

# Explaining Reject Options of Learning Vector Quantization Classifiers <sup>★</sup>

André Artelt<sup>1</sup>, Johannes Brinkrolf<sup>1</sup>, Roel Visser<sup>1</sup> and Barbara Hammer<sup>1</sup>

CITEC - Cognitive Interaction Technology  
Bielefeld University, 33619 Bielefeld, Germany  
{aartelt,jbrinkro,rvisser,bhammer}@techfak.uni-bielefeld.de

**Abstract.** While machine learning models are usually assumed to always output a prediction, there also exist extensions in the form of reject options which allow the model to reject inputs where only a prediction with an unacceptably low certainty would be possible. With the ongoing rise of eXplainable AI, a lot of methods for explaining model predictions have been developed. However, understanding why a given input was rejected, instead of being classified by the model, is also of interest. Surprisingly, explanations of rejects have not been considered so far.

We propose to use counterfactual explanations for explaining rejects and investigate how to efficiently compute counterfactual explanations of different reject options for an important class of models, namely prototype-based classifiers such as learning vector quantization models.

**Keywords:** XAI · Contrasting Explanations · Learning Vector Quantization · Reject Options

## 1 Introduction

Nowadays, machine learning (ML) based decision making systems are increasingly used in safety-critical or high-impact applications like autonomous driving [35], credit (risk) assessment [19] and predictive policing [42]. Because of this, there is an increasing demand for transparency which was also recognized by the policy makers and emphasized in legal regulations like the EU’s GDPR [31]. It is a common approach to realize transparency by explanations – i.e. providing an explanation of why the system behaved in the way it did – which gave rise to the field of explainable artificial intelligence (XAI or eXplainable AI) [36, 43]. Although it is still unclear what exactly makes up a “good” explanation [12, 29], a lot of different explanation methods have been developed [17, 25]. Popular explanations methods [25, 43] are feature relevance/importance methods [15] and examples based methods [1]. Instances of example based methods are contrasting explanations like counterfactual explanations [44, 45] and prototypes &

---

<sup>★</sup> We gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for grant TRR 318/1 2021 438445824, and the VW-Foundation for the project *IMPACT* funded in the frame of the funding line *AI and its Implications for Future Society*.

criticisms [20] – these methods use a set or a single example for explaining the behavior of the system.

Another important aspect, in particular in safety-critical and high-risk applications, is reliability: If the system is not “absolutely” certain about its decision/prediction it might be better to refuse making a prediction and, for instance, pass the input back to human – if making mistakes is “expensive” or critical, the system should reject inputs where it is not certain enough in its prediction. As a consequence, a mechanism called reject options has been pioneered [11], where optimal reject rates are determined based on costs assigned to miss-classifications (e.g. false positives and false negatives) and rejects. Many realizations of reject options are based on probabilities, like class probabilities in classification [11]. However, not all models output probabilities along with their predictions or if they do, their computed probabilities might be of “bad quality” (e.g. just a score in  $[0, 1]$  without any probabilistic/statistical foundation). One possible remedy is to use a general (i.e. model agnostic) post-processing method or wrapper on top of the model for computing reliable certainty scores, as is done in [18], or use a method like conformal prediction [40] in which a non-conformity measure together with a hold-out data set is used for computing certainties and confidences of predictions. Another option is to develop model specific certainty measures and consequently reject options, which for instance was done for prototype-based methods like learning vector quantization (LVQ) models. These are a class of models that convinced with simplicity (and thus interpretability), while still being powerful predictive models, and have also excelled in settings like life-long learning and biomedical applications [8, 13, 14, 21, 28, 30, 47].

We think that while explaining the prediction of the model is important, similarly explaining why a given input was rejected is also important – i.e. explaining why the model “thinks” that it does not know enough for making a proper and reliable prediction. *For instance, consider a biomedical application where a model is supposed to assist in some kind of early cancer detection – such a scenario could be considered a high-risk application where the addition of a reject option to the model would be necessary. In such a scenario, it would be very useful if a reject is accompanied with an explanation of why this sample was rejected, because by this we could learn more about the model and the domain itself – e.g. an explanation could state that some combination of serum values is very unusual compared to what the model has seen before.* Surprisingly, (to the best of our knowledge) explaining rejects have not been considered so far.

*Contributions* In this work, we focus on prototype-based models such as learning vector quantization (LVQ) models and propose to explain LVQ reject options by means of counterfactual explanations – i.e. explaining why a particular sample was rejected by the LVQ model. Whereby we consider different reject options separately and study how to efficiently compute counterfactual explanations in each of these cases.

The remainder of this work is structured as follows: After reviewing the foundations of learning vector quantization (Section 2.1), popular reject options for LVQ (Section 2.2), and counterfactual explanations (Section 2.3), we propose

a modeling and algorithms for efficiently computing counterfactual explanations of different LVQ reject options (Section 3). In Section 4, we empirically evaluate our proposed modelings and algorithms on several different data sets with respect to different aspects. Finally, this work closes with a summary and conclusion in Section 5. Note that for the purpose of readability, all proofs and derivations are given in Appendix A.

## 2 Foundations

In the following, we briefly review the foundations our work is based on.

### 2.1 Learning Vector Quantization

In this work, we focus on learning vector quantization (LVQ) models. In modern variants, these models constitute state-of-the-art classifiers which can also be used for incremental, online, and federated learning [4, 9, 16, 24]. In its basic variant as proposed in [22], it relies on Hebbian learning. However, further extensions based on cost functions have been introduced – these have the benefit that they can be easily extended to a more flexible metric learning scheme [39]. First, we introduce generalized LVQ (GLVQ) [37] and then take a look at extensions for metric learning. In the following, we assume that our the domain is the real valued vector space – i.e. the classification scenario takes place in  $\mathbb{R}^d$  with a fix number of classes  $k$  which are enumerated as  $\{1, \dots, k\} = \mathcal{Y}$ . An LVQ model is characterized by  $m$  labeled prototypes  $(\mathbf{p}_j, c(\mathbf{p}_j)) \in \mathbb{R}^d \times \mathcal{Y}$ ,  $j \in \{1, \dots, m\}$ , whereby the labels  $c(\mathbf{p}_j)$  of the prototypes are fixed. Classification of a sample  $\mathbf{x} \in \mathbb{R}^d$  takes place by a winner takes all scheme:

$$\mathbf{x} \mapsto c(\mathbf{p}_j) \text{ with } j = \arg \min_{j \in \{1, \dots, m\}} d(\mathbf{x}, \mathbf{p}_j) \quad (1)$$

where the squared Euclidean distance is used:

$$d(\mathbf{x}, \mathbf{p}_j) = (\mathbf{x} - \mathbf{p}_j)^\top (\mathbf{x} - \mathbf{p}_j) \quad (2)$$

Note that the prototypes' positions not only allow an interpretable classification but also act as a representation of the data and its underlying classes. Training of LVQ models is done in a supervised fashion – i.e. based on given labeled samples  $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}$  with  $i \in \{1, \dots, N\}$ . For GLVQ [37], the learning rule originates from minimizing the following cost function:

$$E = \sum_{i=1}^N \text{sgd} \left( \frac{d(\mathbf{x}_i, \mathbf{p}_+) - d(\mathbf{x}_i, \mathbf{p}_-)}{d(\mathbf{x}_i, \mathbf{p}_+) + d(\mathbf{x}_i, \mathbf{p}_-)} \right) \quad (3)$$

where  $\text{sgd}(\cdot)$  is a monotonously increasing function – e.g. the logistic function or the identity. Furthermore,  $\mathbf{p}_+$  denotes the closest prototype with the same label – i.e.  $c(\mathbf{x}) = c(\mathbf{p}_+)$  –, and  $\mathbf{p}_-$  denotes the closest prototype which belongs

to a different class. Note that the models' performance heavily depends on the suitability of the Euclidean metric for the given classification problem – note that this is often not the case, in particular in case of different feature relevances. Because of this, a powerful metric learning scheme has been proposed in [39] which substitutes the squared Euclidean metric in Eq. (2) by a weighted alternative:

$$d(\mathbf{x}, \mathbf{p}) = (\mathbf{x} - \mathbf{p})^\top \boldsymbol{\Omega} (\mathbf{x} - \mathbf{p}) \quad (4)$$

where  $\boldsymbol{\Omega} \in \mathcal{S}_+^d$  refers to a positive semidefinite matrix encoding the relevance of each feature. This matrix is treated as an additional parameter which is, together with the prototypes' positions, chosen such that the cost function Eq. (3) is minimized – usually, a gradient based method like LBFGS is used. Due to the parameterization of the metric, this LVQ variant is often named generalized matrix LVQ (GMLVQ). Further details can be found in [39].

