OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Counterfactual Explanations Using Optimization With Constraint Learning

Donato Maragno D.MARAGNO@UVA.NL
Tabea E. Röber T.E.ROBER@UVA.NL
Ş. İlker Birbil S.I.BIRBIL@UVA.NL
Amsterdam Business School, University of Amsterdam, The Netherlands

Abstract

To increase the adoption of counterfactual explanations in practice, several criteria that these should adhere to have been put forward in the literature. We propose counterfactual explanations using optimization with constraint learning (`CE-OCL`), a generic and flexible approach that addresses all these criteria and allows room for further extensions. Specifically, we discuss how we can leverage an optimization with constraint learning framework for the generation of counterfactual explanations, and how components of this framework readily map to the criteria. We also propose two novel modeling approaches to address data manifold closeness and diversity, which are two key criteria for practical counterfactual explanations. We test `CE-OCL` on several datasets and present our results in a case study. Compared against the current state-of-the-art methods, `CE-OCL` allows for more flexibility and has an overall superior performance in terms of several evaluation metrics proposed in related work.

1. Introduction

Interpretability in machine learning (ML) is an ongoing research field that has received increasing attention in recent years. Off the many approaches and tools for interpretability, counterfactual explanations (CEs) are expected to be especially promising due to their resemblance to how we provide explanations in everyday life [17]. It has been established that we do not seek to explain the cause of an event *per se*, but *relative* to some other event that did not occur. Typically, we have a factual instance vector $\hat{x}$ for which the (prediction) outcome $\hat{y}$ relative to some other, desired, outcome $\tilde{y}$ should be explained. The key idea for generating a CE is to find a data point $\tilde{\boldsymbol{x}}$ close to the factual instance $\hat{\boldsymbol{x}}$, such that the prediction outcome for $\tilde{\boldsymbol{x}}$ is $\tilde{y}$. The difference in the features constitutes the explanation. As CEs do not try to explain all possible causes of an event but focus on necessary changes to the environment to reach a certain state, they tend to be simpler, and with that, also easier to understand than those methods which communicate explanations based on the entire feature space [17].

Wachter et al. [30] are the first to propose an optimization-based approach for generating CEs. Having a trained classifier $h(\cdot)$, the aim is to find at least one CE, say $\tilde{\boldsymbol{x}}$, which has the closest distance to the original factual instance $\hat{\boldsymbol{x}}$ such that $h(\tilde{\boldsymbol{x}})$ is equal to a different target $\tilde{y}$. Such a CE can be obtained by solving the following mathematical optimization model:

$$\min_{\boldsymbol{x}} \max_{\lambda} \lambda(h(\boldsymbol{x}) - \tilde{y})^2 + d(\hat{\boldsymbol{x}}, \boldsymbol{x}), \tag{1}$$

where $d(\cdot,\cdot)$ is a distance function and λ acts as a nonnegative balancing weight to ensure $h(\boldsymbol{x}) = \tilde{y}$. Much work has been devoted to refine this problem such that the generated CEs are useful and attainable in practice. From the literature [e.g., 19, 20, 24, 27, 30], we can identify the following eight criteria that a generated CE should fulfill in both theory and practice: **Proximity:** The CE should be as close as possible to the factual instance $\hat{\boldsymbol{x}}$ with respect to the feature values. **Validity:** The prediction for the CE $\tilde{\boldsymbol{x}}$ should be equal to $\tilde{y} \neq \hat{y}$. **Coherence.** When one-hot encoding is used for categorical data, we should be able to map it back to the input feature space to obtain coherent explanations. **Sparsity:** The CE should differ from the factual instance in as few features as possible. **Actionability:** We can distinguish between immutable, mutable but not actionable, and actionable features. **Data manifold closeness:** To ensure the generation of realistic and actionable explanations, the generated CEs should be close to the observed (training) data. **Causality:** Any (known) causal relationships in the data should be respected in the proposed CEs to further ensure realistic explanations. **Diversity:** Any algorithm for the generation of CEs should return a set of CEs which differ in at least one feature.

These criteria have been partially addressed in recent work, see Table 1. For example, Russell [24] and Ustun et al. [26] address coherence and actionability, the latter introducing the notion of immutable, conditionally immutable and mutable features. Further, Russell [24] focuses on diversity and suggests adding constraints greedily by restricting the state of variables altered in previously generated CEs, while Mothilal et al. [19] base their approach to diverse CEs on determinantal point processes [13]. Kanamori et al. [9] attempt to optimize the idea of proximity and data manifold closeness using *Mahalanobis' distance* and the *local outlier factor* to generate CEs close to the empirical distribution of the training data. Poyiadzi et al. [23] base their work on graph theory, and apply a shortest path algorithm to minimize the f-distance quantifying the trade-off between the path length and the density along this path, by that ensuring a solution that lies in a high density region. To address causality, Kanamori et al. [10] discuss the use of a structural causal model (SCM), while others advocate a post-hoc filtering approach [19]. We refer to Verma et al. [27] and Guidotti [7] for an extensive overview of recent works on counterfactual explanations.

To the best of our knowledge ours is the first work that addresses all of these criteria in a combined setting . We propose `CE-OCL`, a generic and flexible approach for generating CEs based on optimization with constraint learning (OCL). OCL is a new and fast-growing research field whose aim is to learn parts of an optimization model (*e.g.*, constraints or objective function) using ML models whenever explicit formulae are not available (see Fajemisin et al. [5] for a recent survey on OCL). We show how all the criteria proposed in the literature can be addressed by an OCL framework. Based on the concept of trust regions, we also propose a new modeling approach to ensure data manifold closeness and coherence. Finally, we propose using incumbent solutions to obtain diverse CEs in a single execution. With our extensive demonstration on standard datasets from the CE literature, we also set new benchmarks for future research.

2. Generation of counterfactual explanations

In an OCL framework, ML models are used to design constraint and objective functions of an optimization model when explicit expressions are unknown. First, the predictive model is trained on historical data and then it is embedded into the optimization model using decision variables as inputs [2, 28, 29]. Although the interplay between optimization and ML has a different aim in OCL than CE generation, we notice that the two frameworks have a similar structure. Recent

Table 1: State-of-the-art methods to generate CEs

	Proximity	Sparsity	Coherence	Actionability	Data Manifold Closeness	Causality	Diversity
Laugel et al. [14]	●	●	●	–	–	–	–
Russell [24]	●	◐	●	–	–	–	●
Ustun et al. [26]	●	●	●	●	–	–	–
Kanamori et al. [9]	●	–	●	–	●	–	–
Mahajan et al. [15]	●	–	●	◐	●	●	–
Karimi et al. [12]	●	–	●	–	–	●	–
Kanamori et al. [10]	●	●	●	●	–	●	●
Mothilal et al. [19]	●	◐	●	●	–	◐	●
Karimi et al. [11]	●	●	●	●	–	–	●
Poyiadzi et al. [23]	●	–	●	●	●	–	–
`CE-OCL`	●	●	●	●	●	●	●

