenter—could further enhance performance and adaptability.

References

[1] Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020

[2] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.

[3] Dissanayake, P., and Dutta, S.. "Model Reconstruction Using Counterfactual Explanations: A Perspective From Polytope Theory." Model Reconstruction Using Counterfactual Explanations: A Perspective From Polytope Theory. In Advances in Neural Information Processing Systems 38 (NeurIPS 2024).

[4] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." Advances in Neural Information Processing Systems (NeurIPS 2013).

[5] Marco Cuturi and Arnaud Doucet. "Fast Computation of Wasserstein Barycenters." International Conference on Machine Learning (ICML), 2014.

Contact:
Xuan Zhao
xu.zhao@fz-juelich.de

Presenter:
Zhuo Cao
z.cao@fz-juelich.de