## 2.2 Reject Options

LVQ schemes provide a classification rule which assigns a label to every possible input no matter how reliable & reasonable such classifications might be. Reject options allow the classifier to reject a sample if a certain prediction is not possible – i.e. the sample is too close to the decision boundary, or it is very different from the observed training data and therefore a classification based on the learned prototypes would not be reasonable. For LVQ models, as reviewed in the previous section, there exist some reject options which allows the model to deny a classification of a sample if a reliable & certain classification is not possible. Generally speaking, we extend the set of possible predictions by a new class which represents a reject – i.e. for a classifier  $h : \mathcal{X} \rightarrow \mathcal{Y}$ , we construct  $h : \mathcal{X} \rightarrow \mathcal{Y} \cup \{R\}$ . Quite a few methods for doing so have been proposed and evaluated in [13]. Those methods use a function  $r : \mathcal{X} \rightarrow \mathbb{R}$  for computing the certainty of a prediction by the classifier [11]. If the certainty is below some threshold  $\theta$ , the sample is rejected, more formally if

$$r(\mathbf{x}) < \theta \quad (5)$$

then  $h(\mathbf{x}) = R$ . In the following, we consider three popular realizations of the certainty function  $r(\cdot)$  in the context of LVQ models.

*Relative similarity:* In [13], a very natural realization of  $r(\cdot)$  called relative similarity (RelSim) yields excellent results for LVQ models:

$$r_{\text{RelSim}}(\mathbf{x}) = \frac{d(\mathbf{x}, \mathbf{p}_-) - d(\mathbf{x}, \mathbf{p}_+)}{d(\mathbf{x}, \mathbf{p}_-) + d(\mathbf{x}, \mathbf{p}_+)} \quad (6)$$

where  $\mathbf{p}_+$  denotes the closest prototype and  $\mathbf{p}_-$  denotes the closest prototype belonging to a different class than  $\mathbf{p}_+$ . Note that this fraction is always non-negative and smaller than 1. Obviously, it is 0 if the distance between the sample and the closest prototype  $\mathbf{p}_+$  equals the distance to  $\mathbf{p}_-$ . At the same time, RelSim gets close to 0 if the sample is far away – i.e. both distances are large [6].

*Decision boundary distance:* Another, similar, realization of a certainty measure is the distance to the decision boundary (Dist). In case of LVQ, this can be formalized as follows [13]:

$$r_{\text{Dist}}(\mathbf{x}) = \frac{|\text{d}(\mathbf{x}, \mathbf{p}_+) - \text{d}(\mathbf{x}, \mathbf{p}_-)|}{2\|\mathbf{p}_+ - \mathbf{p}_-\|_2^2} \quad (7)$$

where  $\mathbf{p}_+$  and  $\mathbf{p}_-$  are defined in the same way as in RelSim Eq. (6). Note that Eq. (7) is not normalized and depends on the prototypes and their distances to the sample  $\mathbf{x}$ . It is worthwhile mentioning that Eq. (7) is closely related to the reject option of a SVM [32] – in case of binary classification problem and a single prototype per class in a LVQ model, both models determine a separating hyperplane.

*Probabilistic certainty:* The third certainty measure is a probabilistic one. The idea is to obtain proper class probabilities  $p(y | \mathbf{x})$  for each class  $y \in \mathcal{Y}$  and reject a sample  $\mathbf{x}$  if the probability for the most probable class is lower than a given threshold  $\theta$ . We denote the probabilistic certainty measure as follows:

$$r_{\text{Proba}}(\mathbf{x}) = \max_{y \in \mathcal{Y}} p(y | \mathbf{x}). \quad (8)$$

In the following, we use the fastest method from [7] for computing class probabilities given a trained LVQ models. The method [33] combines estimates from binary classifiers in order to obtain class probabilities for the final prediction. First, because the estimates of a binary classifiers are necessary, we train a single LVQ model for each pair of classes using only the samples belonging to those two classes. This yields a set of  $|\mathcal{Y}|(|\mathcal{Y}| - 1)/2$  binary classifiers. Next, a data-dependent scaling of the predictions follows to yield pairwise probabilities. Here, we are using RelSim Eq. (6) and fit a sigmoid function to the real-valued scores. This mimics the approach by Platt [32] which has become very popular in the context of SVMs. This post-processing yields estimates  $r_{i,j}(\mathbf{x})$  of pairwise probabilities for every sample  $\mathbf{x}$  and pairs of classes  $i$  and  $j$ :

$$r_{i,j} = p(y = i | y = i \vee j, \mathbf{x}) \quad (9)$$

Given those pairwise probabilities and assuming a symmetric completion of the pairs  $r_{i,j} + r_{j,i} = 1$ , the posterior probabilities are obtained as follows:

$$p(y = i | \mathbf{x}) = \frac{1}{\sum_{j \neq i} \frac{1}{r_{i,j}} - (|\mathcal{Y}| - 2)} \quad (10)$$

where  $i, j \in \mathcal{Y}$  [33]. After computing all probabilities, a normalization step is required such that  $\sum_{i=1}^k p(y = i | \mathbf{x}) = 1$ . We refer to [7] for further details on all these steps.

### 2.3 Counterfactual Explanations

Counterfactual explanations (often just called *counterfactuals*) are a prominent instance of contrasting explanations, which state a change to some features of a given input such that the resulting data point, called the counterfactual, causes a different behavior of the system than the original input does. Thus, one can think of a counterfactual explanation as a suggestion of actions that change the model’s behavior/prediction. One reason why counterfactual explanations are so popular is that there exists evidence that explanations used by humans are often contrasting in nature [10] – i.e. people often ask questions like “*What would have to be different in order to observe a different outcome?*”. For illustrative purposes, consider the example of loan application: *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 had earned 500\$ more per month and if you would not have a second credit card.* Despite their popularity, the missing uniqueness of counterfactuals could pose a problem: Often there exist more than one possible & valid counterfactual – this is called the Rashomon effect [25] – and in such cases, it is not clear which or how many of them should be presented to the user. One common modeling approach (if this problem is not simply ignored) is to enforce uniqueness by a suitable formalization.

In order to keep the explanation (suggested changes) simple – i.e. easy to understand – an obvious strategy is to look for a small number of changes so that the resulting sample (counterfactual) is similar/close to the original sample, which is aimed to be captured by Definition 1.

**Definition 1 ((Closest) Counterfactual Explanation [45]).** Assume a prediction function  $h : \mathbb{R}^d \rightarrow \mathcal{Y}$  is given. Computing a counterfactual  $\mathbf{x}_{cf} \in \mathbb{R}^d$  for a given input  $\mathbf{x}_{orig} \in \mathbb{R}^d$  is phrased as an optimization problem:

$$\arg \min_{\mathbf{x}_{cf} \in \mathbb{R}^d} \ell(h(\mathbf{x}_{cf}), y') + C \cdot \theta(\mathbf{x}_{cf}, \mathbf{x}_{orig}) \quad (11)$$

where  $\ell(\cdot)$  denotes a loss function,  $y'$  the target prediction,  $\theta(\cdot)$  a penalty for dissimilarity of  $\mathbf{x}_{cf}$  and  $\mathbf{x}_{orig}$ , and  $C > 0$  denotes the regularization strength.