●: addressed; ◐: partially addressed; –: absent

advances in OCL successfully reduce the computational burden of embedding fitted ML models into an optimization model [6, 18, 25] and can be easily transferred to the problem of generating CEs. In this regard, we show how the problem of generating CEs, given a fitted model $h(\cdot)$, a factual instance $\hat{\boldsymbol{x}}$, and the desired outcome $\tilde{y}$, can be seen as a special case of *optimization with constraint learning*. In an OCL setting, a dataset $\mathcal{D} = \{(\bar{\boldsymbol{x}}_{\boldsymbol{i}}, \bar{y}_i)\}_{i=1}^N$ with observed feature vector $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ and outcome of interest $\bar{y}_i$ for sample i, is used to train predictive models that are to be constrained or optimized in a larger optimization problem. An OCL model is typically presented as

$$\begin{aligned}
\underset{\boldsymbol{x}\in\mathbb{R}^n, y\in\mathbb{R}}{\text{minimize}} \quad & f(\boldsymbol{x}, y) && \text{(2a)}\\
\text{subject to} \quad & \boldsymbol{g}(\boldsymbol{x}, y) \leq \mathbf{0}, && \text{(2b)}\\
& y = h(\boldsymbol{x}), && \text{(2c)}\\
& \boldsymbol{x} \in \mathcal{X}, && \text{(2d)}
\end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the decision vector with components $x_i \in \mathbb{R}$, $f(\cdot,\cdot) : \mathbb{R}^{n+1} \mapsto \mathbb{R}$ and $\boldsymbol{g}(\cdot,\cdot) : \mathbb{R}^{n+1} \mapsto \mathbb{R}^m$ are known functions possibly also depending on the predicted outcome y, and $h(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ represents the predictive model[1] trained on $\mathcal{D}$. The set $\mathcal{X}$ defines the trust region, *i.e.*, the set of solutions for which we trust the embedded predictive models (see below for details). Formulation (2a-2d) is quite general and encompasses a large body of work that includes CE generation. Now, we characterize the parallelism between some of the eight criteria listed in Section 1 and the structure of the resulting OCL model. We elaborate and discuss the remaining criteria in Appendix A.

Validity. While the trained model $h(\cdot)$ is used in constraint learning to define, completely or partially, the objective function and/or the constraints, in CE generation it is used to enforce the validity constraint. Constraint (2c) is likely to be an encoding of the predictive model. In other words, embedding a trained ML model requires adding multiple constraints and auxiliary variables. When $h(\cdot)$ is a classification model, the CE validity is obtained by constraining the model prediction to be equal to the desired class $\tilde{y}$; that is, we set $y = \tilde{y}$. If $h(\cdot)$ is a regression model, the OCL framework

1. To simplify our exposition, we include only one predictive model. However, a general OCL framework admits multiple learned constraints in the model.

still applies, and an inequality constraint can be used to enforce validity; *e.g.*, $y \leq \tilde{y} - \delta$ or $y \geq \tilde{y} + \delta$ for some fixed $\delta \in \mathbb{R}_+$.

Data manifold closeness. One of the requirements to obtain plausible CEs is that they are close to the data manifold. For this purpose, we can make use of the *trust region* constraints. Maragno et al. [16] define the trust region as the convex hull (CH) of $\mathcal{D}$ in the features space, and they use it in OCL to prevent the trained model from extrapolating, therefore, mitigating the deterioration in predictive performance for points that are farther away from the data points in $\mathcal{D}$. In CE generation, the trust region, or rather *data manifold region*, serves the purpose of ensuring solutions in a high-density region. To this end, we can also denote a CE ($\tilde{\boldsymbol{x}}$) as the convex combination of samples in $\mathcal{D}$, in particular samples belonging to the desired class ($\tilde{y}$).

In case the CH is too restrictive, we can use a relaxed formulation to enlarge the data manifold by including those solutions that are in the ϵ-ball surrounding some feasible solutions in the CH:

$$\epsilon\text{-CH} = \left\{ \boldsymbol{x} \middle| \sum_{i \in \mathcal{I}} \lambda_i \bar{\boldsymbol{x}}_i = \boldsymbol{x} + \boldsymbol{s},\ \sum_{i \in \mathcal{I}} \lambda_i = 1,\ \boldsymbol{\lambda} \geq 0,\ ||\boldsymbol{s}||_p \leq \epsilon \right\}, \tag{3}$$

where $\lambda_i \in [0, 1]$ and $\boldsymbol{s} \in \mathbb{R}^n$ are auxiliary variables, $\epsilon \geq 0$ is a hyperparameter, and $\mathcal{I}$ denotes the indices corresponding to the subset of samples in $\mathcal{D}$ belonging to the desired class $\tilde{y}$. When $\epsilon = 0$, we obtain the trust region as discussed in Maragno et al. [16]. However, $\epsilon > 0$ leads to a less restrictive set of conditions. Further details are available in Appendix A as well as a graphical representation of the data manifold closeness in Figure 2.

Causality. CEs might be inefficient or unrealistic when causal relations are not considered in the generation process. Both these situations are exemplified in Karimi et al. [12], where the authors show the importance of causal relations to obtain CEs that better answer the question "what *should be done* in the future considering the laws governing the world." When a causal model is available, we can formulate the causal relations among variables as extra constraints of the optimization model. When there is not an explicit formulation of the causal relations, we are in a typical constraint learning scenario where an ML model can be trained and embedded into the optimization. We provide the formulation of causality constraints in Appendix A.

Diversity. Most of the methods for generating multiple and diverse CEs in the literature require multiple runs and extra constraints to generate diverse CEs for the same input. Following an iterative approach, we can generate diverse CEs using constraints on the actionability of features [24], or constraints on the distance between the subsequent CE and all the previously generated ones [11]. Again in an iterative way, we can also use the data manifold constraints to generate diverse CEs (i) by finding one CE for each clustered CH, (ii) by enlarging the CH with increasing ϵ whenever the data manifold constraints are active. The use of diversity constraints offers great flexibility at the expense of computation time. As an alternative, we propose to solve one single optimization model and use the pool of *incumbent solutions* as the set of CEs. In mixed-integer optimization, solvers like Gurobi or CPLEX allow retrieving the sub-optimal solutions found during the tree search procedure [3, 8]. In this way, collecting a set of CEs comes at no cost in terms of computation time.

3. Experiments and results

In this section, we demonstrate the effectiveness of OCL through empirical experiments on multiple datasets and comparing the results with other state-of-the-art methods. The experiments are

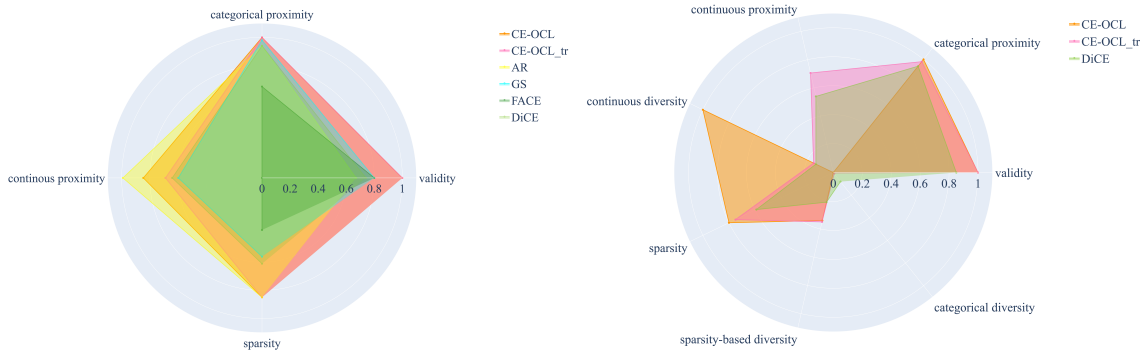

Figure 1: Performance of `CE-OCL` and `CE-OCL_tr` (with trust region) compared to performance of current state-of-the-art methods for generating counterfactual explanations on the COMPAS dataset. Left: we generated one counterfactual for each of 30 factual instances; rf as predictive model. Right: we generated three counterfactuals for the same instances; lr as predictive model. Not all methods support generating several counterfactuals.

executed using `OptiCL` [2] [16], an open-source Python package for optimization with constraint learning. `OptiCL` has been originally designed to help practitioners in modeling an optimization problem whose constraints are partially unknown, but where ML models can be deployed to learn them [16]. However, as detailed in Section 2, the problem of generating CEs directly relates to an OCL problem. `OptiCL` currently supports several MIO-representable predictive models, including logistic regression (lr), support vector machines (svm), (optimal) decision trees (cart), random forests (rf), gradient boosting machines (gbm), and neural networks with ReLU activation functions (mlp). Moreover, `OptiCL` allows for trust region constraints as defined in (3). Whenever a causal model is available but the relations are not explicit, `OptiCL` allows representing the relation using one of the MIO-representable ML models. The open-source implementation for reproducing all our results is available at `https://github.com/tabearoeber/CE-OCL`.

