# Evaluation

Explanations should have few features, as humans pick usually just a few reasons. They should be specific to the problem at hand, and every instance should be explained in the same deterministic way [1]. Deterministic, or consistent feature attributions [2].

Exploration using several visual aids [3].

Other counterfactual methods mostly focus on readily available datasets such as HELOC loan application [4-6], German Credit Risk [7-11], Diabetes [12-14], Iris [13, 15], Wine Quality [15-17], and many others. The issue with such datasets is that they do not represent real-life scenarios. For example, a user study may be done to evaluate how well users can understand predictions on the value of houses according to the size, age, etc.. However, as the users are not real-estate agents, they lack knowledge of this domain.

Which model-agnostic counterfactual explanation techniques exist and what problems can they be applied to?

[12] Wachter

Two reasons that the authors mention which are also relevant to the current case:

1. Understand why a decision was made
2. Understand how a different decision could be made if certain conditions were changed

“Principally, counterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems”

[4] Grath.

They balance the prediction target with the distance measure using a lambda parameter which is optimized using a tolerated mismatch between target and actual prediction.

[18] factual and counterfactual explanations: building a simple classifier around a local neighborhood of the point of interest. This local neighborhood is created with a genetic algorithm, after which this feature space is described using a decision tree. The path down the tree that corresponds to the current label thus explains the actual prediction, while any path that leads to a different outcome could be interpreted as a counterfactual. They measure minimal changes to be the number of split conditions that need to be changed in order to change the prediction.

Evaluated based on correct classifications of the instances with the decision tree mimic model as compared to the original black-box. Evaluated on various (binary) classification datasets, mostly obtaining outputs with higher fidelity than LIME, even for the local neighborhood.

“In summary, ANCHOR shows better precision than LORE at the expenses of generality and stability of the produced explanations.”

[7] CADEX

the gradient of the loss with respect to the input, which can be followed over the input space with an optimizer (e.g. Adam) until we find an instance x\*, that typically lies on the decision boundary between y^ and y\*.

Code: <https://github.com/spore1/cadex>

[19] Multi-Objective Counterfactuals (MOC) method.

Only evaluated on the 4 objectives posed, in comparison to other state-of-the-art methods (dice, recourse, whatif, tweaking).

**As for counterfactuals; they can be applied to any model and data type. Some counterfactuals might not be feasible. It is therefore important to use training data and constraints to guide the process.**

**They recommend to determine clear desiderata:**

* **Identify the target for which the explanation is intended.**
* **Understand what they are looking for.**
* **Understand how they will engage with the explanation.**

[20]: Feasible and Actionable Counterfactual Explanations (FACE). Their only focus for evaluation is to visualize the path taken by their algorithm. They do show that the counterfactuals by Wachter et al. (left picture) do not fall close to the data distribution.

[5]: ViCE: visual counterfactual explanations

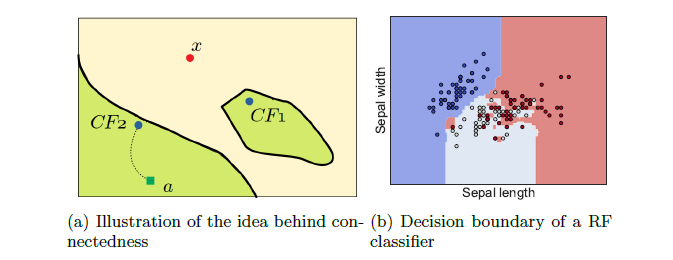
[11] DiCE. Code: <https://github.com/interpretml/DiCE/tree/master/dice_ml>

[17]. Based on the training data. as these counterfactuals are “real experiences” from the problem domain, they have an inherent plausibility”.