The counterfactuals from Definition 1 are also called *closest counterfactuals* because the optimization problem Eq. (11) tries to find an explanation  $\mathbf{x}_{cf}$  that is as close as possible to the original sample  $\mathbf{x}_{orig}$ . However, other aspects like plausibility and actionability are ignored in Definition 1, but are covered in other work [2, 3, 23]. Note that it is not always clear which type of counterfactual is meant when people talk about counterfactuals – in this work, we use the term counterfactuals in the spirit of Definition 1.

## 3 Counterfactual Explanations of Reject

In this section, we elaborate our proposal of using counterfactual explanations for explaining LVQ reject options. First, we introduce the general modeling in

Section 3.1, before then investigating the computational aspects of each reject option in Section 3.2.

### 3.1 Modeling

Because counterfactual explanations (see Section 2.3) proved to be an effective and useful explanation, we propose to use counterfactual explanations for explaining reject options of LVQ models<sup>1</sup> (see Section 2.1). Therefore, a counterfactual explanation of a reject provides the user with actionable feedback of what to change in order to be not rejected. Furthermore, such an explanation also communicates why the model is too uncertain for making a prediction in this particular case.

Since there exist evidence that people prefer low complexity (i.e. “simple”) explanations, we are looking for sparse counterfactuals. Similar to Definition 1, we phrase a counterfactual explanation  $\mathbf{x}_{\text{cf}}$  of a given input  $\mathbf{x}_{\text{orig}}$  as the following optimization problem:

$$\begin{aligned} \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \quad & \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\ \text{s.t.} \quad & r(\mathbf{x}_{\text{cf}}) \geq \theta \end{aligned} \tag{12}$$

where  $r(\cdot)$  denotes the specific reject option and the  $l_1$ -norm objective is supposed to yield a sparse and hence a “low complexity explanation”.

### 3.2 Computational Aspects of LVQ Reject Options

In the following, we propose algorithms which solve convex optimizations problems, for efficiently computing counterfactual explanations of LVQ rejects – whereby we consider each of the three reject options from Section 2.2 separately. For the purpose of readability, we moved all proofs and derivations to Appendix A.1.

**Relative Similarity** In case of the relative similarity reject option Eq. (6), the optimization problem Eq. (12) can be solved by using a divide & conquer approach, where we have to solve a bunch of convex quadratic programs of the following form:

$$\begin{aligned} \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \quad & \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\ \text{s.t.} \quad & \mathbf{x}_{\text{cf}}^\top \boldsymbol{\Omega} \mathbf{x}_{\text{cf}} + \mathbf{q}_j^\top \mathbf{x}_{\text{cf}} + c_j \leq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+) \end{aligned} \tag{13}$$

where  $\mathcal{P}(o_+)$  denotes a set of prototypes (see Appendix A.1 for details). Note that convex quadratic programs can be solved efficiently [5].

The complete algorithm for computing a counterfactual explanation is given in Algorithm 1.

---

<sup>1</sup> Although we focus on reject options for LVQ models, our proposed modeling is applicable to other models and reject options as well.

**Algorithm 1** Counterfactual under the relative similarity reject option**Input:** Original input  $\mathbf{x}_{\text{orig}}$ , reject threshold  $\theta$ , the LVQ model**Output:** Counterfactual  $\mathbf{x}_{\text{cf}}$ 


---

```

1:  $\mathbf{x}_{\text{cf}} = \mathbf{0}$  ▷ Initialize dummy solution
2:  $z = \infty$  ▷ Sparsity of the best solution so far
3: for  $\mathbf{p}_+ \in \mathcal{P}$  do ▷ Loop over every possible prototype
4:   Solving Eq. (13) yields a counterfactual  $\mathbf{x}_{\text{cf}*}$ 
5:   if  $\|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1 < z$  then ▷ Keep this counterfactual if it is sparser
     than the currently “best” counterfactual
6:      $z = \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1$ 
7:      $\mathbf{x}_{\text{cf}} = \mathbf{x}_{\text{cf}*}$ 
8:   end if
9: end for

```

---

**Distance to Decision Boundary** Similar to the relative similarity reject option, we again use a divide & conquer approach for solving Eq. (12). But in contrast to the relative similarity reject option, we have to solve a bunch of linear programs only, which can be solved even faster than convex quadratic programs [5]:

$$\begin{aligned}
& \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
& \text{s.t. } \mathbf{q}_j^\top \mathbf{x}_{\text{cf}} + c_j \geq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+)
\end{aligned} \tag{14}$$

where  $\mathcal{P}(o_+)$  denotes a set of prototypes (see Appendix A.1 for details).

The final algorithm for computing a counterfactual explanation is equivalent to Algorithm 1 for the relative similarity reject option, except that instead of solving Eq. (13) in line 4, we have to solve Eq. (14).

**Probabilistic Certainty Measure** In case of the probabilistic certainty measure as a reject option Eq. (8), we have that:

$$\begin{aligned}
& r_{\text{Proba}}(\mathbf{x}) \geq \theta \\
& \Leftrightarrow \max_{y \in \mathcal{Y}} p(y \mid \mathbf{x}) \geq \theta \\
& \Leftrightarrow \exists i \in \mathcal{Y} : p(y = i \mid \mathbf{x}) \geq \theta
\end{aligned} \tag{15}$$

Applying the divide & conquer paradigm over  $i \in \mathcal{Y}$  for solving Eq. (12), yields optimization problems of the following form:

$$\begin{aligned}
& \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
& \text{s.t. } \sum_{j \neq i} \exp \left( \alpha \frac{d_{i,j}(\mathbf{x}_{\text{cf}}, \mathbf{p}_-) - d_{i,j}(\mathbf{x}_{\text{cf}}, \mathbf{p}_+)}{d_{i,j}(\mathbf{x}_{\text{cf}}, \mathbf{p}_-) + d_{i,j}(\mathbf{x}_{\text{cf}}, \mathbf{p}_+)} + \beta \right) + 1 - \frac{1}{\theta} \leq 0
\end{aligned} \tag{16}$$



While we could solve Eq. (16) directly – yielding Algorithm 2 –, it is a rather difficult optimization problem because of its lack of any structure like convexity – e.g. general (black-box) solvers might be the only applicable choice. We therefore, additionally, propose a convex approximation where we can still guarantee feasibility (i.e. validity of the final counterfactual) at the price of losing closeness – i.e. we might not find the sparsest possible counterfactual, although finding a global optimum of Eq. (16) might be difficult as well. Approximating

---

**Algorithm 2** Counterfactual under the probabilistic certainty reject option

---

**Input:** Original input  $\mathbf{x}_{\text{orig}}$ , reject threshold  $\theta$ , the LVQ model

**Output:** Counterfactual  $\mathbf{x}_{\text{cf}}$

```

1:  $\mathbf{x}_{\text{cf}} = \mathbf{0}$  ▷ Initialize dummy solution
2:  $z = \infty$  ▷ Sparsity of the best solution so far
3: for  $i \in \mathcal{Y}$  do ▷ Loop over every possible class
4:   Solving Eq. (16) yields a counterfactual  $\mathbf{x}_{\text{cf}*}$ 
5:   if  $\|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1 < z$  then ▷ Keep this counterfactual if it is sparser
     than the currently “best” counterfactual
6:      $z = \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1$ 
7:      $\mathbf{x}_{\text{cf}} = \mathbf{x}_{\text{cf}*}$ 
8:   end if
9: end for

```

---

the constraint in Eq. (16) yields a convex quadratic constraint, which then results in a convex quadratic program as a final approximation of Eq. (16) – for details see Appendix A.1:

$$\begin{aligned}
& \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
& \text{s.t. } \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{q}_j^\top \mathbf{x} + c_j \leq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_i)
\end{aligned} \tag{17}$$

Using the approximation Eq. (17) requires us to iterate over every possible prototype, every possible class different from the  $i$ -th class, and finally over every possible class – i.e. finally yielding Algorithm 3. Although our proposed approximation Eq. (17) can be computed quite fast (because it is a convex quadratic program), it comes at the price of “drastically” increasing the complexity of the final divide & conquer algorithm. While we have to solve  $|\mathcal{Y}|$  optimization problems Eq. (16) in Algorithm 2, we get a quadratic complexity (quadratic in the number of classes) for our proposed convex approximation (Algorithm 3):

$$|\mathcal{Y}|^2 \cdot P_N - |\mathcal{Y}| \cdot P_N \tag{18}$$

where  $P_N$  denotes the number of prototypes per class used in the pair-wise classifiers. This quadratic complexity could become a problem in case of a large number of classes and a large number of prototypes per class.