We performed an extensive comparison of our method against four state-of-the-art methods: `Growing Spheres (GS)` [14], `FACE`[23], `Actionable Recourse (AR)` [26], and `DiCE` [19]. Here, we present the results on the COMPAS dataset[3]. We generated CEs for 30 factual instances, then averaged the scores on the evaluation metrics proposed by Mothilal et al. [19]. Our results are presented in Figure 1 (see Appendix B for details on the evaluation metrics). For the sake of clarity, we have rescaled values such that they range from 0 (worst) to 1 (best). Since the majority of the other methods do not generate a set of CEs, we chose to generate only one CE for each factual instance and hence do not report any diversity scores (Figure 1 left). Furthermore, we compare our approach in terms of diversity by generating three CEs for each of the 30 factual instances and compare the results with `DiCE` [19] (Figure 1 right). Extensive results for all datasets and with different predictive models are included in Appendix B. We further demonstrate the generation of CEs in a step-wise manner on the Statlog (German Credit Data) dataset [4], which is one of the standard datasets in the CE literature[4]. The German Credit dataset classifies people described by a set of 20 features as good or bad credit risk, see Table 4 in Appendix C.1 for an overview of the features. For this demonstration, we gradually add constraints to the model and present

2. https://github.com/hwiberg/OptiCL, under the MIT license.
3. Appendix B includes the results for three further datasets: Adult, Give Me Some Credit, and HELOC.
4. We also provide another demonstration on the Statlog (Heart) dataset [4] in Appendix C.

the generated CEs at each step in Table 5 shown in Appendix C.1. The table is divided into six parts (A-F), each showing the set of CEs generated, and a dash is used to represent no change to the corresponding features. In Table 5 in Appendix C.1, we present the evaluation of these CEs using several evaluation metrics proposed by Mothilal et al. [19]: validity, sparsity, categorical and continuous proximity, categorical and continuous diversity, and sparsity-based diversity. The complete mathematical model is detailed in Appendix C.2.

We fit several ML models to the data, all of which performed similarly well. For demonstration purposes, we have chosen a linear support vector machine. The factual instance $\hat{\boldsymbol{x}}$ used for this case study is reported in Table 5. We start the demonstration considering only validity, proximity, and coherence (Part A), and using the ℓ_2-norm as a distance function. The optimal solution suggests several changes in the factual instance and is not actionable in practice due to the negative value for F2 (credit amount). To induce sparsity (Part B), we use auxiliary variables to keep track of the number of features changed and penalize them in the objective function. Multiple and diverse CEs are generated using incumbent solutions (Part C). To ensure that the set of generated CEs is valuable in practice, we add actionability constraints (Part D). Respecting these constraints, the set of generated CEs seems more realistic however, they may still not be attainable in practice. Specifically, if we consider solution (c) of Part D, the only suggested change concerns F4 (age). However, this CE is unlikely to represent a realistic data point, considering the other feature values remain unchanged. In other words, CEs that do not resemble the training data come with the risk of being unattainable in practice. To this end, we use the idea of a *data manifold region*, as detailed in Section 2. As a result, in Part E, we obtain a more realistic set of CEs, although at the expense of sparsity and (categorical) proximity (see the scores reported in Table 5, Appendix C.1). From a qualitative point of view, the three CEs show a more sensible combination of feature values compared to those in Part D. Finally, we can leverage the partial SCM provided by Karimi et al. [12] for this dataset, which shows that F1 (duration) is causally related to F2 (credit amount). This relationship is learned by a multi-layer perceptron (MLP) using 5-fold cross validation. In Part F, we display the set of CEs that satisfy also the learned causality constraints.

4. Discussion

With this work, we propose `CE-OCL`, a generic approach for generating sensible and practical counterfactual explanations. In Section 3, we report the generally superior performance achieved by `CE-OCL` compared to other popular methods. Nevertheless, we acknowledge the limitations of using incumbent solutions as multiple counterfactuals caused by the lack of control over the solutions' diversity. Whenever we have specific diversity requirements to meet, the iterative approaches proposed by Russell [24] and Karimi et al. [11] may suit best. Moreover, owing to the MIO structure of `CE-OCL` and various constraints used to satisfy the established criteria, the feasibility space may shrink to the point of being empty, making the optimization problem infeasible. In the infeasibility case, we recommend following an approach similar to that presented in Section 3, where constraints are added one at a time. Infeasibility problems due to data manifold constraints can be mitigated by enlarging the data manifold region at the (potential) expense of the sensibility of the CEs. For future research, we plan to investigate the effect of clustering and enlargement of the data manifold region on the CE quality and on diversity. We also intend to extend `CE-OCL` with additional criteria like robustness in the sense that the generated CEs are not point solutions, but that they are defined by ranges in the feature values.

Acknowledgments

This work was supported by the Dutch Scientific Council (NWO) grant OCENW.GROOT.2019.015, Optimization for and with Machine Learning (OPTIMAL).

References

[1] Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485*, 2021.

[2] Max Biggs, Rim Hariss, and Georgia Perakis. Optimizing objective functions determined from random forests. *SSRN Electronic Journal*, pages 1–46, 2021. ISSN 1556-5068.

[3] IBM ILOG Cplex. V12. 1: User's manual for CPLEX. *International Business Machines Corporation*, 46(53):157, 2009.

[4] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu.

[5] Adejuyigbe Fajemisin, Donato Maragno, and Dick den Hertog. Optimization with constraint learning: a framework and survey. *arXiv preprint arXiv:2110.02121*, 2021.

[6] Bjarne Grimstad and Henrik Andersson. ReLU networks as surrogate models in mixed-integer linear programs. *Computers & Chemical Engineering*, 131:106580, dec 2019.

[7] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, April 2022.

[8] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2022. URL https://www.gurobi.com.

[9] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. DACE: distribution-aware counterfactual explanation by mixed-integer linear optimization. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2855–2862, 2020.

[10] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. Ordered counterfactual explanation by mixed-integer linear optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11564–11574, May 2021.

[11] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 895–905. PMLR, 26–28 Aug 2020.

[12] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

[13] Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

[14] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

[15] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[16] Donato Maragno, Holly Wiberg, Dimitris Bertsimas, S. Ilker Birbil, Dick den Hertog, and Adejuyigbe Fajemisin. Mixed-integer optimization with constraint learning. *arXiv preprint arXiv:2111.04469*, 2021.

[17] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2018. ISSN 0004-3702.

[18] Velibor V. Mišić. Optimization of tree ensembles. *Operations Research*, 68(5):1605–1624, 2020. ISSN 15265463.

[19] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[20] Guillermo Navas-Palencia. Optimal counterfactual explanations for scorecard modelling. *arXiv preprint arXiv:2104.08619*, 2021.

[21] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms, 2021.

[22] Judea Pearl. Structural counterfactuals: a brief introduction. *Cognitive Science*, 37(6):977–985, 2013. ISSN 1551-6709.

[23] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, feb 2020.

[24] Chris Russell. Efficient search for diverse coherent explanations. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[25] Artur M. Schweidtmann and Alexander Mitsos. Deterministic global optimization with artificial neural networks embedded. *Journal of Optimization Theory and Applications*, 180(3): 925–948, 2019. ISSN 15732878.

[26] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[27] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: a review. *arXiv preprint arXiv:2010.10596*, 2020.

