

On the computation of counterfactual explanations - A survey

André Artelt* and Barbara Hammer †

CITEC - Cognitive Interaction Technology
Bielefeld University - Faculty of Technology
Inspiration 1, 33619 Bielefeld - Germany

Abstract. Due to the increasing use of machine learning in practice it becomes more and more important to be able to explain the prediction and behavior of machine learning models. An instance of explanations are counterfactual explanations which provide an intuitive and useful explanations of machine learning models.

In this survey we review model-specific methods for efficiently computing counterfactual explanations of many different machine learning models and propose methods for models that have not been considered in literature so far.

1 Introduction

Due to recent advances in machine learning (ML), ML methods are increasingly use in real world scenarios [1–4]. Especially, ML technology is nowadays used in critical situations like predictive policing [5] and loan approval [6]. In order to increase trust and acceptance of these kind of technology, it is important to be able to explain the behaviour and prediction of these models [7] - in particular answer questions like “Why did the model do that? And why not smth. else?”. This becomes even more important in view to legal regulations like the EU regulation on GDPR [8], that grants the user a right to an explanation.

A popular method for explaining models [7, 9–11] are counterfactual explanations (often just called counterfactuals) [12]. A counterfactual explanation states changes to some features that lead to a different (specified) behaviour or prediction of the model. Thus, counterfactual explanation can be interpreted as a recommendation what to do in order to achieve a requested goal. This is why counterfactual explanations are that popular - they are intuitive and user-friendly [7, 12].

Counterfactual explanations are an instance of model-agnostic methods. Therefore, counterfactuals are not tailored to a particular model but can be computed for all possible models (in theory). Other instances of model-agnostic methods are feature interaction methods [13], feature importance methods [14], partial dependency plots [15] and local methods that approximates the model locally by an explainable model (e.g. a decisiontree) [16, 17]. The nice thing about model-agnostic methods is that they (in theory) do not need access to model

*corresponding author: aartelt@techfak.uni-bielefeld.de

†We gratefully acknowledge funding from the VW-Foundation for project *IMPACT* funded in the frame of the funding line *AI and its Implications for Future Society*.

internals and/or training data - it is sufficient to have an interface where we can pass data points to the model and observe the output/predictions of the model.

However, it turns out that efficiently computing high quality counterfactual explanations of black-box models can be very difficult [18]. Therefore, it is beneficial to develop model-specific methods - that use model internals - for efficiently computing counterfactual explanations. Whenever we have access to model internals, we can use the model-specific method over the model-agnostic method for efficiently computing counterfactual explanations. In this work we focus on such model-specific methods.

In particular, our contributions are:

- We review model-specific methods for efficiently computing counterfactual explanations of different ML models.
- We propose model-specific methods for efficiently computing counterfactual explanations of models that have not been considered in literature so far.

The remainder of this paper is structured as follows: First, we briefly review counterfactual explanations (section 2). Then, in section 3 we review and propose model-specific methods for computing counterfactual explanations. Finally, section 5 summarizes this papers. All derivations and mathematical details can be found in the appendix (section 6).

2 Counterfactual explanations

Counterfactual explanations [12] (often just called counterfactuals) are an instance of example-based explanations [19]. Other instances of example-based explanations [7] are influential instances [20] and prototypes & criticisms [21].

A counterfactual states a change to some features/dimensions of a given input such that the resulting data point (called counterfactual) has a different (specified) prediction than the original input. Using a counterfactual instance for explaining the prediction of the original input is considered to be fairly intuitive, human-friendly and useful because it tells people what to do in order to achieve a desired outcome [7, 12].

A classical use case of counterfactual explanations is loan application [6, 7]: *Imagine you applied for a credit at a bank. Unfortunately, the bank rejects your application. Now, you would like to know why. In particular, you would like to know what would have to be different so that your application would have been accepted. A possible explanation might be that you would have been accepted if you would earn 500\$ more per month and if you would not have a second credit card.*

Although counterfactuals constitute very intuitive explanation mechanisms, there do exist a couple of problems.

One problem is that there often exist more than one counterfactual - this is called *Rashomon effect* [7]. If there are more than one possible explanation (counterfactual), it is not clear which one should be selected.

An alternative - but very similar in the spirit - to counterfactuals [12] is the Growing Spheres method [22]. However, this method suffers from the curse of dimensionality because it has to draw samples from the input space, which can become difficult if the input space is high-dimensional.

According to [12], we formally define the finding of a counterfactual as follows: Assume a prediction function $h : \mathcal{X} \mapsto \mathcal{Y}$ is given. Computing a counterfactual $\vec{x}' \in \mathbb{R}^d$ of a given input $\vec{x} \in \mathbb{R}^d$ can be interpreted as an optimization problem:

$$\arg \min_{\vec{x}' \in \mathbb{R}^d} \ell(h(\vec{x}'), y') + C \cdot \theta(\vec{x}', \vec{x}) \quad (1)$$

where $\ell()$ denotes a loss function that penalizes deviation of the prediction $h(\vec{x}')$ from the requested prediction y' . $\theta()$ denotes a regularization that penalizes deviations from the original input \vec{x} and the hyperparameter C denotes the regularization strength.

Two common regularizations are the weighted Manhattan distance and the generalized L2 distance. The weighted Manhattan distance is defined as

$$\theta(\vec{x}', \vec{x}) = \sum_j \alpha_j \cdot |(\vec{x})_j - (\vec{x}')_j| \quad (2)$$

where $\alpha_j > 0$ denote the feature wise weights. A popular choice [12] for α_j is the inverse median absolute deviation of the j -th feature median in the training data set \mathcal{D} :

$$\begin{aligned} \alpha_j &= \frac{1}{\text{MAD}_j} \\ \text{where} \quad & \quad \quad \quad (3) \\ \text{MAD}_j &= \text{median}_{\vec{x} \in \mathcal{D}} \left(\left| (\vec{x})_j - \text{median}_{\vec{x} \in \mathcal{D}} ((\vec{x})_j) \right| \right) \end{aligned}$$

The weights α_j compensate for the (potentially) different variability of the features. However, because we need access to the training data set \mathcal{D} , this regularization is not a truly model-agnostic method - it is not usable if we only have access to a prediction interface of a black-box model.

Although counterfactual explanations are a model-agnostic method, the computation of a counterfactual becomes much more efficient when having access to the internals of the model. In this work we assume that we have access to all needed model internals as well as access to the training data set - we will only need the training data for computing the weights α_j in the weighted Manhattan distance Eq. 2. We do not need access to the training data if we do not use the weighted Manhattan distance or if we use some other methods for computing the weights α_j (e.g. setting all weights to 1).

¹We restrict ourself to \mathbb{R}^d , but in theory one could use an arbitrary domain \mathcal{X} .

A slightly modified version of Eq. 1 was proposed in [23]. The authors claim that the original formalization in Eq. 1 does not take into account that the counterfactual should lie on the data manifold - the counterfactual should be a plausible data instance. To deal with this issue, the authors propose to add two additional terms to the original objective Eq. 1:

1. The distance/norm between the counterfactual \bar{x}' and the reconstructed version of it that has been computed by using a pretrained autoencoder.
2. The distance/norm between the encoding of the counterfactual \bar{x}' and the mean encoding of training samples that belong to the requested class y' .

The first term is supposed to make sure that the counterfactual \bar{x}' lies on the data manifold and thus is a plausible data instance. The second term is supposed to accelerate the solver for computing the solution of the final optimization problem. Both claims have been evaluated empirically [23].