---

**Algorithm 3** Counterfactual under the probabilistic certainty reject option – Approximation

---

**Input:** Original input  $\mathbf{x}_{\text{orig}}$ , reject threshold  $\theta$ , the LVQ model

**Output:** Counterfactual  $\mathbf{x}_{\text{cf}}$

---

```

1:  $\mathbf{x}_{\text{cf}} = \mathbf{0}$  ▷ Initialize dummy solution
2:  $z = \infty$  ▷ Sparsity of the best solution so far
3: for  $i \in \mathcal{Y}$  do ▷ Loop over every possible class
4:   for  $j \in \mathcal{Y} \setminus \{i\}$  do ▷ Loop over all other classes
5:     for  $\mathbf{p}_i \in \mathcal{P}$  do ▷ Loop over every possible prototype
6:       Solving Eq. (17) yields a counterfactual  $\mathbf{x}_{\text{cf}*}$ 
7:       if  $\|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1 < z$  then ▷ Keep this counterfactual if it is
         sparser than the currently “best” counterfactual
8:          $z = \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}*}\|_1$ 
9:          $\mathbf{x}_{\text{cf}} = \mathbf{x}_{\text{cf}*}$ 
10:      end if
11:    end for
12:  end for
13: end for

```

---

To summarize, we propose two divide & conquer algorithms for computing counterfactual explanations of the probabilistic certainty reject option Eq. (8). In Algorithm 2, we have to solve a rather complicated (i.e. unstructured) optimization problem, but we have to do this only a few times, whereas in Algorithm 3 we have to solve many convex quadratic programs. Although we have to solve many more optimization problems in Algorithm 3 than in Algorithm 2, solving the optimization problems in Algorithm 3 is much easier and it is also possible to easily extend the optimization problems with additional constraints, such as plausibility constraints as proposed in [2]. We think that in general, both algorithms have their areas of application and that practitioners should choose between them depending on their needs and the specific scenario.

## 4 Experiments

We empirically evaluate all our proposed algorithms for computing counterfactual explanations of different reject options (see Sections 2.1, 2.2) for GMLVQ on several different data sets. In our empirical evaluation, we consider two different properties:

- We evaluate algorithmic properties like sparsity, validity (in case of black-box solvers) and overlap (degree of agreement) between counterfactuals computed in different ways.
- We also evaluate whether counterfactual explanations of reject (as proposed in this work) are able to find and explain known ground truth reasons for rejects – i.e. evaluating the goodness of the explanations themselves.

All experiments are implemented in Python and the source code is available on GitHub<sup>2</sup>.

#### 4.1 Data Sets

We consider the following data sets for our empirical evaluation – all data sets are scaled and standardized except when using Algorithm 3<sup>3</sup>:

**Wine** The “Wine data set” [34] is used for predicting the cultivator of given wine samples based on their chemical properties. The data set contains 178 samples and 13 numerical features such as alcohol, hue and color intensity.

**Breast cancer** The “Breast Cancer Wisconsin (Diagnostic) Data Set” [46] is used for classifying breast cancer samples into benign and malignant. The data set contains 569 samples and 30 numerical features such as area, smoothness and compactness.

**Flip** This data set is used for the prediction of fibrosis. The set consists of samples of 118 patients and 12 numerical features such as blood glucose, BMI and total cholesterol, and was provided by the Department of Gastroenterology, Hepatology and Infectiology of the University Magdeburg [41]. As the data set contains some rows with missing values, we chose to replace these missing values with the corresponding feature mean.

**T21** This data set is used for early diagnosis of chromosomal abnormalities, such as trisomy 21, in pregnant women. The data set consists of 18 numerical features such as heart rate and weight, and contains over 50000 samples but only 0.8 percent abnormal samples (e.g. cases of trisomy 21) – i.e. it is highly imbalanced. It was collected by the Fetal Medicine Centre at King’s College Hospital and University College London Hospital in London [27].

#### 4.2 Setup

For finding the best model parameters we perform a grid search on the GMLVQ hyperparameters and reject option model’s rejection thresholds. In order to reduce the risk of overfitting, we cross validate the GMLVQ models’ accuracy and rejection rates on the different reject thresholds using a 5-fold cross validation.

For each GMLVQ model hyperparameterization, the rejection rates for each of the thresholds is computed as well as the impact of this on the accuracy of

<sup>2</sup> [https://github.com/andreArtelt/explaining\\_lvq\\_reject](https://github.com/andreArtelt/explaining_lvq_reject)

<sup>3</sup> In all cases, except for Algorithm 3, standardizing the data is necessary to get the methods working (i.e. avoiding numerical problems) – however, for Algorithm 3 the opposite is true: without standardizing the method works fine but if the data is standardized, the method does not work at all – i.e. we observe numerical problems.

the model. Based on these rejection rates and accuracies, the accuracy-rejection curve (ARC) [26] can be computed. The area under the ARC measure (AU-ARC) gives an indication of how well the reject model performs given the GMLVQ model type and its hyperparameterization. For each combination of data set, GMLVQ model type, and reject option method, we can then determine the best GMLVQ model parameters and rejection threshold. We do this by selecting the GMLVQ hyperparameters with the highest accuracy. Another alternative method would be the inclusion of the AU-ARC as an indication of how well the model performs. Furthermore, we could combine these two measures by using some type of weighting strategy.

Following the selection of the best GMLVQ hyperparameters, the best rejection threshold is found by determining the “optimal” threshold in the ARC by finding the so-called knee-point using the Kneedle algorithm [38]. In reality, the truly optimal threshold depends on the data set and the real-world application in which it is used. However, for the purposes of our evaluation, we are primarily interested in those GMLVQ hyperparameters and reject option thresholds that allow us to compare the overall performance and impact of the different methods for computing counterfactuals. Once the optimal hyperparameters have been obtained, the counterfactuals of the different reject options using different algorithms, can be computed, compared, and evaluated. By finding the optimal hyperparameters for each model, our evaluation of the different model types will be less dependent on potentially poor hyperparameterization.

We evaluate two properties separately: Algorithmic properties like sparsity and goodness of the explanations. We run the experiments (5-fold cross validation) for each data set, each reject option and each method for computing the counterfactuals – i.e. black-box solver for solving Eq. (12)<sup>4</sup>, our proposed algorithms from Section 3) and the closest sample from the training data set which is not rejected (this “naive” way of computing a counterfactual serves as a baseline).

**Algorithmic Properties** We evaluate the sparsity of the counterfactual explanations by using the  $l_0$ -norm<sup>5</sup>. Since our methods and algorithms are guaranteed to output feasible (i.e. valid) counterfactuals, we only evaluate validity of the counterfactuals computed by the black-box solver. Furthermore, we also compare the pairwise overlap/agreement of the counterfactuals computed by the different methods – i.e. how large is the overlap in the selected (non-zero) features of the counterfactuals computed by two different methods.

**Goodness of Counterfactual Explanations** For evaluating the goodness of the counterfactuals, we create scenarios with ground truth as follows: For each

<sup>4</sup> We use the penalty method (with equal weighting) to turn Eq. (12) into an unconstrained problem, solve it by using the Nelder-Mead black-box method.