[28] S. Verwer, Y. Zhang, and Q. C. Ye. Auction optimization using regression trees and linear models as integer programs. *Artificial Intelligence*, 244:368–395, 2017.

[29] G. Villarrubia, J. F. De Paz, P. Chamoso, and F. De la Prieta. Artificial neural networks used in optimization problems. *Neurocomputing*, 272:10–16, 2018.

[30] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2), 2018.

Appendix A. Generate counterfactuals

In Section 2, we describe how criteria such as validity, closeness, causality, and diversity can be fulfilled exploiting OCL components. Likewise, other criteria can be mathematically represented in the following way:

Proximity. By definition, a CE has to be in the proximity of the factual instance according to some user-defined distance function. To obtain a CE $\tilde{\boldsymbol{x}}$ in the proximity of $\hat{\boldsymbol{x}}$, we can write the objective function (2a) as a distance function $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$. In the literature, this function is represented by ℓ_1-norm, ℓ_2-norm, or as the Mahalanobis' distance.

Coherence. When one-hot encoding is used to deal with categorical features, we can use the constraints proposed by Russell [24] to obtain coherent CEs. That is, we write for k categorical features the following constraints:

$$\sum_{j' \in \mathcal{C}_j} x_{j'} = 1, \;\; j = 1, \ldots, k, \tag{4}$$

where $\mathcal{C}_j$ is a set of indices referring to the dummy (binary) variables used to represent the categorical feature j. The use of a data manifold region (with a sufficiently small ϵ) has an interesting impact on CE coherence because constraints (4) become redundant. To exemplify how data manifold constraints guarantee coherence, we consider a set of samples represented by the set of indices $\mathcal{I}$, and a categorical feature *diet* that can assume only three values: *vegan*, *vegetarian*, or *omnivore*. We use one-hot encoding to replace the feature *diet* and describe a CE with the dummy (binary) variables x_{vegan}, $x_{vegetarian}$, $x_{omnivore}$. From (3), we have

$$x_j = \sum_{i \in \mathcal{I}} \lambda_i \bar{x}_{i,j}, \;\; j \in \{vegan, vegetarian, omnivore\},$$

with $\sum_{i \in \mathcal{I}} \lambda_i = 1$. One of the dummy variables, say x_{vegan}, can assume value 1 only if it is the convex combination of data points $\bar{\boldsymbol{x}}_i$ with $\bar{x}_{i,vegan} = 1$ and $\bar{x}_{i,vegeterian} = \bar{x}_{i,omnivore} = 0$. Thus, $\lambda_i > 0$ only when $\bar{x}_{i,vegan} = 1$, and consequently, we obtain $x_{vegetarian} = x_{omnivore} = 0$.

Sparsity. The sparsity can be handled by enforcing the following set of constraints:

$$|x_j - \hat{x}_j| \leq M z_j, \;\; j = 1, \ldots, n, \tag{5a}$$

$$\sum_{i=1}^{n} z_i \leq K, \tag{5b}$$

where $z_j \in \{0, 1\}$, $j = 1, \ldots, n$ are auxiliary variables that are simply used to count the number of features in $\boldsymbol{x}$ that differ from $\hat{\boldsymbol{x}}$, and K is an upper bound on the number of allowed changes. Alternatively, constraints (5b) can be relaxed and moved to the objective function with a scaling penalty factor $\alpha > 0$. That is, we obtain the new objective function $f(\boldsymbol{x}, y) + \alpha \sum_{i=1}^{n} z_i$. Though simpler, this relaxation does not guarantee to lead to an optimal solution with less than or equal to K changes.

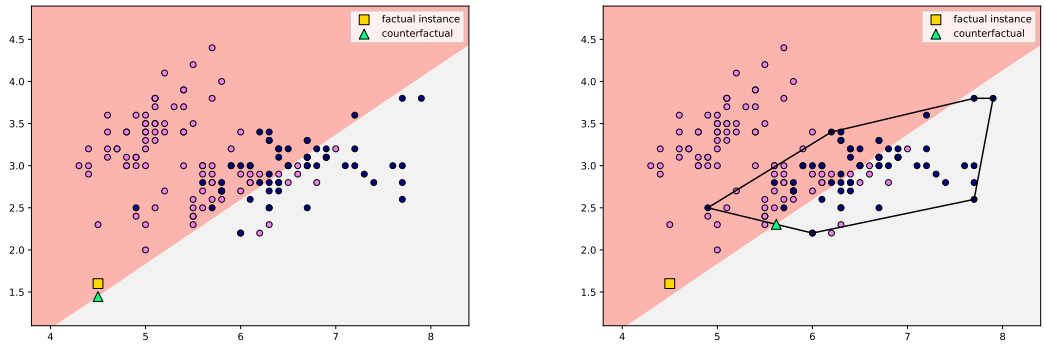

Figure 2: The effect of the data manifold region on the generated CE. The left figure shows the factual instance and its closest counterfactual without closeness constraints. The right figure shows the same factual instance with the CE constrained to be within the data manifold region.

Actionability. As a recommended CE should never change the immutable features, we can restrict the CE to be equal to the factual instance for all the immutable features. Suppose that the set of immutable features is represented by $\mathcal{I}_m$, then we simply add the following constraints:

$$x_i = \hat{x}_i, \quad i \in \mathcal{I}_m. \tag{6}$$

Other feasibility constraints might concern actionable variables that cannot take certain values, such as *age*, which can only be increased, or *has_phd*, which can only change from false to true. These conditions can be added exactly like immutable features.

Data manifold region. Figure 2 shows how the CH of $\mathcal{D}$ in the features space ensures a solution closer to the data manifold, leading to more plausible CEs. In some cases, the CH may be too restrictive, which is why we introduce formulation 3 to enlarge the data manifold region by including solutions that are in the ϵ-ball around some feasible solutions in the CH. Being able to enlarge the data manifold region represents a solution to the criticism by Balestriero et al. [1]: "[...] interpolation[5] almost surely never occurs in high-dimensional spaces (> 100) regardless of the underlying intrinsic dimension of the data manifold." Aside from the bound on the norm of $\boldsymbol{s}$, all constraints in (3) are linear. Fortunately, the most common norms used to constraint $\boldsymbol{s}$ are ℓ_1-, ℓ_2-, or ℓ_∞-norm. These norms lead to convex conic constraints that can be handled easily with off-the-shelf optimization solvers. The effectiveness of the data manifold region might be hampered by the fact that the CH includes low-density regions. In this case, Maragno et al. [16] advocate a two-step approach: first, clustering is used to identify distinct high-density regions, and then, the data manifold region is represented as the union of the (enlarged) convex hulls of the individual clusters.

Causality The causality constraints are modelled by applying the Abduction-Action-Prediction steps [22], Karimi et al. [12] define the endogenous variables (with indices in the set $\mathcal{E}$) as

$$x_i = \hat{x}_i + c_i(\boldsymbol{p}_i) - c_i(\hat{\boldsymbol{p}}_i), \quad i \in \mathcal{E}, \tag{7}$$

where $c_i(\boldsymbol{p}_i)$ is a function of the parents of x_i, namely the predecessors of the feature i in the SCM. Both $\hat{x}_i$ and $c_i(\hat{\boldsymbol{p}}_i)$ are known before the optimization and therefore treated as parameters. When

5. Interpolation occurs for a sample $\boldsymbol{x}$ whenever this sample belongs to the CH of a set of data points.

there is not an explicit formulation of $c_i(\cdot)$, we are in a constraint learning scenario where an ML model can be trained and embedded into the optimization as $c_i = h_i(\boldsymbol{p}_i)$ for all $i \in \mathcal{E}$.