Recently, another approach for computing plausible/feasible counterfactual explanations was proposed [24]. Instead of computing a single counterfactual, the authors propose to compute a path of intermediate counterfactuals that lead to the final counterfactual. The idea behind this path of intermediate counterfactuals is to provide the user with a set of intermediate goals that finally lead to the desired goal - it might be more feasible to “go into the direction” of the final goal step by step instead of accomplishing it in a single step. In order to compute such a path of intermediate counterfactuals, the authors propose different strategies for constructing a graph on the training data set - including the query point. In this graph, two samples are connected by a weighted edge if they are “sufficient close to each other” - the authors propose different measurements for closeness (e.g. based on density estimation). The path of intermediate counterfactuals is equal to the shortest path between the query point and a point that satisfies the desired goal - this is the final counterfactual. Therefore the final counterfactual as well as all intermediate counterfactuals are elements from the training data set.

Despite the highlighted issues [23, 24] of the original formalization Eq. 1, we stick to it and leave further investigations on the computation of feasible & plausible counterfactuals as future research. However, many of the approaches for computing counterfactuals - that are discussed in this paper - can be augmented to restrict the space of potential counterfactuals. These restrictions provide an opportunity for encoding domain knowledge that lead to more plausible and feasible counterfactuals.

3 Computation of counterfactuals

In the subsequent sections we explore model-specific methods for efficiently computing counterfactual explanations of many different ML models. But before looking at model-specific methods, we first (section 3.1) discuss methods for dealing with arbitrary types of models - gradient based as well as gradient free methods.

Note that for the purpose of better readability and due to space constraints, we put all derivations in the appendix (section 6).

3.1 The general case

We can compute a counterfactual explanation of any model we like by plugging the prediction function h of the model into Eq. 1 and choosing a loss (eq. 0-1 loss) and regularization (e.g. Manhattan distance) function. Depending on the model, loss and regularization function, the resulting optimization problem might be differentiable or not. If it is differentiable, we can use gradient-based methods like (L-)BFGS and conjugate gradients for solving the optimization problem. If Eq. 1 is not differentiable, we can use gradient-free methods like the Downhill-Simplex method or an evolutionary algorithm like CMA-ES or CERTIFAI [25] - the nice thing about evolutionary algorithms is that they can easily deal with categorical features. Another approach, limited to linear classifiers, for handling continuous and discrete features is to use mixed-integer programming (MIP) [26]. Unfortunately, solving a MIP is NP-hard. However, there exist solvers that can compute an approximate solution very efficiently. Popular methods are branch-and-bound and branch-and-cut algorithms [27].

When developing model-specific methods for computing counterfactuals, we always consider untransformed inputs only - since a non-linear feature transformation usually makes the problem non-convex. Furthermore, we only consider the Euclidean distance and the weighted Manhattan distance as candidates for the regularization function $\theta(\cdot)$.

3.2 Separating hyperplane models

A model whose prediction function h can be written as:

$$h(\vec{x}) = \text{sign}(\vec{w}^\top \vec{x} + b) \quad (4)$$

is called a separating hyperplane model. Popular instances of separating hyperplane models are SVM, LDA, perceptron and logistic regression.

Without loss of generality, we assume $\mathcal{Y} = \{-1, 1\}$. Then, the optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned} \arg \min_{\vec{x}' \in \mathbb{R}^d} \quad & \theta(\vec{x}', \vec{x}) \\ \text{s.t.} \quad & \\ & \vec{q}^\top \vec{x}' + c < 0 \end{aligned} \quad (5)$$

where

$$\vec{q} = -y' \vec{w} \quad (6)$$

$$c = -by' \quad (7)$$

Depending on the regularization, the optimization problem Eq. 5 becomes either a linear program (LP) - if the weighted Manhattan distance is used - or a convex

quadratic program (QP) with linear constraints - if the Euclidean distance is used. More details can be found in the appendix (section 6.2).

If we would have some discrete features instead of continuous features only, we would obtain a MIP or MIQP as described in [26].

3.3 Generalized linear model

In a generalized linear model we assume that the distribution of the response variable belongs to the exponential family. The expected value is connected to a linear combination of features by a link function, where different distributions have different link functions.

In the subsequent sections, we explore how to efficiently compute counterfactual explanations of popular instances of the generalized model.

3.3.1 Logistic regression

In logistic regression we model the response variable as a Bernoulli distribution. The prediction function h of a logistic regression model is given as

$$h(\vec{x}) = \begin{cases} 1 & \text{if } p(y = 1 \mid \vec{x}) \geq t \\ -1 & \text{otherwise} \end{cases} \quad (8)$$

where t is the discrimination threshold (often $t = 0.5$) and

$$p(y = 1 \mid \vec{x}) = \frac{1}{1 + \exp(-\vec{w}^\top \vec{x} - b)} \quad (9)$$

When ignoring all probabilities and setting $t = 0.5$, the prediction function h of a logistic regression model becomes a separating hyperplane:

$$h(\vec{x}) = \text{sign}(\vec{w}^\top \vec{x} + b) \quad (10)$$

Therefore, computing a counterfactual of a logistic regression model is exactly the same as for a separating hyperplane model (section 3.2).

3.3.2 Softmax regression

In softmax regression we model the distribution of the response variable as a generalized Bernoulli distribution. The prediction function h of a softmax regression model is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \frac{\exp(\vec{w}_i^\top \vec{x} + b_i)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} \quad (11)$$

In this case, the optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \text{s.t.} \\ & \vec{q}_{ij}^\top \vec{x}' + c_{ij} < 0 \quad \forall j \in \mathcal{Y}, j \neq i = y' \end{aligned} \quad (12)$$

where

$$\vec{q}_{ij} = \vec{w}_j - \vec{w}_i \quad (13)$$

$$c_{ij} = b_j - b_i \quad (14)$$

Depending on the regularization, the optimization problem Eq. 12 becomes either a LP - if the weighted Manhattan distance is used - or a convex QP with linear constraints - if the Euclidean distance is used. More information can be found in the appendix (section 6.3.1).

3.3.3 Linear regression

In linear regression we model the distribution of the response variable as a Gaussian distribution. The prediction function f of a linear regression model is given as:

$$f(\vec{x}) = \vec{w}^\top \vec{x} + b \quad (15)$$

The optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \text{s.t.} \\ & \vec{w}^\top \vec{x}' + c \leq \epsilon \\ & -\vec{w}^\top \vec{x}' - c \leq \epsilon \end{aligned} \quad (16)$$

where

$$c = b - y' \quad (17)$$

and $\epsilon \geq 0$ denotes the tolerated deviation from the requested prediction y' .

Depending on the regularization, the optimization problem Eq. 16 becomes either a LP (if the weighted Manhattan distance is used) or a convex QP with linear constraints (if the Euclidean distance is used). More information can be found in the appendix (section 6.3.2).

3.3.4 Poisson regression

In Poisson regression we model the distribution of the response variable as a Poisson distribution. The prediction function f of a Poisson regression model is given as:

$$f(\vec{x}) = \exp(\vec{w}^\top \vec{x} + b) \quad (18)$$

In this case, the optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \text{s.t.} \\ & \vec{w}^\top \vec{x}' + c \leq \epsilon \\ & -\vec{w}^\top \vec{x}' - c \leq \epsilon \end{aligned} \quad (19)$$

where

$$c = b - \log(y') \quad (20)$$

and $\epsilon \geq 0$ denotes the tolerated deviation from the requested prediction y' .

Depending on the regularization, the optimization problem Eq. 19 becomes either a LP (if the weighted Manhattan distance is used) or a convex QP with linear constraints (if the Euclidean distance is used). More information can be found in the appendix (section 6.3.3).

3.3.5 Exponential regression

In exponential regression we model the distribution of the response variable as a exponential distribution. The prediction function f of an exponential regression model is given as:

$$f(\vec{x}) = -\frac{1}{\vec{w}^\top \vec{x} + b} \quad (21)$$

Then, the optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \text{s.t.} \\ & \vec{w}^\top \vec{x}' + c \leq \epsilon \\ & -\vec{w}^\top \vec{x}' - c \leq \epsilon \end{aligned} \quad (22)$$

where

$$c = b + \frac{1}{y'} \quad (23)$$

and $\epsilon \geq 0$ denotes the tolerated deviation from the requested prediction y' .