<sup>5</sup> For numerical reasons, we use a threshold at which we consider a floating point number to be equal to zero – see provided source code for details.

data set, we select a random subset of features (30%) and perturb these in the test set by adding Gaussian noise – we then check which of these samples are rejected due to the noise (i.e. applying the reject option before and after applying the perturbation), and compute counterfactual explanations of these samples only. We then evaluate for each counterfactual how many of the relevant features (from the known ground truth) are detected and included in the counterfactual explanation.

### 4.3 Results & Discussion

When reporting the results, we use the following abbreviations: *BbCfFeasibility* – Feasibility of the counterfactuals computed by the black-box solver, in case of the probabilistic reject option Eq. (8), we report the results of using Algorithm 2 (the results for the “true” black-box solver can be found in Appendix B); *BbCf* – Counterfactuals computed by the black-box solver, *TrainCf* – Counterfactuals by selecting the closest sample from the training set which is not rejected; *ClosestCf* – Counterfactuals computed by our proposed algorithms.

Note that we round all values to two decimal points.

**Algorithmic Properties** In Table 1, we report the mean sparsity (along with the variance) of the counterfactuals for different reject options and different data sets. The results (mean and variance) for evaluating the number of overlapping features are reported in Table 2. We observe that our proposed methods for

Table 1: Algorithmic properties – Mean sparsity (incl. variance) of different counterfactuals, smaller values are better.

	<i>DataSet</i>	<i>BbCfFeasibility</i>	<i>BbCf</i>	<i>TrainCf</i>	<i>ClosestCf</i>
Relative Similarity Eq. (6)	Wine	$1.0 \pm 0.0$	$11.14 \pm 1.41$	$13.0 \pm 0.0$	$5.68 \pm 11.36$
	Breast Cancer	$0.99 \pm 0.0$	$27.64 \pm 4.38$	$30.0 \pm 0.0$	$12.69 \pm 52.31$
	t21	$0.99 \pm 0.0$	$16.14 \pm 1.91$	$11.0 \pm 1.0$	$4.39 \pm 12.79$
	Flip	$1.0 \pm 0.0$	$10.52 \pm 1.38$	$12.0 \pm 0.0$	$4.07 \pm 8.02$
Distance DecisionBoundary Eq. (7)	Wine	$1.0 \pm 0.0$	$10.98 \pm 1.12$	$13.0 \pm 0.0$	$2.8 \pm 9.13$
	Breast Cancer	$1.0 \pm 0.0$	$27.5 \pm 2.78$	$30.0 \pm 0.0$	$4.16 \pm 48.19$
	t21	$1.0 \pm 0.0$	$16.02 \pm 1.43$	$11.0 \pm 1.0$	$2.12 \pm 5.19$
	Flip	$1.0 \pm 0.0$	$10.31 \pm 1.2$	$12.0 \pm 0.0$	$1.88 \pm 3.11$
Probabilistic Certainty Eq. (8)	Wine	$0.45 \pm 0.21$	$13.0 \pm 0.0$	$13.0 \pm 0.0$	$12.67 \pm 0.22$
	Breast Cancer	$0.94 \pm 0.02$	$30.0 \pm 0.0$	$30.0 \pm 0.0$	$26.04 \pm 60.5$
	t21	$0.4 \pm 0.24$	$17.39 \pm 0.64$	$11.0 \pm 0.0$	$12.55 \pm 7.41$
	Flip	$0.4 \pm 0.24$	$11.75 \pm 0.23$	$12.0 \pm 0.0$	$11.5 \pm 1.47$

computing counterfactual explanation of reject options is consistently able to compute very sparse (i.e. low complexity) counterfactuals – however, the variance is often very large which suggests that there exist a few outliers in the data set

Table 2: Algorithmic properties – Mean overlap (incl. variance) between different counterfactuals.

	<i>DataSet</i>	BbCfFeasibility	BbCf	TrainCf	ClosestCf
Relative Similarity Eq. (6)	Wine	$1.0 \pm 0.0$	$4.68 \pm 7.26$	$11.16 \pm 1.45$	$5.27 \pm 9.02$
	Breast Cancer	$0.99 \pm 0.0$	$10.86 \pm 54.34$	$27.34 \pm 5.18$	$11.37 \pm 59.55$
	t21	$0.99 \pm 0.0$	$3.92 \pm 10.5$	$9.61 \pm 2.14$	$3.52 \pm 9.42$
	Flip	$1.0 \pm 0.0$	$3.8 \pm 6.94$	$10.11 \pm 1.55$	$4.19 \pm 8.73$
Distance DecisionBoundary Eq. (7)	Wine	$1.0 \pm 0.0$	$2.18 \pm 5.14$	$11.08 \pm 1.13$	$2.39 \pm 7.21$
	Breast Cancer	$1.0 \pm 0.0$	$4.66 \pm 51.26$	$27.83 \pm 3.13$	$4.85 \pm 56.74$
	t21	$1.0 \pm 0.0$	$1.67 \pm 1.75$	$10.0 \pm 1.87$	$1.53 \pm 1.39$
	Flip	$1.0 \pm 0.0$	$1.59 \pm 1.17$	$9.95 \pm 1.57$	$1.68 \pm 1.35$
Probabilistic Certainty Eq. (8)	Wine	$0.45 \pm 0.21$	$11.09 \pm 2.41$	$11.0 \pm 2.1$	$12.83 \pm 0.14$
	Breast Cancer	$0.94 \pm 0.02$	$24.75 \pm 8.77$	$22.0 \pm 1.43$	$29.42 \pm 0.93$
	t21	$0.4 \pm 0.24$	$9.26 \pm 1.8$	$8.47 \pm 2.23$	$11.54 \pm 9.55$
	Flip	$0.4 \pm 0.24$	$11.71 \pm 0.29$	$11.69 \pm 0.34$	$12.0 \pm 0.0$

for which it is not possible to compute a sparse counterfactuals. As it was to be expected, we observe the worst performance when choosing a sample from the training set as a counterfactual. The counterfactuals computed by a black-box solver are often a bit better than those from the training set but still far away from the counterfactuals computed by our proposed algorithms. While the black-box solver works quite well in case of the relative similarity and distance to decision boundary reject options, the performance drops significantly (for many but not all data sets) in case of the probabilistic certainty reject option. We think this might be due to the increased complexity of the reject option, compared to the other two reject options which have much simpler mathematical form. For this reject option, our proposed algorithm is still able to consistently yield the sparsest counterfactuals but the difference to other counterfactuals is not that significant like it is the case for the other two reject options.

Concerning the feature overlap of the different counterfactuals, we observe that usually all methods agree upon the same counterfactual features.

**Goodness of Counterfactual Explanations** The mean recall (along with the variance) of recovered (ground truth) relevant features for different counterfactuals, data sets and reject options, is given in Table 3. We observe that all methods are able to identify the relevant features that caused the reject. However, the counterfactuals computed by our proposed algorithms sometimes miss a few relevant features which is most likely due to the fact that these methods are designed to find the sparsest possible counterfactual – this indicates that sparsity alone is not always sufficient for coming up with the “best” explanation.

Table 3: Goodness of counterfactual explanations – Mean and variance recall of identified relevant features (larger numbers are better).

	<i>DataSet</i>	BbCfFeasible	BbCf	TrainCf	Cf
Relative Similarity Eq. (6)	Wine	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$
	Breast Cancer	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.99 \pm 0.01$
	t21	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.98 \pm 0.0$
	Flip	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.96 \pm 0.02$
Distance DecisionBoundary Eq. (7)	Wine	$0.8 \pm 0.16$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.72 \pm 0.15$
	Breast Cancer	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.45 \pm 0.18$
	t21	$0.6 \pm 0.24$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.63 \pm 0.15$
	Flip	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.62 \pm 0.22$
Probabilistic Certainty Eq. (8)	Wine	$0.8 \pm 0.16$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$
	Breast Cancer	$0.8 \pm 0.16$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$1.0 \pm 0.0$
	t21	$0.85 \pm 0.02$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.94 \pm 0.01$
	Flip	$0.73 \pm 0.06$	$1.0 \pm 0.0$	$1.0 \pm 0.0$	$0.75 \pm 0.19$

## 5 Summary & Conclusion

In this work we proposed to explain reject options of LVQ models by means of counterfactual explanations. We considered three popular reject options and for each, proposed (extendable) modelings and algorithms for efficiently computing counterfactual explanations under the particular reject option. We empirically evaluated all our proposed methods under different aspects – in particular, we demonstrated that our algorithms delivers sparse (i.e. “low-complexity”) explanations, and that counterfactual explanations in general seem to be able to detect and highlight relevant features in scenarios where the ground truth is known.