Appendix B. Comparison against other methods

We compared `CE-OCL` to four open-source tools for generating CEs: `Growing Spheres` [14], `FACE`[23], `Actionable Recourse` [26], and `DiCE` [19]. The experiments are performed using `CARLA` [21], a Python library to benchmark counterfactual explanation and recourse models. The predictive model used in the experiments is a random forest and the evaluation is performed by generating a counterfactual for 30 different factual instances on four datasets available in `CARLA`: Adult, Give Me Some Credit, COMPAS, and HELOC. We average the results for the evaluation metrics proposed by Mothilal et al. [19] and present them together with the standard error (s.e.) in Table 2. Validity, sparsity, categorical proximity, categorical diversity, and sparsity-based diversity range in the interval [0,1], where 0 and 1 represent the worst and the best scores ($\uparrow_0^1$), respectively. Continuous diversity is a positive number, and the higher it is, the better ($\uparrow_0^+$). Continuous proximity is a negative number, and the closer it is to 0, the better ($\uparrow_-^0$).

While `CE-OCL` can deal with causality and closeness constraints, this does not apply to `DiCE` which uses a post-hoc filtering approach to remove unrealistic CEs. In addition to causality and closeness constraints, `Actionable Recourse`, and Growing Sphere cannot generate more than one counterfactual for each instance. `FACE` does not support diversity and causality constraints but it is able to generate CEs close to the data manifold region. Therefore, in Table 2 we report both the results obtained with `CE-OCL` including validity, proximity, coherence, sparsity, and immutability constraints, and the results obtained including also the closeness constraints, `CE-OCL_tr`. The results show that, across all datasets, both `CE-OCL` and `CE-OCL_tr` exhibit better performance in terms of validity, categorical proximity, and sparsity. `Actionable Recourse` and `CE-OCL`/`CE-OCL_tr` perform equally well in terms of continuous proximity.

We performed a more thorough comparison between `CE-OCL` and `DiCE` on the same four datasets but this time generating three CEs for each instance and using all the predictive models supported by both `OptiCL` and `DiCE`. In Table 3, we report the results obtained with `CE-OCL` including validity, proximity, coherence, sparsity, diversity, and actionability together with the results obtained considering also the data manifold closeness, (`CE-OCL_tr`). The results clearly show how `CE-OCL` outperforms `DiCE` in terms of validity, categorical proximity, continuous proximity, and sparsity. While both methods have a categorical diversity score very close to zero in every scenario, `DiCE` has a generally better performance in terms of continuous diversity. Similarly, `DiCE` has a better sparsity-based diversity score with the exception of the COMPAS dataset. The addition of closeness constraints (`CE-OCL`_tr) has a negative effect on the sparsity and proximity scores but it positively affects the diversity scores when compared to `CE-OCL`. This was to be expected, as the data manifold region forces solutions to be located in a high-density region, which might lead to optimal solutions with more feature changes. While the sparsity decreases, this loss comes at a high potential of more valuable counterfactuals.

Table 2: Comparison of CE-OCL with DiCE (genetic), Algorithmic Recourse, Growing spheres, and FACE using Random Forest as predictive model.

		validity ($\uparrow_0^1$) mean (s.e.)	**cat. proximity** ($\uparrow_0^1$) mean (s.e.)	**cont. proximity** ($\uparrow_-^0$) mean (s.e.)	**sparsity** ($\uparrow_0^1$) mean (s.e.)
ADULT	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	-4844.35 (575.93)	**0.93** (0.00)
	CE-OCL_tr	**1.00** (0.00)	0.97 (0.02)	-21785.50 (7506.35)	0.86 (0.01)
	DiCE	**1.00** (0.00)	0.74 (0.03)	-84278.61 (11613.30)	0.50 (0.02)
	Actionable Recourse	**1.00** (0.00)	0.78 (0.07)	**0.00** (0.00)	0.89 (0.04)
	Growing Spheres	0.80 (0.07)	0.95 (0.01)	-78901.08 (10395.08)	0.59 (0.01)
	FACE	0.80 (0.07)	0.65 (0.03)	-108614.33 (18804.05)	0.47 (0.02)
COMPAS	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	-15.23 (5.84)	**0.85** (0.01)
	CE-OCL_tr	**1.00** (0.00)	1.00 (0.00)	-30.83 (16.00)	**0.85** (0.01)
	DiCE	0.74 (0.09)	0.94 (0.03)	-35.58 (9.59)	0.61 (0.01)
	Actionable Recourse	0.67 (0.11)	0.94 (0.03)	**-0.87** (0.10)	**0.85** (0.01)
	Growing spheres	0.80 (0.07)	0.98 (0.02)	-39.88 (6.58)	0.56 (0.01)
	FACE	0.80 (0.07)	0.65 (0.05)	-98.83 (21.18)	0.37 (0.03)
HELOC	CE-OCL	**1.00** (0.00)	–	**-12.21** (2.67)	**0.94** (0.01)
	CE-OCL_tr	**1.00** (0.00)	–	-92.71 (13.75)	0.75 (0.02)
	DiCE	0.97 (0.03)	–	-203.83 (13.70)	0.22 (0.02)
	Actionable Recourse*	–	–	–	–
	Growing spheres	0.77 (0.08)	–	-87.36 (13.16)	0.00 (0.00)
	FACE	0.77 (0.08)	–	-361.80 (25.20)	0.17 (0.02)
CREDIT	CE-OCL	**1.00** (0.00)	–	**-1.18** (0.66)	**0.90** (0.01)
	CE-OCL_tr	**1.00** (0.00)	–	-120.61 (118.54)	0.87 (0.01)
	DiCE	**1.00** (0.00)	–	-1618.33 (305.53)	0.25 (0.02)
	Actionable Recourse	0.83 (0.17)	–	-8.47 (7.96)	0.88 (0.02)
	Growing spheres	0.63 (0.09)	–	-47.73 (27.24)	0.10 (0.00)
	FACE	0.63 (0.09)	–	-3001.73 (430.61)	0.11 (0.02)

For the comparison, one counterfactual was generated for each of 30 factual instances.
The scores were averaged over all instances, and the standard error was derived.
* For the Heloc dataset, Actionable Recourse did not yield any counterfactuals for any of the thirty factual instances.

Table 3: Comparison of CE-OCL, CE-OCL with trust region, and DiCE (genetic) with a range of preditive models.