Depending on the regularization, the optimization problem Eq. 22 becomes either a LP (if the weighted Manhattan distance is used) or a convex QP with linear constraints (if the Euclidean distance is used). More information can be found in the appendix (section 6.3.4).

3.4 Gaussian naive Bayes

The Gaussian naive Bayes model makes the assumption that all features are independent of each other and follow a normal distribution. The prediction function h of a Gaussian naive Bayes model is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \prod_{k=1}^d \mathcal{N}(\vec{x} \mid \mu_{ik}, \sigma_{ik}^2) \pi_i \quad (24)$$

where π_i denotes the a-priori probability of the i -th class.

The optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned}
& \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\
& \text{s.t.} \\
& \vec{x}'^\top \mathbf{A}_{ij} \vec{x}' + \vec{q}_{ij}^\top \vec{x}' + c_{ij} < 0 \quad \forall j \in \mathcal{Y}, j \neq i = y'
\end{aligned} \tag{25}$$

where

$$\mathbf{A}_{ij} = \text{diag} \left(\frac{1}{2\sigma_{ik}^2} - \frac{1}{2\sigma_{jk}^2} \right) \tag{26}$$

$$\vec{q}_{ij} = \left(\frac{\mu_{j1}}{\sigma_{j1}^2} - \frac{\mu_{i1}}{\sigma_{i1}^2}, \dots, \frac{\mu_{jd}}{\sigma_{jd}^2} - \frac{\mu_{id}}{\sigma_{id}^2} \right)^\top \tag{27}$$

$$c_{ij} = \log \left(\frac{\pi_j}{\pi_i} \right) + \sum_{k=1}^d \log \left(\frac{\sqrt{2\pi\sigma_{ik}^2}}{\sqrt{2\pi\sigma_{jk}^2}} \right) - \frac{\mu_{jk}^2}{2\sigma_{jk}^2} + \frac{\mu_{ik}^2}{2\sigma_{ik}^2} \tag{28}$$

Because we can not make any statement about the definiteness of \mathbf{A}_{ij} , the quadratic constraints in Eq. 25 are non-convex. Therefore, the optimization problem Eq. 25 is a non-convex quadratically constrained quadratic program (QCQP).

We can approximately solve Eq. 25 by using an approximation method like the Suggest-Improve framework [28]. Furthermore, if we have a binary classification problem, we can solve a semi-definite program (SDP) whose solution is equivalent to Eq. 25. More details can be found in the appendix (sections 6.4, 6.9.1 and 6.9.2).

3.5 Quadratic discriminant analysis

In quadratic discriminant analysis (QDA) we model each class distribution as an independent Gaussian distribution - note that in contrast to LDA each class distribution has its own covariance matrix. The prediction function h of a QDA model is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \mathbf{\Sigma}_i) \pi_i \tag{29}$$

where π_i denotes the a-priori probability of the i -th class.

In this case, the optimization problem for computing a counterfactual explanation Eq. 1 can be rewritten as:

$$\begin{aligned}
& \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\
& \text{s.t.} \\
& \frac{1}{2} \vec{x}'^\top \mathbf{A}_{ij} \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + c_{ij} < 0 \quad \forall j \in \mathcal{Y}, j \neq i = y'
\end{aligned} \tag{30}$$

where

$$\mathbf{A}_{ij} = \Sigma_i^{-1} - \Sigma_j^{-1} \quad (31)$$

$$\vec{q}_{ij} = \Sigma_j^{-1} \vec{\mu}_j - \Sigma_i^{-1} \vec{\mu}_i \quad (32)$$

$$c_{ij} = \frac{1}{2} \left(\vec{\mu}_i^\top \Sigma_i^{-1} \vec{\mu}_i - \vec{\mu}_j^\top \Sigma_j^{-1} \vec{\mu}_j \right) + \frac{1}{2} \log \left(\frac{\det(\Sigma_i)}{\det(\Sigma_j)} \right) + \log \left(\frac{\pi_j}{\pi_i} \right) \quad (33)$$

Because we can not make any statement about the definiteness of \mathbf{A}_{ij} , the quadratic constraints in Eq. 30 are non-convex. Thus, like in Gaussian naive Bayes (section 3.4), the optimization problem Eq. 30 is a non-convex QCQP.

Like in the case of the previous non-convex QCQPs, we can approximately solve Eq. 30 by using an approximation method. Furthermore, if we have a binary classification problem, we can solve a SDP whose solution is equivalent to Eq. 30. More details can be found in the appendix (sections 6.5, 6.9.1 and 6.9.2).

3.6 Learning vector quantization models

Learning vector quantization (LVQ) models [29] compute a set of labeled prototypes $\{(\vec{p}_i, o_i)\}$ from a given training data set - we refer to the i -th prototype as \vec{p}_i and the corresponding label as o_i . The prediction function h of a LVQ model is given as:

$$\begin{aligned} h(\vec{x}) &= o_i \\ \text{s.t. } \min d(\vec{x}, \vec{p}_i) \end{aligned} \quad (34)$$

where $d()$ denotes a function for computing the distance between a data point and a prototype - usually this is the Euclidean distance:

$$d(\vec{x}, \vec{p}) = (\vec{x} - \vec{p})^\top \mathbb{I} (\vec{x} - \vec{p}) \quad (35)$$

There exist LVQ models like (L)GMLVQ [30] and (L)MRSLVQ [31] that learn a custom (class or prototype specific) distance matrix $\mathbf{\Omega}_p$ that is used instead of the identity \mathbb{I} when computing the distance between a data point and a prototype. This gives rise to the generalized L2 distance:

$$d(\vec{x}, \vec{p}) = (\vec{x} - \vec{p})^\top \mathbf{\Omega}_p (\vec{x} - \vec{p}) \quad (36)$$

Because a LVQ model assigns the label of the nearest prototype to a given input, the nearest prototype of a counterfactual must be a prototype \vec{p}_i with $o_i = y'$. According to [18], for computing a counterfactual, it is sufficient to solve the following optimization problem for each prototype \vec{p}_i with $o_i = y'$ and select the counterfactual \vec{x}' yielding the smallest value of $\theta(\vec{x}', \vec{x})$:

$$\begin{aligned} \arg \min_{\vec{x}' \in \mathbb{R}^d} & \theta(\vec{x}', \vec{x}) \\ \text{s.t.} & \\ d(\vec{x}', \vec{p}_i) & < d(\vec{x}', \vec{p}_j) \quad \forall \vec{p}_j \in \mathcal{P}(y') \end{aligned} \quad (37)$$

where $\mathcal{P}(y')$ denotes the set of all prototypes not labeled as y' . Note that the feasible region of Eq. 37 is always non-empty - the prototype \vec{p}_i is always a feasible solution.

In the subsequent sections we explore the type of constraints of Eq. 37 for different LVQ models.

3.6.1 (Generalized matrix) LVQ

In case of a (generalized matrix) LVQ model - all prototypes use the same distance matrix $\mathbf{\Omega}$, the optimization problem Eq. 37 becomes [18]:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \text{s.t.} \\ & \vec{x}'^\top \vec{q}_{ij} + c_{ij} < 0 \quad \forall \vec{p}_j \in \mathcal{P}(y') \end{aligned} \quad (38)$$

where

$$\vec{q}_{ij} = \frac{1}{2} \mathbf{\Omega} (\vec{p}_j - \vec{p}_i) \quad (39)$$

$$c_{ij} = \frac{1}{2} \left(\vec{p}_i^\top \mathbf{\Omega} \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega} \vec{p}_j \right) \quad (40)$$

Depending on the regularization, the optimization problem Eq. 38 becomes either a LP (if the Euclidean distance is used) or a convex QP with linear constraints (if the weighted Manhattan distance is used). More information can be found in the appendix (section 6.6).