Although our proposed idea of using counterfactual explanations for explaining rejects is rather general, our proposed methods and algorithms are completely tailored towards LVQ models and thus not applicable to other ML models. Therefore, it would be of interest to see other, either model specific or more general methods for computing counterfactual explanations of reject under other ML models.

Our evaluation focused on algorithmic properties such as sparsity and feature relevances for assessing the goodness of the computed counterfactuals. However, it is still unclear how and if these kinds of explanations of reject are useful and helpful to humans – since it is difficult to implement “human usefulness” as a scoring function, a proper user study for evaluating the usefulness is necessary.

We leave these aspects as future work.

## References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and systemapproaches. AI communications (1994)

2. Artelt, A., Hammer, B.: Convex density constraints for computing plausible counterfactual explanations. 29th International Conference on Artificial Neural Networks (ICANN) (2020)
3. Artelt, A., Hammer, B.: Convex optimization for actionable & plausible counterfactual explanations. CoRR **abs/2105.07630** (2021), <https://arxiv.org/abs/2105.07630>
4. Biehl, M., Hammer, B., Villmann, T.: Prototype-based models in machine learning. Wiley interdisciplinary reviews. Cognitive science **7**(2), 92–111 (2016). <https://doi.org/10.1002/wcs.1378>
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, NY, USA (2004)
6. Brinkrolf, J., Hammer, B.: Interpretable machine learning with reject option. Automatisierungstechnik **66**(4), 283–290 (2018). <https://doi.org/10.1515/auto-2017-0123>
7. Brinkrolf, J., Hammer, B.: Time integration and reject options for probabilistic output of pairwise lvq. Neural Computing and Applications pp. 1–14 (2019). <https://doi.org/10.1007/s00521-018-03966-0>, <https://doi.org/10.1007/s00521-018-03966-0>
8. Brinkrolf, J., Hammer, B.: Time integration and reject options for probabilistic output of pairwise LVQ. Neural Comput. Appl. **32**(24), 18009–18022 (2020). <https://doi.org/10.1007/s00521-018-03966-0>, <https://doi.org/10.1007/s00521-018-03966-0>
9. Brinkrolf, J., Hammer, B.: Federated learning vector quantization. In: Verleysen, M. (ed.) Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2021). i6doc.com (2021)
10. Byrne, R.M.J.: Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In: IJCAI-19 (2019)
11. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Trans. Inf. Theor. **16**(1), 4146 (sep 1970). <https://doi.org/10.1109/TIT.1970.1054406>, <https://doi.org/10.1109/TIT.1970.1054406>
12. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
13. Fischer, L., Hammer, B., Wersing, H.: Efficient rejection strategies for prototype-based classification. Neurocomputing **169**, 334–342 (2015). <https://doi.org/10.1016/j.neucom.2014.10.092>, <https://doi.org/10.1016/j.neucom.2014.10.092>
14. Fischer, L., Hammer, B., Wersing, H.: Optimum reject options for prototype-based classification. CoRR **abs/1503.06549** (2015), <http://arxiv.org/abs/1503.06549>
15. Fisher, A., Rudin, C., Dominici, F.: 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 arXiv:1801.01489 (Jan 2018)
16. Gepperth, A., Hammer, B.: Incremental learning algorithms and applications. In: 24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016 (2016), <http://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-19.pdf>
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (Aug 2018)
18. Herbei, R., Wegkamp, M.H.: Classification with reject option. Canadian Journal of Statistics **34**(4), 709–721 (2006).



- <https://doi.org/https://doi.org/10.1002/cjs.5550340410>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.5550340410>
19. Khandani, A.E., Kim, A.J., Lo, A.: Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* **34**(11) (2010)
  20. Kim, B., Koyejo, O., Khanna, R.: Examples are not enough, learn to criticize! criticism for interpretability. In: *Advances in Neural Information Processing Systems* 29 (2016)
  21. Kirstein, S., Wersing, H., Gross, H.M., Körner, E.: A life-long learning vector quantization approach for interactive learning of multiple categories. *Neural networks : the official journal of the International Neural Network Society* **28**, 90–105 (04 2012). <https://doi.org/10.1016/j.neunet.2011.12.003>
  22. Kohonen, T.: *Self-Organizing Maps*, Second Edition, Springer Series in Information Sciences, vol. 30. Springer (1997). <https://doi.org/10.1007/978-3-642-97966-8>, <https://doi.org/10.1007/978-3-642-97966-8>
  23. Looveren, A.V., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. *CoRR* **abs/1907.02584** (2019), <http://arxiv.org/abs/1907.02584>
  24. Losing, V., Hammer, B., Wersing, H.: Incremental on-line learning: A review and comparison of state of the art algorithms. *Neurocomputing* **275**, 1261–1274 (2018). <https://doi.org/10.1016/j.neucom.2017.06.084>, <https://doi.org/10.1016/j.neucom.2017.06.084>
  25. Molnar, C.: *Interpretable Machine Learning* (2019)
  26. Nadeem, M.S.A., Zucker, J.D., Hanczar, B.: Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In: *Machine Learning in Systems Biology*. pp. 65–81. PMLR (2009)
  27. Nicolaidis, K.H., Spencer, K., Avgidou, K., Faiola, S., Falcon, O.: Multicenter study of first-trimester screening for trisomy 21 in 75 821 pregnancies: results and estimation of the potential impact of individual risk-orientated two-stage first-trimester screening. *Ultrasound in Obstetrics & Gynecology* **25**(3), 221–226 (2005). <https://doi.org/https://doi.org/10.1002/uog.1860>, <https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.1860>
  28. Nova, D., Estévez, P.A.: A review of learning vector quantization classifiers. *Neural Comput. Appl.* **25**(3-4), 511–524 (Sep 2014). <https://doi.org/10.1007/s00521-013-1535-3>
  29. Offert, F.: ”i know it when i see it”. visualization and intuitive interpretability (2017)
  30. Owomugisha, G., Nuwamanya, E., Quinn, J.A., Biehl, M., Mwebaze, E.: Early detection of plant diseases using spectral data. In: Petkov, N., Strisciuglio, N., Travieso-Gonzalez, C.M. (eds.) *APPIS 2020: 3rd International Conference on Applications of Intelligent Systems, APPIS 2020, Las Palmas de Gran Canaria Spain, 7-9 January 2020*. ACM (2020). <https://doi.org/10.1145/3378184.3378222>, <https://doi.org/10.1145/3378184.3378222>
  31. parliament, E., council: General data protection regulation: Regulation (eu) 2016/679 of the european parliament (2016)
  32. Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. pp. 61–74. MIT Press (1999)
  33. Price, D., Knerr, S., Personnaz, L., Dreyfus, G.: Pairwise Neural Network Classifiers with Probabilistic Outputs. In: *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*. pp. 1109–1116 (1994), <http://papers.nips.cc/paper/883-pairwise-neural-network-classifiers-with-probabilistic-outputs>