		validity($\uparrow^1_0$) mean (s.e.)	**cat. proximity** ($\uparrow^1_0$) mean (s.e.)	**cont. proximity**($\uparrow^0_-$) mean (s.e.)	**sparsity**($\uparrow^1_0$) mean (s.e.)	**cat. diversity** ($\uparrow^1_0$) mean (s.e.)	**cont. diversity** ($\uparrow^+_0$) mean (s.e.)	**sparsity-based diversity**($\uparrow^1_0$) mean (s.e.)
Adult dataset								
	CE-OCL	**1.00** (0.00)	**0.99** (0.00)	**-6775.61** (953.24)	**0.92** (0.00)	0.01 (0.01)	5796.61 (1056.70)	0.11 (0.01)
rf	CE-OCL_tr	**1.00** (0.00)	0.97 (0.02)	-25044.04 (7645.82)	0.86 (0.01)	0.00 (0.00)	13288.83 (3236.28)	0.13 (0.01)
	DiCE	1.00 (0.00)	0.74 (0.02)	-81581.97 (8003.12)	0.51 (0.01)	**0.19** (0.02)	**79595.54** (7496.70)	**0.26** (0.02)
	CE-OCL	**1.00** (0.00)	**0.99** (0.01)	**-4226.05** (794.33)	**0.89** (0.01)	0.01 (0.01)	8042.59 (1418.66)	0.19 (0.01)
lr	CE-OCL_tr	**1.00** (0.00)	0.96 (0.02)	-23288.53 (6837.94)	0.84 (0.01)	0.05 (0.02)	23421.06 (7376.79)	0.22 (0.01)
	DiCE	0.71 (0.05)	0.68 (0.03)	-111661.60 (10261.28)	0.47 (0.02)	**0.30** (0.03)	**110668.36** (11019.56)	**0.35** (0.02)
	CE-OCL	**1.00** (0.00)	0.88 (0.02)	**-13994.62** (4198.07)	0.83 (0.02)	0.21 (0.04)	26889.72 (8343.13)	**0.27** (0.03)
cart	CE-OCL_tr	**1.00** (0.00)	**0.94** (0.01)	-19490.98 (5725.22)	**0.84** (0.01)	0.08 (0.02)	20313.32 (5702.44)	0.19 (0.01)
	DiCE	0.65 (0.06)	0.77 (0.02)	-91507.16 (10271.17)	0.55 (0.02)	**0.23** (0.02)	**84111.77** (11802.96)	0.25 (0.01)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-10553.10** (1842.00)	**0.88** (0.01)	0.00 (0.00)	2538.19 (669.37)	0.15 (0.02)
mlp	CE-OCL_tr	**1.00** (0.00)	0.97 (0.02)	-21467.01 (7465.77)	0.86 (0.01)	0.00 (0.00)	6229.99 (2385.40)	0.14 (0.01)
	DiCE	0.62 (0.06)	0.67 (0.03)	-89314.87 (10646.04)	0.47 (0.02)	**0.26** (0.03)	**83848.59** (11476.56)	**0.31** (0.02)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-2488.82** (865.72)	**0.91** (0.00)	0.00 (0.00)	4475.59 (1737.16)	0.13 (0.01)
gbm	CE-OCL_tr	**1.00** (0.00)	0.97 (0.01)	-22732.34 (7545.41)	0.88 (0.01)	0.02 (0.01)	10609.72 (3583.71)	0.15 (0.01)
	DiCE	0.91 (0.04)	0.70 (0.02)	-113297.47 (11606.39)	0.49 (0.01)	**0.25** (0.02)	**75728.74** (10042.23)	**0.27** (0.02)
COMPAS dataset								
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-18.79** (6.15)	**0.85** (0.00)	0.00 (0.00)	8.87 (4.30)	0.17 (0.01)
rf	CE-OCL_tr	**1.00** (0.00)	**1.00** (0.00)	-38.02 (16.46)	**0.85** (0.01)	0.00 (0.00)	9.41 (4.49)	0.16 (0.01)
	DiCE	0.81 (0.04)	0.96 (0.01)	-40.94 (7.06)	0.60 (0.01)	**0.06** (0.02)	**22.30** (5.78)	**0.20** (0.01)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	-121.74 (13.72)	**0.80** (0.01)	0.00 (0.00)	**229.89** (28.27)	0.34 (0.01)
lr	CE-OCL_tr	**1.00** (0.00)	0.98 (0.01)	**-35.84** (11.03)	0.75 (0.01)	0.01 (0.01)	34.68 (7.79)	**0.35** (0.01)
	DiCE	0.85 (0.05)	0.94 (0.02)	-56.05 (12.83)	0.59 (0.01)	**0.08** (0.02)	30.61 (9.48)	0.21 (0.02)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-23.14** (6.03)	**0.84** (0.01)	0.00 (0.00)	**33.38** (11.83)	**0.19** (0.01)
cart	CE-OCL_tr	**1.00** (0.00)	**1.00** (0.00)	-28.43 (9.16)	0.83 (0.01)	0.00 (0.00)	31.16 (10.10)	**0.19** (0.01)
	DiCE	0.77 (0.08)	0.96 (0.01)	-32.99 (6.04)	0.60 (0.01)	**0.07** (0.02)	24.31 (6.63)	0.18 (0.01)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-16.55** (2.11)	0.81 (0.01)	0.00 (0.00)	16.20 (4.62)	**0.22** (0.01)
mlp	CE-OCL_tr	**1.00** (0.00)	**1.00** (0.00)	-27.75 (10.25)	**0.82** (0.01)	0.00 (0.00)	7.29 (4.09)	0.18 (0.01)
	DiCE	0.82 (0.06)	0.96 (0.01)	-59.11 (13.01)	0.58 (0.01)	**0.06** (0.02)	**24.95** (5.78)	**0.22** (0.02)
	CE-OCL	**1.00** (0.00)	**1.00** (0.00)	**-10.10** (2.60)	**0.86** (0.00)	0.00 (0.00)	13.64 (3.09)	**0.21** (0.01)
gbm	CE-OCL_tr	**1.00** (0.00)	**1.00** (0.00)	-25.70 (10.66)	0.85 (0.01)	0.00 (0.00)	13.49 (5.08)	0.20 (0.01)
	DiCE	0.59 (0.07)	0.96 (0.01)	-42.64 (6.32)	0.60 (0.01)	**0.08** (0.02)	**24.51** (5.81)	0.20 (0.01)
Heloc dataset								
	CE-OCL	**1.00** (0.00)	–	**-13.53** (2.35)	**0.93** (0.00)	–	9.94 (2.74)	0.09 (0.01)
rf	CE-OCL_tr	**1.00** (0.00)	–	-94.24 (13.68)	0.75 (0.02)	–	18.93 (4.62)	0.24 (0.02)
	DiCE	0.90 (0.03)	–	-231.05 (11.17)	0.21 (0.02)	–	**223.91** (14.16)	**0.61** (0.02)
	CE-OCL	**1.00** (0.00)	–	**-99.09** (14.22)	**0.88** (0.01)	–	188.16 (28.38)	0.21 (0.01)
lr	CE-OCL_tr	**1.00** (0.00)	–	-138.29 (16.52)	0.72 (0.02)	–	72.51 (8.39)	0.34 (0.02)
	DiCE	0.70 (0.06)	–	-232.39 (12.87)	0.21 (0.02)	–	**207.02** (11.34)	**0.61** (0.02)
	CE-OCL	**1.00** (0.00)	–	**-13.12** (1.40)	**0.95** (0.00)	–	19.72 (2.41)	0.08 (0.00)
cart	CE-OCL_tr	**1.00** (0.00)	–	-99.05 (13.45)	0.73 (0.02)	–	41.03 (6.53)	0.31 (0.02)
	DiCE	0.80 (0.07)	–	-216.70 (13.47)	0.22 (0.02)	–	**234.89** (16.10)	**0.61** (0.02)
	CE-OCL	**1.00** (0.00)	–	**-25.09** (7.57)	**0.92** (0.01)	–	21.30 (4.18)	0.12 (0.01)
mlp	CE-OCL_tr	**1.00** (0.00)	–	-98.94 (15.87)	0.75 (0.02)	–	15.41 (5.52)	0.26 (0.02)
	DiCE	0.67 (0.07)	–	-252.56 (14.17)	0.20 (0.02)	–	**246.96** (16.31)	**0.61** (0.02)
	CE-OCL	**1.00** (0.00)	–	**-8.41** (2.45)	**0.94** (0.00)	–	16.31 (4.92)	0.10 (0.00)
gbm	CE-OCL_tr	**1.00** (0.00)	–	-89.91 (14.70)	0.76 (0.02)	–	18.70 (6.87)	0.25 (0.02)
	DiCE	0.73 (0.08)	–	-234.96 (11.60)	0.22 (0.02)	–	**248.95** (17.34)	**0.59** (0.02)
Give me some credit dataset								
	CE-OCL	**1.00** (0.00)	–	**-6.77** (4.43)	**0.90** (0.00)	–	9.14 (5.53)	0.15 (0.01)
rf	CE-OCL_tr	**1.00** (0.00)	–	-97.01 (95.21)	0.89 (0.01)	–	115.65 (113.90)	0.16 (0.01)
	DiCE	**1.00** (0.00)	–	-2166.72 (318.36)	0.23 (0.02)	–	**2446.71** (455.17)	**0.32** (0.01)
	CE-OCL	**1.00** (0.00)	–	**-3.79** (1.24)	**0.88** (0.01)	–	7.50 (2.49)	0.24 (0.01)
lr	CE-OCL_tr	**1.00** (0.00)	–	-614.00 (202.97)	0.83 (0.01)	–	1107.84 (381.92)	0.25 (0.01)
	DiCE	0.92 (0.05)	–	-1946.86 (256.37)	0.21 (0.02)	–	**1909.26** (187.85)	**0.29** (0.01)
	CE-OCL	**1.00** (0.00)	–	-1.85 (0.23)	**0.87** (0.00)	–	1.91 (0.23)	0.17 (0.00)
cart	CE-OCL_tr	**1.00** (0.00)	–	-212.82 (100.91)	0.85 (0.01)	–	285.60 (121.56)	0.22 (0.01)
	DiCE	0.00 (0.00)	–	-1895.95 (230.21)	0.25 (0.02)	–	**2214.51** (319.98)	**0.32** (0.01)
	CE-OCL	**1.00** (0.00)	–	**-24.21** (8.71)	**0.89** (0.00)	–	38.15 (13.75)	0.15 (0.01)
mlp	CE-OCL_tr	**1.00** (0.00)	–	-996.30 (370.04)	0.85 (0.01)	–	971.37 (447.54)	0.17 (0.01)
	DiCE	0.97 (0.03)	–	-2526.22 (265.46)	0.20 (0.02)	–	**3205.58** (427.42)	**0.32** (0.01)
	CE-OCL	**1.00** (0.00)	–	**-175.98** (74.32)	**0.89** (0.01)	–	296.26 (134.93)	0.17 (0.01)
gbm	CE-OCL_tr	**1.00** (0.00)	–	-219.19 (131.96)	0.87 (0.01)	–	123.61 (82.17)	0.16 (0.01)
	DiCE	0.93 (0.04)	–	-2222.50 (277.89)	0.22 (0.02)	–	**2749.12** (409.75)	**0.31** (0.02)