3.6.2 (Localized generalized matrix) LVQ

In case of a (localized generalized matrix) LVQ model - there are different, class or prototype specific, distance matrices $\mathbf{\Omega}_p$, the optimization problem Eq. 37 becomes [18]:

$$\begin{aligned} & \arg \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\ & \frac{1}{2} \vec{x}'^\top \mathbf{A}_{ij} \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + c_{ij} < 0 \quad \forall \vec{p}_j \in \mathcal{P}(y') \end{aligned} \quad (41)$$

where

$$\mathbf{A}_{ij} = \mathbf{\Omega}_i - \mathbf{\Omega}_j \quad (42)$$

$$\vec{q}_{ij} = \frac{1}{2} (\mathbf{\Omega}_j \vec{p}_j - \mathbf{\Omega}_i \vec{p}_i) \quad (43)$$

$$c_{ij} = \frac{1}{2} \left(\vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j \right) \quad (44)$$

Because we can not make any statement about the definiteness of \mathbf{A}_{ij} , the quadratic constraints in Eq. 41 are non-convex. Thus, like in Gaussian naive

Bayes (section 3.4) and QDA (section 3.5), the optimization problem Eq. 41 is a non-convex QCQP.

Like the previous non-convex QCQPs, we can approximately solve Eq. 41 by using an approximation method. Furthermore, if we have a binary classification problem and each class is represented by a single prototype, we can solve a SDP whose solution is equivalent to Eq. 41. More details can be found in the appendix (sections 6.6, 6.9.1 and 6.9.2).

3.7 Tree based models

Tree based models are very popular in data science because they often achieve a high predictive-accuracy [32]. In the subsequent sections we discuss how to compute counterfactual explanations of tree based models. In particular, we consider decision/regression trees and tree based ensembles like random forest models.

3.7.1 Decision trees

In case of decision/regression tree models, we can compute a counterfactual by enumerating all possible paths that lead to the requested prediction [17, 33]. However, it might happen that some requested predictions are not possible because all possible predictions of the tree are encoded in the leafs. In this case one might define an interval of acceptable predictions so that a counterfactual exists.

The procedure for computing a counterfactual of a decision/regression tree is described in Algorithm 1.

Algorithm 1 Computing a counterfactual of a decision/regression tree

Input: Original input \vec{x} , requested prediction y' of the counterfactual, the tree model

Output: Counterfactual \vec{x}'

- 1: Enumerate all leafs with prediction y'
 - 2: For each leaf, enumerate all paths reaching the leaf
 - 3: For each path, compute the minimal change to \vec{x} that yields the path
 - 4: Sort all paths according to regularization of the change to \vec{x}
 - 5: Select the path and the corresponding change to \vec{x} that minimizes the regularization
-

3.7.2 Tree based ensembles

Popular instances of tree based ensembles are random forest and gradient boosting regression trees. It turns out that the problem of computing a counterfactual explanation of such models is NP-hard [33].

The following heuristic for computing a counterfactual explanation of a random forest model was proposed in [34]: First, we compute a counterfactual of a

model from the ensemble. Next, we use this counterfactual as a starting point for minimizing the number of trees that do not output the requested prediction by using a gradient-free optimization method like the Downhill-Simplex method. The idea behind this approach is that the counterfactual of a tree from the ensemble is close to the decision boundary of the ensemble so that computing a counterfactual of the ensemble becomes easier. By doing this for all trees in the ensemble, we get many counterfactuals and we can select the one that minimizes the regularization the most. This heuristic seems to work well in practice [34].

Another approach for computing counterfactual explanations of an ensemble of trees was proposed in [33] - although the authors do not call it counterfactuals, they actually compute counterfactuals. Their algorithm works as follows: We iterate over all trees in the ensemble that do not yield the requested prediction. Next, we compute all possible counterfactuals of each of these trees (see section 3.7.1). If this counterfactual turns out to be counterfactual of the ensemble, we store it so that in the end we can select the counterfactual with the smallest deviation from the original input. However, it can not be guaranteed that a counterfactual of the ensemble is found because it might happen that by changing the data point so that it becomes a counterfactual of a particular tree, the prediction of other trees in the ensemble change as well. According to the authors, this algorithm/heuristic works well in practice. Unfortunately, the worst-case complexity is exponential in the number of features and thus it is not suitable for high dimensional data.

4 Implementation

The gradient-based and gradient free methods, as well as the model specific methods for tree based models are already implemented in CEML [34]. The implementation of the LVQ specific methods are provided by the authors of [18]. The Python [35] implementation of our proposed methods is available on GitHub² and is based on the Python packages scikit-learn [36], numpy [37] and cvxpy [38].

We plan to add these model-specific methods to CEML [34] in the near future.

5 Conclusion

In this survey we extensively studied how to compute counterfactual explanations of many different ML models. We reviewed known methods from literature and proposed methods (mostly LPs and (QC)QPs) for computing counterfactuals of ML models that have not been considered in literature so far.

²<https://github.com/andreArtelt/OnTheComputationOfCounterfactualExplanations>

6 Appendix

6.1 Relaxing strict inequalities

When modeling the problem of computing counterfactuals, we often obtain strict inequalities like

$$g(\vec{x}) < 0 \quad (45)$$

Strict inequalities are not allowed in convex programming because the feasible region would become an open set. However, we could turn the $<$ into a \leq by adding a small number to the left side of the inequality:

$$g(\vec{x}) + \epsilon \leq 0 \quad (46)$$

where $\epsilon > 0$ is a small number.

In practice, when implementing our methods, we found that we can often safely replace all $<$ by \leq without changing anything else - this might be because of the numerics (like round-off errors) of fixed size floating-point numbers.

6.2 Separating hyperplane

Recall that the prediction function h is given as:

$$h(\vec{x}) = \text{sign}(\vec{w}^\top \vec{x} + b) \quad (47)$$

If we multiply the projection $\vec{w}^\top \vec{x} + b$ by the requested prediction y^3 , the result is positive if and only if the classification $h(\vec{x})$ is equal to y . Therefore, the linear constraint for predicting class y is given as

$$\begin{aligned} y (\vec{w}^\top \vec{x} + b) &> 0 \\ \Leftrightarrow \vec{q}^\top \vec{x} + c &< 0 \end{aligned} \quad (48)$$

where

$$\vec{q} = -y\vec{w} \quad (49)$$

$$c = -by \quad (50)$$

6.3 Generalized linear models

6.3.1 Softmax regression

Recall that the prediction function h is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \frac{\exp(\vec{w}_i^\top \vec{x} + b_i)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} \quad (51)$$

Thus, the constraint for obtaining a specific prediction y' is given as:

$$\frac{\exp(\vec{w}_i^\top \vec{x} + b_i)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} > \frac{\exp(\vec{w}_j^\top \vec{x} + b_j)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} \quad \forall j \neq i = y' \quad (52)$$

³Note that we assume $\mathcal{Y} = \{-1, 1\}$.

Holding i and j fixed, we can simplify Eq. 52:

$$\begin{aligned}
\frac{\exp(\vec{w}_i^\top \vec{x} + b_i)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} &> \frac{\exp(\vec{w}_j^\top \vec{x} + b_j)}{\sum_k \exp(\vec{w}_k^\top \vec{x} + b_k)} \\
&\Leftrightarrow \\
\exp(\vec{w}_i^\top \vec{x} + b_i) &> \exp(\vec{w}_j^\top \vec{x} + b_j) \\
&\Leftrightarrow \\
\vec{w}_i^\top \vec{x} + b_i &> \vec{w}_j^\top \vec{x} + b_j \\
&\Leftrightarrow \\
\vec{q}_{ij}^\top \vec{x}' + c_{ij} &< 0
\end{aligned} \tag{53}$$

where

$$\vec{q}_{ij} = \vec{w}_j - \vec{w}_i \tag{54}$$

$$c_{ij} = b_j - b_i \tag{55}$$

Therefore, we can rewrite Eq. 52 as a set of linear inequalities.