34. S. Aeberhard, D.C., de Vel, O.: Comparison of classifiers in high dimensional settings. Tech. Rep. no. 92-02 (1992)
35. Sallab, A.E., Abdou, M., Perot, E., Yogamani, S.: Deep reinforcement learning framework for autonomous driving. *Electronic Imaging* **2017**(19), 70–76 (2017)
36. Samek, W., Wiegand, T., Müller, K.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR* **abs/1708.08296** (2017)
37. Sato, A., Yamada, K.: Generalized Learning Vector Quantization. In: *Advances in Neural Information Processing Systems 8*, NIPS, Denver, CO, November 27–30, 1995. pp. 423–429 (1995), <http://papers.nips.cc/paper/1113-generalized-learning-vector-quantization>
38. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: *2011 31st International Conference on Distributed Computing Systems Workshops*. pp. 166–171 (2011). <https://doi.org/10.1109/ICDCSW.2011.20>
39. Schneider, P., Biehl, M., Hammer, B.: Adaptive Relevance Matrices in Learning Vector Quantization. *Neural Computation* **21**(12), 3532–3561 (2009). <https://doi.org/10.1162/neco.2009.11-08-908>, <http://dx.doi.org/10.1162/neco.2009.11-08-908>
40. Shafer, G., Vovk, V.: A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**, 371–421 (2008), <https://dl.acm.org/citation.cfm?id=1390693>
41. Sowa, J.P., Heider, D., Bechmann, L.P., Gerken, G., Hoffmann, D., Canbay, A.: Novel algorithm for non-invasive assessment of fibrosis in nafl. *PLOS ONE* **8**(4), 1–6 (04 2013). <https://doi.org/10.1371/journal.pone.0062439>, <https://doi.org/10.1371/journal.pone.0062439>
42. Stalidis, P., Semertzidis, T., Daras, P.: Examining deep learning architectures for crime classification and prediction **abs/1812.00602** (2018)
43. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR* **abs/1907.07374** (2019)
44. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review (2020)
45. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* **31**, 841 (2017)
46. William H. Wolberg, W. Nick Street, O.L.M.: Breast cancer wisconsin (diagnostic) data set. [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)) (1995)
47. Xu, Y., Furo, S., Hasegawa, O., Zhao, J.: An online incremental learning vector quantization. In: *Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27–30, 2009, Proceedings*. pp. 1046–1053 (2009). [https://doi.org/10.1007/978-3-642-01307-2\\_112](https://doi.org/10.1007/978-3-642-01307-2_112), [https://doi.org/10.1007/978-3-642-01307-2\\_112](https://doi.org/10.1007/978-3-642-01307-2_112)

## A Proofs & Derivations

### A.1 Counterfactual Explanations

**Relative Similarity** In order for a sample  $\mathbf{x} \in \mathbb{R}^d$  to be classified, it must hold that:

$$r_{\text{RelSim}}(\mathbf{x}) \geq \theta \quad (19)$$

Using the shorter notation  $d^+(\mathbf{x}) = d(\mathbf{x}, \mathbf{p}_+)$  and  $d^-(\mathbf{x}) = d(\mathbf{x}, \mathbf{p}_-)$ , respectively, where  $\mathbf{p}_+$  refers to the closest prototype and  $\mathbf{p}_-$  refers to the closest one belongs to a different class, this further translates into:

$$\begin{aligned}
r_{\text{RelSim}}(\mathbf{x}) &\geq \theta \\
&\Leftrightarrow \frac{d^-(\mathbf{x}) - d^+(\mathbf{x})}{d^-(\mathbf{x}) + d^+(\mathbf{x})} \geq \theta \\
&\Leftrightarrow d^-(\mathbf{x}) - d^+(\mathbf{x}) \geq (d^-(\mathbf{x}) + d^+(\mathbf{x})) \theta \\
&\Leftrightarrow d^-(\mathbf{x}) - d^-(\mathbf{x})\theta \geq d^+(\mathbf{x}) + d^+(\mathbf{x})\theta \\
&\Leftrightarrow (1 - \theta)d^-(\mathbf{x}) - (1 + \theta)d^+(\mathbf{x}) \geq 0
\end{aligned} \tag{20}$$

Assuming that the closest prototype  $\mathbf{p}_+$  is fixed, we can rewrite Eq. (20) as follows:

$$(1 - \theta)d(\mathbf{x}, \mathbf{p}_j) - (1 + \theta)d(\mathbf{x}, \mathbf{p}_+) \geq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+) \tag{21}$$

where  $\mathcal{P}(o_+)$  denotes the set of all prototypes that are not labeled as  $o_+$ . The intuition behind the translation of Eq. (20) into Eq. (21) is that, if Eq. (20) holds for all possible  $\mathbf{p}_-$ , then also for the particular choice of  $\mathbf{p}_-$  in Eq. (19).

These constraints can be rewritten as the following convex quadratic constraints – for a given  $\mathbf{p}_j$ :

$$\begin{aligned}
(1 - \theta)d(\mathbf{x}, \mathbf{p}_j) - (1 + \theta)d(\mathbf{x}, \mathbf{p}_+) &\geq 0 \\
&\Leftrightarrow d(\mathbf{x}, \mathbf{p}_j) - d(\mathbf{x}, \mathbf{p}_+) - \theta(d(\mathbf{x}, \mathbf{p}_j) + d(\mathbf{x}, \mathbf{p}_+)) \geq 0 \\
&\Leftrightarrow 2(\mathbf{p}_+^\top \Omega - \mathbf{p}_j^\top \Omega)\mathbf{x} + \mathbf{p}_j^\top \Omega \mathbf{p}_j - \mathbf{p}_+^\top \Omega \mathbf{p}_+ - \\
&\quad \theta(2\mathbf{x}^\top \Omega \mathbf{x} - 2(\mathbf{p}_j^\top \Omega + \mathbf{p}_+^\top \Omega)\mathbf{x} + \mathbf{p}_j^\top \Omega \mathbf{p}_j + \mathbf{p}_+^\top \Omega \mathbf{p}_+) \geq 0 \\
&\Leftrightarrow -2(\mathbf{p}_+^\top \Omega - \mathbf{p}_j^\top \Omega)\mathbf{x} - \mathbf{p}_j^\top \Omega \mathbf{p}_j + \mathbf{p}_+^\top \Omega \mathbf{p}_+ + \\
&\quad \theta(2\mathbf{x}^\top \Omega \mathbf{x} - 2(\mathbf{p}_j^\top \Omega + \mathbf{p}_+^\top \Omega)\mathbf{x} + \mathbf{p}_j^\top \Omega \mathbf{p}_j + \mathbf{p}_+^\top \Omega \mathbf{p}_+) \leq 0 \\
&\Leftrightarrow \mathbf{x}^\top \Omega \mathbf{x} + \mathbf{q}_j^\top \mathbf{x} + c_j \leq 0
\end{aligned} \tag{22}$$

where

$$\mathbf{q}_j^\top = \left(-\frac{1}{\theta} - 1\right) \mathbf{p}_+^\top \Omega + \left(\frac{1}{\theta} - 1\right) \mathbf{p}_j^\top \Omega \tag{23}$$

$$c_j = \frac{1}{2} \left( \left(1 - \frac{1}{\theta}\right) \mathbf{p}_j^\top \Omega \mathbf{p}_j + \left(1 + \frac{1}{\theta}\right) \mathbf{p}_+^\top \Omega \mathbf{p}_+ \right) \tag{24}$$

We therefore get the following convex quadratic optimization problem – note that convex quadratic programs can be solved efficiently [5]:

$$\begin{aligned}
&\min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
&\text{s.t. } \mathbf{x}_{\text{cf}}^\top \Omega \mathbf{x}_{\text{cf}} + \mathbf{q}_j^\top \mathbf{x}_{\text{cf}} + c_j \leq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+)
\end{aligned} \tag{25}$$

We simply solve this optimization problem Eq. (25) for all possible target prototypes and select the counterfactual which is the closest to the original sample – since every possible prototype could be a potential closest prototype  $\mathbf{p}_+$ , we have to solve as many optimization problems as we have prototypes. However, one could introduce some kind of early stopping by adding an additional constraint on the distance of the counterfactual to the original sample – i.e. use the currently best known solution as a upper bound on the objective, which might result in an infeasible program and hence will be aborted quickly. Alternatively, one could solve the different optimization problems in parallel because they are no dependencies between them.