Appendix C. Case studies

We reserve this appendix for the details of our case study, Statlog (German Credit Data) dataset, and for the additional demonstration on the Statlog (Heart) dataset.

C.1. German Credit Data tables

We report an overview of the Statlog (German Credit Data) dataset features and the CEs generated at each step detailed in Section 3 in Table 4 and Table 5, respectively.

Table 4: Information on Statlog (German Credit Data) Data Set [4]

Label	Variable name	Description	Domain*	Constraint
F1	duration	Duration in months	real	≥ 0
F2	credit_amount	Credit amount	real	≥ 0
F3	instalment_commitment	Installment rate in percentage of disposable income	real	≥ 0
F4	age	Age in years	real	$x_{age} \geq \hat{\boldsymbol{x}}_{age}$
F5	residence_since	Present residence since X years	integer	$x_{residence\_since} \geq \hat{\boldsymbol{x}}_{residence\_since}$
F6	existing_credits	Number of existing credits at this bank	integer	≥ 0
F7	num_dependents	Number of people being liable to provide maintenance for	integer	≥ 0
F8	checking_status	Status of existing checking account, in Deutsche Mark	binary	–
F9	credit_history	Credit history (credits taken, paid back duly, delays, critical accounts)	binary	–
F10	employment	Present employment, in number of years.	binary	conditionally immutable
F11	foreign_worker	Foreign worker (yes,no)	binary	immutable
F12	housing	Housing (rent, own,...)	binary	–
F13	job	Job	binary	–
F14	other_parties	Other debtors / guarantors	binary	–
F15	other_payment_plans	Other installment plans (banks, stores)	binary	–
F16	own_telephone	Telephone (yes,no)	binary	–
F17	personal_status	Personal status (married, single,...) and sex	binary	immutable
F18	property_magnitude	Property (e.g. real estate)	binary	–
F19	purpose	Purpose of the credit (car, television,...)	binary	immutable
F20	saving_status	Status of savings account/bonds, in Deutsche Mark.	binary	–

* All categorical are one-hot encoded and therefore considered binary.

Table 5: `CE-OCL` demo on the Statlog (German Credit Data) Data Set [4].

(a) Counterfactual explanations generated for enriching the optimization model step by step with the constraint presented in Section 2

	F1	F2	F3	F4	F8*	F10	F12	F14	F16	F18*	F20*
$\hat{x}$	24.0	1371.26	4.0	25.0	A	1≤X<4	rent	none	none	A	A
Part A: validity, proximity, coherence											
(a)	15.02	-333.52	3.86	27.04	–	–	–	–	–	–	–
Part B: validity, proximity, coherence, sparsity											
(a)	7.12	–	–	–	–	–	–	–	–	–	–
Part C: validity, proximity, coherence, sparsity, diversity											
(a)	7.12	–	–	–	–	–	–	–	–	–	–
(b)	–	-2873.47	–	30.06	–	–	–	–	–	–	–
(c)	–	–	1.96	26.63	–	–	–	–	–	–	–
Part D: validity, proximity, coherence, sparsity, diversity, actionability											
(a)	7.12	–	–	–	–	–	–	–	–	–	–
(b)	–	–	1.96	26.63	–	–	–	–	–	–	–
(c)	–	–	–	75.52	–	–	–	–	–	–	–
Part E: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness											
(a)	22.0	1283.52	–	–	B	4≤X<7	–	–	–	B	–
(b)	10	1363.43	2.0	64.0	B	–	own	–	yes	C	B
(c)	12.0	1893.04	–	29.0	–	–	own	guarantor	yes	B	B
Part F: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness, causality											
(a)	–	–	–	–	B	4<=X<7	–	–	–	B	–
(b)	22.0	990.51	–	–	B	4<=X<7	–	–	–	B	–
(c)	26.83	1910.28	–	–	B	4<=X<7	–	–	–	B	–

F1–F20 represent the 20 features of the dataset. See Table 4 in Appendix C for a description.
The dash (–) represents no change in a feature with respect to the factual instance.
F5, F6, F7, F9, F11, F13, F15, F17, F19: None of the counterfactual explanations proposed a change in these variables. For space reasons they are not displayed here.
* **F8:** A: <0, B: no checking; **F18:** A: real estate, B: life insurance, C: car ; **F20:** A: no known savings, B: <100

(b) Evaluation* of counterfactuals generated for a single factual instance, with constraints added gradually.

	categorical proximity($\uparrow_0^1$)	**continuous proximity**($\uparrow_-^0$)	**sparsity**($\uparrow_0^1$)	**categorical diversity**($\uparrow_0^1$)	**continuous diversity**($\uparrow_0^+$)	**sparsity-based diversity**($\uparrow_0^1$)
Part A	1.00	-1715.94	0.8	–	–	–
Part B	1.00	-16.88	0.95	–	–	–
Part C	1.00	-1423.45	0.92	0.00	2845.81	0.15
Part D	1.00	-23.69	0.93	0.00	46.29	0.12
Part E	0.67	-230.12	0.63	0.36	441.68	0.42
Part F	0.77	-308.20	0.78	0.00	616.40	0.10

Part A: validity, proximity, coherence; **Part B:** validity, proximity, coherence, sparsity; **Part C:** validity, proximity, coherence, sparsity, diversity; **Part D:** validity, proximity, coherence, sparsity, diversity, actionability; **Part E:** validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness; **Part F:** validity, proximity, coherence, sparsity, diversity, actionability. data manifold closeness, causality
* **validity** ($\uparrow_0^1$)**:** 1.00 in all cases