6.3.2 Linear regression

Recall that the prediction function f is given as:

$$f(\vec{x}) = \vec{w}^\top \vec{x} + b \tag{56}$$

By introducing the parameter $\epsilon \geq 0$ that specifies the maximum tolerated deviation from the requested prediction - we set $\epsilon = 0$ if we do not allow any deviations - the constraint for obtaining the requested prediction y' is given as

$$\begin{aligned}
|f(\vec{x}') - y'| &\leq \epsilon \\
\Leftrightarrow |\vec{w}^\top \vec{x}' + b - y'| &\leq \epsilon \\
\Leftrightarrow |\vec{w}^\top \vec{x}' + c| &\leq \epsilon
\end{aligned} \tag{57}$$

where

$$c = b - y' \tag{58}$$

Finally, we can rewrite Eq. 57 as two linear inequality constraints:

$$\begin{aligned}
\vec{w}^\top \vec{x}' + c &\leq \epsilon \\
-\vec{w}^\top \vec{x}' - c &\leq \epsilon
\end{aligned} \tag{59}$$

6.3.3 Poisson regression

Recall that the prediction function f is given as

$$f(\vec{x}) = \exp(\vec{w}^\top \vec{x} + b) \tag{60}$$

The constraint for exactly obtaining the requested prediction y' is

$$\begin{aligned}
f(\vec{x}') &= y' \\
&\Leftrightarrow \exp(\vec{w}^\top \vec{x}' + b) = y' \\
&\Leftrightarrow \vec{w}^\top \vec{x}' + b - \log(y') = 0 \\
&\Leftrightarrow \vec{w}^\top \vec{x}' + c = 0
\end{aligned} \tag{61}$$

where

$$c = b - \log(y') \tag{62}$$

Finally, we obtain the following set of linear inequality constraints:

$$\begin{aligned}
\vec{w}^\top \vec{x}' + c &\leq \epsilon \\
-\vec{w}^\top \vec{x}' - c &\leq \epsilon
\end{aligned} \tag{63}$$

where we introduced the parameter $\epsilon \geq 0$ that specifies the maximum tolerated deviation from the requested prediction - we set $\epsilon = 0$ if we do not allow any deviations.

6.3.4 Exponential regression

Recall that the prediction function f is given as:

$$f(\vec{x}) = -\frac{1}{\vec{w}^\top \vec{x} + b} \tag{64}$$

The constraint for a specific prediction y' is given as:

$$\begin{aligned}
f(\vec{x}') &= y' \\
&\Leftrightarrow -\frac{1}{\vec{w}^\top \vec{x}' + b} = y' \\
&\Leftrightarrow \vec{w}^\top \vec{x}' + b + \frac{1}{y'} = 0 \\
&\Leftrightarrow \vec{w}^\top \vec{x}' + c = 0
\end{aligned} \tag{65}$$

where

$$c = b + \frac{1}{y'} \tag{66}$$

Finally, we obtain the following set of linear inequality constraints:

$$\begin{aligned}
\vec{w}^\top \vec{x}' + c &\leq \epsilon \\
-\vec{w}^\top \vec{x}' - c &\leq \epsilon
\end{aligned} \tag{67}$$

where we introduced the parameter $\epsilon \geq 0$ that specifies the maximum tolerated deviation from the requested prediction - we set $\epsilon = 0$ if we do not allow any deviations.

6.4 Gaussian naive bayes

Recall that the prediction function h is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \prod_{k=1}^d \mathcal{N}(\vec{x} \mid \mu_{ik}, \sigma_{ik}^2) \pi_i \quad (68)$$

We note that Eq. 68 is equivalent to

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \sum_{k=1}^d \log(\mathcal{N}(\vec{x} \mid \mu_{ik}, \sigma_{ik}^2)) + \log(\pi_i) \quad (69)$$

Simplifying the term in Eq. 69 yields

$$\begin{aligned} \log(\pi_i) + \sum_{k=1}^d \log(\mathcal{N}(\vec{x} \mid \mu_{ik}, \sigma_{ik}^2)) &= \log(\pi_i) + \sum_{k=1}^d \log\left(\frac{1}{\sqrt{2\pi\sigma_{ik}^2}}\right) + \\ &\quad \sum_{k=1}^d -\frac{1}{2\sigma_{ik}^2} \|(\vec{x})_k - \mu_{ik}\|_2^2 \\ &= \log(\pi_i) + \sum_{k=1}^d \log\left(\frac{1}{\sqrt{2\pi\sigma_{ik}^2}}\right) - \\ &\quad \sum_{k=1}^d \frac{1}{2\sigma_{ik}^2} \left((\vec{x})_k^2 + \mu_{ik}^2 - 2(\vec{x})_k \mu_{ik} \right) \\ &= c_i - \vec{x}^\top \mathbf{A}_i \vec{x} + \vec{q}_i^\top \vec{x} \end{aligned} \quad (70)$$

where

$$c_i = \log(\pi_i) + \sum_{k=1}^d \log\left(\frac{1}{\sqrt{2\pi\sigma_{ik}^2}}\right) - \frac{\mu_{ik}^2}{2\sigma_{ik}^2} \quad (71)$$

$$\mathbf{A}_i = \text{diag}\left(\frac{1}{2\sigma_{ik}^2}\right) \quad (72)$$

$$\vec{q}_i = \left(\frac{\mu_{i1}}{\sigma_{i1}^2}, \dots, \frac{\mu_{id}}{\sigma_{id}^2}\right)^\top \quad (73)$$

For a sample \vec{x} , in order to be classified as the i -th class, the following set of strict inequalities must hold:

$$c_i - \vec{x}^\top \mathbf{A}_i \vec{x} + \vec{q}_i^\top \vec{x} > c_j - \vec{x}^\top \mathbf{A}_j \vec{x} + \vec{q}_j^\top \vec{x} \quad \forall j \neq i \quad (74)$$

By rearranging terms in Eq. 74, we get the final constraints

$$\vec{x}^\top \mathbf{A}_{ij} \vec{x} + \vec{q}_{ij}^\top \vec{x} + c_{ij} < 0 \quad \forall j \neq i \quad (75)$$

where

$$\mathbf{A}_{ij} = \mathbf{A}_i - \mathbf{A}_j = \text{diag} \left(\frac{1}{2\sigma_{ik}^2} - \frac{1}{2\sigma_{jk}^2} \right) \quad (76)$$

$$\vec{q}_{ij} = \vec{q}_j - \vec{q}_i = \left(\frac{\mu_{j1}}{\sigma_{j1}^2} - \frac{\mu_{i1}}{\sigma_{i1}^2}, \dots, \frac{\mu_{jd}}{\sigma_{jd}^2} - \frac{\mu_{id}}{\sigma_{id}^2} \right)^\top \quad (77)$$

$$\begin{aligned} c_{ij} &= c_j - c_i = \log(\pi_j) - \log(\pi_i) + \\ &\quad \sum_{k=1}^d \log \left(\frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \right) - \frac{\mu_{jk}^2}{2\sigma_{jk}^2} - \log \left(\frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \right) + \frac{\mu_{ik}^2}{2\sigma_{ik}^2} \\ &= \log \left(\frac{\pi_j}{\pi_i} \right) + \sum_{k=1}^d \log \left(\frac{\sqrt{2\pi\sigma_{ik}^2}}{\sqrt{2\pi\sigma_{jk}^2}} \right) - \frac{\mu_{jk}^2}{2\sigma_{jk}^2} + \frac{\mu_{ik}^2}{2\sigma_{ik}^2} \end{aligned} \quad (78)$$

Because we can not make any statement about the definiteness of the diagonal matrix \mathbf{A}_{ij} , the constraint Eq. 75 is a non-convex quadratic inequality constraint.