[10] MACE: <https://github.com/amirhk/mace>

[9] GeCo: Compares to MACE, CERTIFAI, WIT (what if tool)

A study on justification of counterfactuals [21]. “*Given a classifier f : X ! Y trained on a dataset X, a counterfactual example e <in> X is justified by an instance a <in> X correctly predicted if f(e) = f(a) and if there exists a continuous path h between e and a such that no decision boundary of f is met. Formally, e is justified by a <in> X if: <there is an> h : [0; 1] ! X such that: (i) h is continuous, (ii) h(0) = a, (iii) h(1) = e and (iv) <all>t <in> [0; 1]; f(h(t)) = f(e).”*



They argue that CF2 is better than CF1; in B you can see an application for it (overfit classifier). “CF2 can be connected to a ground-truth instance a without crossing the decision boundary of f and is therefore justified.”

**So instead of choosing instances from the training data per se; they make sure that a counterfactual instance can be “connected” to a ground truth data point without crossing the decision boundary.**

They have written software to asses if counterfactuals are connected to ground-truth data. They sample a circle around an instance, and cluster them using DBSCAN. If they belong to the same cluster as the ground-truth instance, they are justified.

Shows that LORE often creates justified explanations.

<https://github.com/thibaultlaugel/truce>

[22]:

Challenges:

* The evaluation for counterfactual explanations must be done using a user study.
* Counterfactual explanations should be integrated with visualization features.
* Generate robust counterfactual explanations (not overfit to model).
* Unify counterfactuals with traditional XAI

1. Lakkaraju, H., et al., *Interpretable & Explorable Approximations of Black Box Models.* 2017.

2. Lundberg, S. and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*. 2017.

3. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.

4. Grath, R.M., et al., *Interpretable Credit Application Predictions With Counterfactual Explanations.* ArXiv, 2018. **abs/1811.05245**.

5. Gomez, O., et al. *ViCE*. 2020.

6. Russell, C. *Efficient search for diverse coherent explanations*. 2019.

7. Moore, J., N. Hammerla, and C. Watkins, *Explaining deep learning models with constrained adversarial examples*. 2019. p. 43-56.

8. Sokol, K. and P. Flach, *One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency.* KI - Kunstliche Intelligenz, 2020. **34**(2): p. 235-250.

9. Schleich, M., et al., *GeCo: Quality Counterfactual Explanations in Real Time*. 2021.

10. Karimi, A.-H., et al. *Model-agnostic counterfactual explanations for consequential decisions*. in *International Conference on Artificial Intelligence and Statistics*. 2020. PMLR.

11. Mothilal, R.K., A. Sharma, and C. Tan. *Explaining machine learning classifiers through diverse counterfactual explanations*. 2020.

12. Wachter, S., B. Mittelstadt, and C. Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.* Harvard journal of law & technology, 2018. **31**: p. 841-887.

13. White, A. and A. Garcez. *Measurable Counterfactual Local Explanations for Any Classifier*. in *ECAI*. 2020.

14. Sharma, S., J. Henderson, and J. Ghosh. *CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models*. 2020.

15. Rathi, S., *Generating Counterfactual and Contrastive Explanations using SHAP*. 2019.

16. Adhikari, A., et al. *LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models*. in *IEEE International Conference on Fuzzy Systems*. 2019.

17. Keane, M. and B. Smyth, *Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)*. 2020.

18. Guidotti, R., et al., *Factual and Counterfactual Explanations for Black Box Decision Making.* IEEE Intelligent Systems, 2019. **34**(6): p. 14-23.

19. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature – PPSN XVI*. 2020. Cham: Springer International Publishing.

20. Poyiadzi, R., et al., *FACE: Feasible and Actionable Counterfactual Explanations*. 2020. 344-350.

21. Laugel, T., et al., *Unjustified Classification Regions and Counterfactual Explanations in Machine Learning*. 2020. p. 37-54.

22. Verma, S., J.P. Dickerson, and K. Hines, *Counterfactual Explanations for Machine Learning: A Review.* ArXiv, 2020. **abs/2010.10596**.