C.2. German Credit Data CE generation model

For the case study in Section 3, we made use of the Statlog (German Credit Data) dataset [4] [6]. Table 4 provides an overview of features in this dataset, alongside a short description and the measurement level. For conciseness, we labelled the features F1-F20, and use those labels throughout the manuscript. Table 4 also displays the actionability constraints we imposed on the features. The following mathematical model is used to generate CEs and contains all the constraints – criteria – presented in Section 2:

$$
\begin{aligned}
\underset{\boldsymbol{x},\boldsymbol{z},\boldsymbol{s}\in\mathbb{R}^n,\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{I}|}_{\geq 0}}{\text{minimize}} \quad & \ell_2(\boldsymbol{x},\hat{\boldsymbol{x}}) + \alpha\sum_i z_i + \beta\ell_1(\boldsymbol{s},\tilde{\boldsymbol{s}}) && \text{(8a, proximity, sparsity, and closeness)}\\
\text{subject to} \quad & h(\boldsymbol{x}) = 1 && \text{(8b, validity)}\\
& |\boldsymbol{x}-\hat{\boldsymbol{x}}| \leq M\boldsymbol{z}, && \text{(8c, sparsity)}\\
& \sum_{i\in\mathcal{I}}\lambda_i\bar{\boldsymbol{x}}_i = \boldsymbol{x}+\boldsymbol{s}, && \text{(8d, data manifold closeness)}\\
& \sum_{i\in\mathcal{I}}\lambda_i = 1, && \text{(8e, data manifold closeness)}\\
& x_i \geq 0, \;\; i\in\{F1,F2,F3,F6,F7\} && \text{(8f, actionability)}\\
& x_i \geq \hat{x}_i, \;\; i\in\{F4,F5\} && \text{(8g, actionability)}\\
& x_i = \hat{x}_i, \;\; i\in\{F11,F17,F19\} && \text{(8h, immutability)}\\
& x_{F10}\in\mathcal{C}_{F10}, && \text{(8i, conditional immutability)}\\
& x_{F1} = \hat{x}_{F1} + h_{causality}(x_{F2}) - h_{causality}(\hat{x}_{F2}), && \text{(8j, causality)}\\
& \boldsymbol{x}\in\mathcal{L}, && \text{(8k, Domain (real, integer, binary))}
\end{aligned}
$$

6. Preprocessed from https://datahub.io/machine-learning/credit-g.

C.3. Heart tables

Similarly to the German Credit Data case study, we report Table 6 with the CEs generated at each step and the scores for the evaluation metrics, and Table 7 with an overview of the features.

Table 6: CE-OCL demo on the Statlog (Heart) Data Set [4]

(a) Counterfactual explanations generated for enriching the optimization model step by step with the constraint presented in Section 2

	age	bp	sch	mhrt	opk	chp	ecg	exian	fbs	sex	slope	thal	vessel
$\hat{x}$	49.0	130.0	265.98	171.01	0.6	atypical angina	normal	no	false	male	upsloping	normal	0
Part A: validity, proximity, coherence													
(a)	48.82	139.28	328.09	153.98	1.05	-	-	-	-	-	-	-	-
Part B: validity, proximity, coherence, sparsity													
(a)	-	-	407.24	-	-	-	-	-	-	-	-	-	-
Part C: validity, proximity, coherence, sparsity, diversity													
(a)	-	-	407.24	-	-	-	-	-	-	-	-	-	-
(b)	-	-	393.92	175.14	-	-	-	-	-	-	-	-	-
(c)	-	-	404.04	-	0.47	-	-	-	-	-	-	-	-
Part D: validity, proximity, coherence, sparsity, diversity, actionability													
(a)	-	-	407.24	-	-	-	-	-	-	-	-	-	-
(b)	-	-	393.92	175.14	-	-	-	-	-	-	-	-	-
(c)	-	-	-	124.37	-	-	-	-	-	-	-	-	-
Part E: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness													
(a)	-	111.77	253.81	152.7	0.0	nonanginal pain	-	-	-	-	-	-	-
(b)	-	137.0	258.5	147.01	1.55	asymptomatic	left ventricular hypertrophy	-	-	-	flat	reversible defect	-
(c)	-	140.61	274.7	128.61	0.49	asymptomatic	left ventricular hypertrophy	yes	-	-	-	reversible defect	-

See Table 7 for a description of each feature
The dash (–) represents no change in a feature with respect to the factual instance.

(b) Evaluation* of counterfactuals generated for a single factual instance, with constraints added gradually.

	categorical proximity($\uparrow_0^1$)	**continuous proximity**($\uparrow_-^0$)	**sparsity**($\uparrow_0^1$)	**categorical diversity**($\uparrow_0^1$)	**continuous diversity**($\uparrow_0^+$)	**sparsity-based diversity**($\uparrow_0^1$)
Part A	1.00	-89.05	0.62	-	-	-
Part B	1.00	-141.26	0.92	-	-	-
Part C	1.00	-137.17	0.87	0.00	11.72	0.18
Part D	1.00	-106.66	0.90	0.00	128.02	0.15
Part E	0.62	-50.19	0.46	0.42	50.25	0.56

Part A: validity, proximity, coherence; **Part B:** validity, proximity, coherence, sparsity; **Part C:** validity, proximity, coherence, sparsity, diversity; **Part D:** validity, proximity, coherence, sparsity, diversity, actionability; **Part E:** validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness; **Part F:** validity, proximity, coherence, sparsity, diversity, actionability. data manifold closeness, causality
*__validity__ ($\uparrow_0^1$)**:** 1.00 in all cases

Table 7: Information on Statlog (Heart) Data Set [4]

Variable name	Description	Domain*	Constraint
age	Patient age in years	real	immutable
sex	Gender	binary	immutable
chp	Chest pain type	binary	-
bp	Resting blood pressure	real	≥ 0
sch	Serum cholesterol	real	≥ 0
fbs	Fasting blood sugar >120 mg/dL	binary	-
ecg	Resting electrocardiographic result	binary	-
mhrt	Maximum heart rate	real	≥ 0
exian	Exercise induced angina	binary	-
opk	Old peak	real	≥ 0
slope	Slope of peak exercise ST segment	binary	-
vessel	Number of major vessels	binary	-
thal	Defect type	binary	-

* All categorical are one-hot encoded and therefore considered binary.

C.4. Heart CE generation model

The following mathematical model is used to generate CEs and contains all the constraints – criteria – presented in the Section 2:

$$
\begin{aligned}
\underset{\boldsymbol{x},\boldsymbol{z},\boldsymbol{s}\in\mathbb{R}^n,\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{I}|}_{\geq 0}}{\text{minimize}} \quad & \ell_2(\boldsymbol{x},\hat{\boldsymbol{x}}) + \alpha \sum_i z_i + \beta \ell_1(\boldsymbol{s},\tilde{\boldsymbol{s}}) && \text{(9a, proximity, sparsity, and closeness)} \\
\text{subject to} \quad & h(\boldsymbol{x}) = 1 && \text{(9b, validity)} \\
& |\boldsymbol{x} - \hat{\boldsymbol{x}}| \leq M\boldsymbol{z}, && \text{(9c, sparsity)} \\
& \sum_{i\in\mathcal{I}} \lambda_i \bar{\boldsymbol{x}}_i = \boldsymbol{x} + \boldsymbol{s}, && \text{(9d, data manifold closeness)} \\
& \sum_{i\in\mathcal{I}} \lambda_i = 1, && \text{(9e, data manifold closeness)} \\
& x_i \geq 0, \;\; i \in \{bp, sch, mhrt, opk\} && \text{(9f, actionability)} \\
& x_i = \hat{x}_i, \;\; i \in \{age, sex\} && \text{(9g, immutability)} \\
& \boldsymbol{x} \in \mathcal{L}, && \text{(9h, Domain (real, binary))}
\end{aligned}
$$

The predictive model used for this demo is a neural network with one hidden layer of 50 nodes and ReLU activation functions. A description of the Statlog (Heart) dataset used in the experiment is given in Table 7. The experiments have the same structure described in Section 3, and the results are reported in Table 6.