6.5 Quadratic discriminant analysis

Recall that the prediction function h is given as:

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \mathbf{\Sigma}_i) \pi_i \quad (79)$$

We can rewrite Eq. 79 as

$$h(\vec{x}) = \arg \max_{i \in \mathcal{Y}} \log \left(\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \mathbf{\Sigma}_i) \pi_i \right) \quad (80)$$

Working on the log term yields

$$\begin{aligned} \log \left(\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \mathbf{\Sigma}_i) \pi_i \right) &= \log \left(\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \mathbf{\Sigma}_i) \right) + \log(\pi_i) \\ &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log \left(\det(\mathbf{\Sigma}_i^{-1}) \right) - \frac{1}{2} (\vec{x} - \vec{\mu}_i)^\top \mathbf{\Sigma}_i^{-1} (\vec{x} - \vec{\mu}_i) \\ &\quad + \log(\pi_i) \\ &= -\frac{1}{2} \vec{x}^\top \mathbf{\Sigma}_i^{-1} \vec{x} + \vec{x}^\top \vec{q}_i + c_i \end{aligned} \quad (81)$$

where

$$\vec{q}_i = \mathbf{\Sigma}_i^{-1} \vec{\mu}_i \quad (82)$$

$$c_i = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log \left(\det(\mathbf{\Sigma}_i) \right) - \frac{1}{2} \vec{\mu}_i^\top \mathbf{\Sigma}_i^{-1} \vec{\mu}_i + \log(\pi_i) \quad (83)$$

For a sample \vec{x} , in order to be classified as the i -th class, the following set of strict inequalities must hold:

$$-\frac{1}{2}\vec{x}^\top \mathbf{\Sigma}_i^{-1} \vec{x} + \vec{x}^\top \vec{q}_i + c_i > -\frac{1}{2}\vec{x}^\top \mathbf{\Sigma}_j^{-1} \vec{x} + \vec{x}^\top \vec{q}_j + c_j \quad \forall j \neq i \quad (84)$$

Rearranging Eq. 84 yields

$$\frac{1}{2}\vec{x}^\top \mathbf{A}_{ij} \vec{x} + \vec{x}^\top \vec{q}_{ij} + c_{ij} < 0 \quad \forall j \neq i \quad (85)$$

where

$$\mathbf{A}_{ij} = \mathbf{\Sigma}_i^{-1} - \mathbf{\Sigma}_j^{-1} \quad (86)$$

$$\vec{q}_{ij} = \vec{q}_j - \vec{q}_i = \mathbf{\Sigma}_j^{-1} \vec{\mu}_j - \mathbf{\Sigma}_i^{-1} \vec{\mu}_i \quad (87)$$

$$\begin{aligned} c_{ij} &= c_j - c_i \\ &= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(\det(\mathbf{\Sigma}_j)) - \frac{1}{2} \vec{\mu}_j^\top \mathbf{\Sigma}_j^{-1} \vec{\mu}_j + \log(\pi_j) + \\ &\quad \frac{d}{2} \log(2\pi) + \frac{1}{2} \log(\det(\mathbf{\Sigma}_i)) + \frac{1}{2} \vec{\mu}_i^\top \mathbf{\Sigma}_i^{-1} \vec{\mu}_i - \log(\pi_i) \\ &= \frac{1}{2} \left(\vec{\mu}_i^\top \mathbf{\Sigma}_i^{-1} \vec{\mu}_i - \vec{\mu}_j^\top \mathbf{\Sigma}_j^{-1} \vec{\mu}_j \right) + \frac{1}{2} \log \left(\frac{\det(\mathbf{\Sigma}_i)}{\det(\mathbf{\Sigma}_j)} \right) + \log \left(\frac{\pi_j}{\pi_i} \right) \end{aligned} \quad (88)$$

The final constraint Eq. 85 is a non-convex quadratic constraint because we can not make any statement about the definiteness of \mathbf{A}_{ij} .

6.6 Learning vector quantization

Note: The subsequent sections are taken from [18].

6.6.1 Enforcing a specific prototype as the nearest neighbor

By using the following set of inequalities, we can force the prototype \vec{p}_i to be the nearest neighbor of the counterfactual \vec{x}' - which would cause \vec{x}' to be classified as o_i :

$$d(\vec{x}', \vec{p}_i) < d(\vec{x}', \vec{p}_j) \quad \forall \vec{p}_j \in \mathcal{P}(y') \quad (89)$$

We consider a fixed pair of i and j :

$$\begin{aligned} d(\vec{x}', \vec{p}_i) &< d(\vec{x}', \vec{p}_j) \\ \Leftrightarrow \|\vec{x}' - \vec{p}_i\|_{\mathbf{\Omega}_i}^2 &< \|\vec{x}' - \vec{p}_j\|_{\mathbf{\Omega}_j}^2 \\ \Leftrightarrow (\vec{x}' - \vec{p}_i)^\top \mathbf{\Omega}_i (\vec{x}' - \vec{p}_i) &< (\vec{x}' - \vec{p}_j)^\top \mathbf{\Omega}_j (\vec{x}' - \vec{p}_j) \\ \Leftrightarrow \vec{x}'^\top \mathbf{\Omega}_i \vec{x}' - 2\vec{x}'^\top \mathbf{\Omega}_i \vec{p}_i + \vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i &< \vec{x}'^\top \mathbf{\Omega}_j \vec{x}' - 2\vec{x}'^\top \mathbf{\Omega}_j \vec{p}_j + \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j \\ \Leftrightarrow \vec{x}'^\top \mathbf{\Omega}_i \vec{x}' - \vec{x}'^\top \mathbf{\Omega}_j \vec{x}' - 2\vec{x}'^\top \mathbf{\Omega}_i \vec{p}_i + 2\vec{x}'^\top \mathbf{\Omega}_j \vec{p}_j + \vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j &< 0 \quad (90) \\ \Leftrightarrow \vec{x}'^\top (\mathbf{\Omega}_i - \mathbf{\Omega}_j) \vec{x}' + \vec{x}'^\top (-2\mathbf{\Omega}_i \vec{p}_i + 2\mathbf{\Omega}_j \vec{p}_j) + (\vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j) & \\ \Leftrightarrow \frac{1}{2} \vec{x}'^\top (\mathbf{\Omega}_i - \mathbf{\Omega}_j) \vec{x}' + \frac{1}{2} \vec{x}'^\top (\mathbf{\Omega}_j \vec{p}_j - \mathbf{\Omega}_i \vec{p}_i) + \frac{1}{2} (\vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j) &< 0 \\ \Leftrightarrow \frac{1}{2} \vec{x}'^\top \mathbf{A}_{ij} \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + c_{ij} &< 0 \end{aligned}$$

where

$$\mathbf{A}_{ij} = \mathbf{\Omega}_i - \mathbf{\Omega}_j \quad (91)$$

$$\vec{q}_{ij} = \frac{1}{2} (\mathbf{\Omega}_j \vec{p}_j - \mathbf{\Omega}_i \vec{p}_i) \quad (92)$$

$$c_{ij} = \frac{1}{2} (\vec{p}_i^\top \mathbf{\Omega}_i \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega}_j \vec{p}_j) \quad (93)$$

If we only have one global distance matrix $\mathbf{\Omega}$, we find that $\mathbf{A}_{ij} = \mathbf{0}$ and the inequality Eq. 90 simplifies:

$$\begin{aligned} d(\vec{x}, \vec{p}_i) &< d(\vec{x}, \vec{p}_j) \\ \Leftrightarrow \vec{x}'^\top \vec{q}_{ij} + c_{ij} &< 0 \end{aligned} \quad (94)$$

where

$$\vec{q}_{ij} = \frac{1}{2} \mathbf{\Omega} (\vec{p}_j - \vec{p}_i) \quad (95)$$

$$c_{ij} = \frac{1}{2} (\vec{p}_i^\top \mathbf{\Omega} \vec{p}_i - \vec{p}_j^\top \mathbf{\Omega} \vec{p}_j) \quad (96)$$

If we do not use a custom distance matrix, we have $\mathbf{\Omega} = \mathbb{I}$ and Eq. 90 becomes:

$$\begin{aligned} d(\vec{x}, \vec{p}_i) &< d(\vec{x}, \vec{p}_j) \\ \Leftrightarrow \vec{x}'^\top \vec{q}_{ij} + c_{ij} &< 0 \end{aligned} \quad (97)$$

where

$$\vec{q}_{ij} = \frac{1}{2} (\vec{p}_j - \vec{p}_i) \quad (98)$$

$$c_{ij} = \frac{1}{2} (\vec{p}_i^\top \vec{p}_i - \vec{p}_j^\top \vec{p}_j) \quad (99)$$

6.7 Minimizing the Euclidean distance

Minimizing the Euclidean distance (Eq. 36) yields a *quadratic* objective.

First, we expand the Euclidean distance (Eq. 36):

$$\begin{aligned} \|\vec{x}' - \vec{x}\|_2^2 &= (\vec{x}' - \vec{x})^\top (\vec{x}' - \vec{x}) \\ &= \vec{x}'^\top \vec{x}' - \vec{x}'^\top \vec{x} - \vec{x}^\top \vec{x}' + \vec{x}^\top \vec{x} \\ &= \vec{x}'^\top \vec{x}' - 2\vec{x}^\top \vec{x}' + \vec{x}^\top \vec{x} \end{aligned} \quad (100)$$

Next, we note that that we can drop the constant $\vec{x}^\top \vec{x}$ when optimizing with respect to \vec{x}' :

$$\begin{aligned} \min_{\vec{x}' \in \mathbb{R}^d} \|\vec{x}' - \vec{x}\|_2^2 \\ \Leftrightarrow \\ \min_{\vec{x}' \in \mathbb{R}^d} \frac{1}{2} \vec{x}'^\top \vec{x}' - \vec{x}^\top \vec{x}' \end{aligned} \quad (101)$$

6.8 Minimizing the weighted Manhattan distance

Minimizing the weighted Manhattan distance (Eq. 2) yields a *linear* objective.

First, we transform the problem of minimizing the weighted Manhattan distance (Eq. 2) into epigraph form:

$$\begin{aligned}
& \min_{\vec{x}' \in \mathbb{R}^d} \sum_j \alpha_j \cdot |(\vec{x}')_j - (\vec{x})_j| \\
& \Leftrightarrow \\
& \min_{\vec{x}' \in \mathbb{R}^d, \beta \in \mathbb{R}} \beta \\
& \text{s.t.} \quad \sum_j \alpha_j \cdot |(\vec{x}')_j - (\vec{x})_j| \leq \beta \\
& \quad \beta \geq 0
\end{aligned} \tag{102}$$

Next, we separate the dimensions:

$$\begin{aligned}
& \min_{\vec{x}' \in \mathbb{R}^d, \beta \in \mathbb{R}} \beta \\
& \text{s.t.} \quad \sum_j \alpha_j \cdot |(\vec{x}')_j - (\vec{x})_j| \leq \beta \\
& \quad \beta \geq 0 \\
& \Leftrightarrow \\
& \min_{\vec{x}', \vec{\beta} \in \mathbb{R}^d} \sum_j (\vec{\beta})_j \\
& \text{s.t.} \quad \alpha_j \cdot |(\vec{x}')_j - (\vec{x})_j| \leq (\vec{\beta})_j \quad \forall j \\
& \quad (\vec{\beta})_j \geq 0 \quad \forall j
\end{aligned} \tag{103}$$

After that, we remove the absolute value function:

$$\begin{aligned}
& \min_{\vec{x}', \vec{\beta} \in \mathbb{R}^d} \sum_j (\vec{\beta})_j \\
& \text{s.t.} \quad \alpha_j \cdot |(\vec{x}')_j - (\vec{x})_j| \leq (\vec{\beta})_j \quad \forall j \\
& \quad (\vec{\beta})_j \geq 0 \quad \forall j \\
& \Leftrightarrow \\
& \min_{\vec{x}', \vec{\beta} \in \mathbb{R}^d} \sum_j (\vec{\beta})_j \\
& \text{s.t.} \quad \alpha_j (\vec{x}')_j - \alpha_j (\vec{x})_j \leq (\vec{\beta})_j \quad \forall j \\
& \quad -\alpha_j (\vec{x}')_j + \alpha_j (\vec{x})_j \leq (\vec{\beta})_j \quad \forall j \\
& \quad (\vec{\beta})_j \geq 0 \quad \forall j
\end{aligned} \tag{104}$$

Finally, we rewrite everything in matrix-vector notation:

$$\begin{aligned}
& \min_{\vec{x}', \vec{\beta} \in \mathbb{R}^d} \vec{1}^\top \vec{\beta} \\
& \text{s.t.} \\
& \Upsilon \vec{x}' - \Upsilon \vec{x} \leq \vec{\beta} \\
& -\Upsilon \vec{x}' + \Upsilon \vec{x} \leq \vec{\beta} \\
& \vec{\beta} \geq \vec{0}
\end{aligned} \tag{105}$$

where

$$\Upsilon = \text{diag}(\alpha_j) \tag{106}$$

6.9 Solving a non-convex QCQP

Solving a non-convex QCQP is known to be NP-hard [28, 39].

In section 6.9.1 we discuss a method for approximately solving a non-convex QCQP and in section 6.9.2 we describe how to solve the special case of a non-convex QCQP having a single constraint.

6.9.1 Approximately solving a non-convex QCQP

Recall the non-convex quadratic constraint:

$$\frac{1}{2} \vec{x}'^\top \mathbf{A}_{ij} \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + r_{ij} \leq 0 \tag{107}$$

In this paper, we always defined the matrix \mathbf{A}_{ij} as the difference of two s.p.s.d. matrices \mathbf{A}_i and \mathbf{A}_j :

$$\mathbf{A}_{ij} = \mathbf{A}_i - \mathbf{A}_j \tag{108}$$

By making use of Eq. 108, we can rewrite Eq. 107 as:

$$\begin{aligned}
& \frac{1}{2} \vec{x}'^\top \mathbf{A}_i \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + r_{ij} - \frac{1}{2} \vec{x}'^\top \mathbf{A}_j \vec{x}' \leq 0 \\
& \Leftrightarrow f(\vec{x}') - g(\vec{x}') \leq 0
\end{aligned} \tag{109}$$

where

$$f(\vec{x}') = \frac{1}{2} \vec{x}'^\top \mathbf{A}_i \vec{x}' + \vec{x}'^\top \vec{q}_{ij} + r_{ij} \tag{110}$$

$$g(\vec{x}') = \frac{1}{2} \vec{x}'^\top \mathbf{A}_j \vec{x}' \tag{111}$$

Under the assumption that our regularization function $\theta(\cdot)$ is a convex function⁴, we can rewrite a generic version of the non-convex QCQP Eq. 41 as follows:

$$\begin{aligned}
& \min_{\vec{x}' \in \mathbb{R}^d} \theta(\vec{x}', \vec{x}) \\
& \text{s.t.} \\
& f(\vec{x}') - g(\vec{x}') \leq 0
\end{aligned} \tag{112}$$

⁴The weighted Manhattan distance and the Euclidean distance are convex functions!

Because \mathbf{A}_i and \mathbf{A}_j are s.p.s.d. matrices, we know that $f(\vec{x}')$ and $g(\vec{x}')$ are convex functions. Therefore, Eq. 112 is a difference-of-convex program (DCP).

This allows us to use the penalty convex-concave procedure (CCP) [28] for computing an approximate solution of Eq. 112, yielding an approximate solution of the original non-convex QCQP. For using the penalty CCP, we need the first order Taylor approximation of $g(\vec{x}')$ around a current point \vec{x}_k :

$$\begin{aligned}\hat{g}(\vec{x}')_{\vec{x}_k} &= g(\vec{x}_k) + (\nabla_{\vec{x}'} g)(\vec{x}_k)^\top (\vec{x}' - \vec{x}_k) \\ &= \frac{1}{2} \vec{x}_k^\top \mathbf{A}_j \vec{x}_k + (\mathbf{A}_j \vec{x}_k)^\top (\vec{x}' - \vec{x}_k) \\ &= (\mathbf{A}_j \vec{x}_k)^\top \vec{x}' + \frac{1}{2} \vec{x}_k^\top \mathbf{A}_j \vec{x}_k - (\mathbf{A}_j \vec{x}_k)^\top \vec{x}_k \\ &= \vec{\rho}_{jk}^\top \vec{x}' + \tilde{c}_{jk}\end{aligned}\tag{113}$$

where

$$\vec{\rho}_{jk} = \mathbf{A}_j \vec{x}_k \tag{114}$$

$$\tilde{c}_{jk} = -\frac{1}{2} \vec{x}_k^\top \mathbf{A}_j \vec{x}_k \tag{115}$$

In order to run the convex-concave procedure, we have to provide an initial (feasible) solution. We could either use the original data point as an initial *infeasible* solution, some data point yielding the requested prediction as an initial *feasible* solution or some other “smart” initialization.

As an alternative, we could use other methods for approximately computing a solution of the non-convex QCQP like the Suggest-Improve framework [28] - actually, the methods we described in the previous paragraph is an instance of the Suggest-Improve framework.

6.9.2 Solving a non-convex QCQP with just one constraint

We consider the general QCQP

$$\begin{aligned}\min_{\vec{x}' \in \mathbb{R}^d} \quad & \vec{x}'^\top \mathbf{Q} \vec{x}' + \vec{q}^\top \vec{x}' + c \\ \text{s.t.} \quad & \\ & \vec{x}'^\top \mathbf{A} \vec{x}' + \vec{b}^\top \vec{x}' + r \leq 0\end{aligned}\tag{116}$$

where $\mathbf{Q}, \mathbf{A} \in \mathbb{R}^{d \times d}$, $\vec{q}, \vec{b} \in \mathbb{R}^d$ and $c, r \in \mathbb{R}$.

If \mathbf{Q} and \mathbf{A} are not symmetric positive semi-definite, Eq. 116 is a non-convex QCQP. However, despite the non-convexity, we can solve Eq. 116 efficiently by solving the dual of Eq. 116 and observing that the duality gap is zero [39] - under the assumption that Eq. 116 is strictly feasible⁵.

⁵We can always achieve strict feasibility by moving a non-strict feasible point away from the decision boundary.

Therefore, solving Eq. 116 is equivalent to solving the following semi-definite program (SDP) [39]:

$$\begin{aligned}
& \arg \min_{\mathbf{X} \in \mathcal{S}^d, \vec{x}' \in \mathbb{R}^d} \quad \text{trace}(\mathbf{Q}\mathbf{X}) + \vec{q}^\top \vec{x}' + c \\
& \text{s.t.} \\
& \text{trace}(\mathbf{A}\mathbf{X}) + \vec{b}^\top \vec{x}' + r \leq 0 \\
& \begin{pmatrix} \mathbf{X} & \vec{x}' \\ \vec{x}'^\top & 1 \end{pmatrix} \succeq 0
\end{aligned} \tag{117}$$

where we introduced an additional variable⁶ \mathbf{X} that can be discarded afterwards.

References

- [1] Victor Talpaert, Ibrahim Sobh, B. Ravi Kiran, Patrick Mannion, Senthil Yogamani, Ahmad El Sallab, and Patrick Perez. Exploring applications of deep reinforcement learning for real-world autonomous driving systems. *CoRR*, abs/1901.01536, 2019.
- [2] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara N. Sainath. Deep learning for audio signal processing. *CoRR*, abs/1905.00078, 2019.
- [3] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. A survey of the usages of deep learning in natural language processing. *CoRR*, abs/1807.10854, 2018.
- [4] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *CoRR*, abs/1907.09408, 2019.
- [5] Panagiotis Stalidis, Theodoros Semertzidis, and Petros Daras. Examining deep learning architectures for crime classification and prediction. abs/1812.00602, 2018.
- [6] Amir E. Khandani, Adlar J. Kim, and Andrew Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.
- [7] Christoph Molnar. *Interpretable Machine Learning*. 2019. <https://christophm.github.io/interpretable-ml-book/>.
- [8] European parliament and council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016.
- [9] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019.
- [10] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018*, pages 80–89, 2018.
- [11] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017.
- [12] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

⁶ \mathcal{S}^d denotes the set of symmetric $\mathbb{R}^{d \times d}$ matrices

- [13] Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. A simple and effective model-based variable importance measure. CoRR, abs/1805.04755, 2018.
- [14] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. arXiv e-prints, page arXiv:1801.01489, Jan 2018.
- [15] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. Journal of Business & Economic Statistics, 0(ja):1–19, 2019.
- [16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.
- [17] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. CoRR, abs/1805.10820, 2018.
- [18] André Artelt and Barbara Hammer. Efficient computation of counterfactual explanations of LVQ models. arXiv e-prints, page arXiv:1908.00735, Aug 2019.
- [19] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications, 1994.
- [20] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning, ICMML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 1885–1894, 2017.
- [21] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2280–2288, 2016.
- [22] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part I, pages 100–111, 2018.
- [23] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. CoRR, abs/1907.02584, 2019.
- [24] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodriguez, Tijl De Bie, and Peter A. Flach. FACE: feasible and actionable counterfactual explanations. CoRR, abs/1909.09369, 2019.
- [25] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. CoRR, abs/1905.07857, 2019.
- [26] Chris Russell. Efficient search for diverse coherent explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019, pages 20–28, 2019.
- [27] Laurence A. Wolsey. Mixed Integer Programming, pages 1–10. American Cancer Society, 2008.
- [28] Jaehyun Park and Stephen Boyd. General heuristics for nonconvex quadratically constrained quadratic programming. arXiv preprint arXiv:1703.07870, 2017.
- [29] David Nova and Pablo A. Estévez. A review of learning vector quantization classifiers. Neural Comput. Appl., 25(3-4):511–524, September 2014.
- [30] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. Neural Computation, 21(12):3532–3561, 2009. PMID: 19764875.
- [31] Petra Schneider, Michael Biehl, and Barbara Hammer. Distance learning in discriminative vector quantization. Neural Computation, 21(10):2942–2969, 2009. PMID: 19635012.

- [32] Himani Sharma and Sunil Kumar. A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5, 04 2016.
- [33] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, pages 465–474, New York, NY, USA, 2017. ACM.
- [34] André Artelt. Ceml: Counterfactuals for explaining machine learning models - a python toolbox. <https://www.github.com/andreArtelt/ceml>, 2019.
- [35] Guido Van Rossum and Fred L Drake Jr. Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
- [37] Stéfan van der Walt, S. Chris Colbert, and Gaël Varoquaux. The numpy array: A structure for efficient numerical computation. Computing in Science and Engineering, 13(2):22–30, 2011.
- [38] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research, 17(83):1–5, 2016.
- [39] Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, New York, NY, USA, 2004.