**Distance to decision boundary** In order for a sample  $\mathbf{x} \in \mathbb{R}^d$  to be accepted (i.e. not being rejected), it must hold that:

$$r_{\text{Dist}}(\mathbf{x}) \geq \theta \quad (26)$$

This further translates into:

$$\begin{aligned} r_{\text{Dist}}(\mathbf{x}) &\geq \theta \\ \Leftrightarrow \frac{|\mathbf{d}^+(\mathbf{x}) - \mathbf{d}^-(\mathbf{x})|}{2\|\mathbf{p}_+ - \mathbf{p}_-\|_2^2} &\geq \theta \\ \Leftrightarrow |\mathbf{d}^+(\mathbf{x}) - \mathbf{d}^-(\mathbf{x})| &\geq 2\theta\|\mathbf{p}_+ - \mathbf{p}_-\|_2^2 \\ \Leftrightarrow \mathbf{d}^-(\mathbf{x}) - \mathbf{d}^+(\mathbf{x}) - 2\theta\|\mathbf{p}_+ - \mathbf{p}_-\|_2^2 &\geq 0 \end{aligned} \quad (27)$$

Assuming that the closest prototype  $\mathbf{p}_+$  is fixed, we get:

$$\mathbf{d}(\mathbf{x}, \mathbf{p}_j) - \mathbf{d}(\mathbf{x}, \mathbf{p}_+) - 2\theta\|\mathbf{p}_+ - \mathbf{p}_j\|_2^2 \geq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+) \quad (28)$$

where  $\mathcal{P}(o_+)$  denotes the set of prototypes that are not labeled as  $o_+$ . Again, the intuition behind this translation is that, if Eq. (27) holds for all possible  $\mathbf{p}_-$ , then also for the particular choice of  $\mathbf{p}_-$  in Eq. (26).

These constraints can be rewritten as the following linear constraints – for a given  $\mathbf{p}_j$ :

$$\begin{aligned} &\mathbf{d}(\mathbf{x}, \mathbf{p}_j) - \mathbf{d}(\mathbf{x}, \mathbf{p}_+) - 2\theta\|\mathbf{p}_+ - \mathbf{p}_j\|_2^2 \geq 0 \\ \Leftrightarrow &(\mathbf{x} - \mathbf{p}_j)^\top \boldsymbol{\Omega}(\mathbf{x} - \mathbf{p}_j) - (\mathbf{x} - \mathbf{p}_+)^\top \boldsymbol{\Omega}(\mathbf{x} - \mathbf{p}_+) - \\ &2\theta(\mathbf{p}_+ - \mathbf{p}_j)^\top \boldsymbol{\Omega}(\mathbf{p}_+ - \mathbf{p}_j) \geq 0 \\ \Leftrightarrow &\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{p}_j - \mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j - \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{p}_+ + \\ &\mathbf{p}_+^\top \boldsymbol{\Omega} \mathbf{x} - \mathbf{p}_+^\top \boldsymbol{\Omega} \mathbf{p}_+ - 2\theta(\mathbf{p}_+ - \mathbf{p}_j)^\top \boldsymbol{\Omega}(\mathbf{p}_+ - \mathbf{p}_j) \geq 0 \\ \Leftrightarrow &(2\mathbf{p}_+^\top \boldsymbol{\Omega} - 2\mathbf{p}_j^\top \boldsymbol{\Omega}) \mathbf{x} + \mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j - \mathbf{p}_+^\top \boldsymbol{\Omega} \mathbf{p}_+ - \\ &2\theta(\mathbf{p}_+ - \mathbf{p}_j)^\top \boldsymbol{\Omega}(\mathbf{p}_+ - \mathbf{p}_j) \geq 0 \\ \Leftrightarrow &\mathbf{q}_j^\top \mathbf{x} + c_j \geq 0 \end{aligned} \quad (29)$$

where

$$\mathbf{q}_j^\top = 2\mathbf{p}_+^\top \boldsymbol{\Omega} - 2\mathbf{p}_j^\top \boldsymbol{\Omega} \quad c_j = \mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j - \mathbf{p}_+^\top \boldsymbol{\Omega} \mathbf{p}_+ - 2\theta(\mathbf{p}_+ - \mathbf{p}_j)^\top \boldsymbol{\Omega}(\mathbf{p}_+ - \mathbf{p}_j) \quad (30)$$

Finally, we get the following linear optimization problem – note that linear programs can be solved even faster than convex quadratic programs [5]:

$$\begin{aligned} \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \quad & \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\ \text{s.t.} \quad & \mathbf{q}_j^\top \mathbf{x}_{\text{cf}} + c_j \geq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_+) \end{aligned} \quad (31)$$

Again, we try all possible target prototypes  $\mathbf{p}_+$  and select the best counterfactual – everything from the relative similarity case applies (see Section A.1).

**Probabilistic certainty measure** In order for a sample  $\mathbf{x} \in \mathbb{R}^d$  to be classified (i.e. not being rejected), it must hold that:

$$r_{\text{Proba}}(\mathbf{x}) \geq \theta \quad (32)$$

This further translates into:

$$\begin{aligned} r_{\text{Proba}}(\mathbf{x}) &\geq \theta \\ \Leftrightarrow \max_{y \in \mathcal{Y}} p(y \mid \mathbf{x}) &\geq \theta \\ \Leftrightarrow \exists i \in \mathcal{Y} : p(y = i \mid \mathbf{x}) &\geq \theta \end{aligned} \quad (33)$$

For the moment, we assume that  $i$  is fixed – i.e. using a divide & conquer approach. It follows that:

$$\begin{aligned} p(y = i \mid \mathbf{x}) &\geq \theta \\ \Leftrightarrow \frac{1}{\sum_{j \neq i} \frac{1}{p(y=i \mid (i, j), \mathbf{x})} - (|\mathcal{Y}| - 2)} &\geq \theta \\ \Leftrightarrow 1 &\geq \theta \left( \sum_{j \neq i} \frac{1}{p(y = i \mid (i, j), \mathbf{x})} - (|\mathcal{Y}| - 2) \right) \\ \Leftrightarrow \sum_{j \neq i} \frac{1}{p(y = i \mid (i, j), \mathbf{x})} - c &\leq 0 \end{aligned} \quad (34)$$

where

$$c = |\mathcal{Y}| - 2 + \frac{1}{\theta} \quad (35)$$

Further simplifications of Eq. (34) yield:

$$\begin{aligned}
& \sum_{j \neq i} \frac{1}{p(y = i \mid (i, j), \mathbf{x})} - c \leq 0 \\
& \Leftrightarrow \sum_{j \neq i} 1 + \exp(\alpha \cdot r_{i,j}(\mathbf{x}) + \beta) - c \leq 0 \\
& \Leftrightarrow \sum_{j \neq i} \exp(\alpha \cdot r_{i,j}(\mathbf{x}) + \beta) + c' \leq 0 \\
& \Leftrightarrow \sum_{j \neq i} \exp \left( \alpha \frac{d_{i,j}^-(\mathbf{x}_{\text{cf}}) - d_{i,j}^+(\mathbf{x}_{\text{cf}})}{d_{i,j}^-(\mathbf{x}_{\text{cf}}) + d_{i,j}^+(\mathbf{x}_{\text{cf}})} + \beta \right) + c' \leq 0
\end{aligned} \tag{36}$$

where

$$\begin{aligned}
c' &= |\mathcal{Y}| - 1 - c \\
&= |\mathcal{Y}| - 1 - |\mathcal{Y}| + 2 - \frac{1}{\theta} \\
&= 1 - \frac{1}{\theta}
\end{aligned} \tag{37}$$

We therefore get the following optimization problem:

$$\begin{aligned}
& \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
& \text{s.t. } \sum_{j \neq i} \exp \left( \alpha \frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} + \beta \right) + c' \leq 0
\end{aligned} \tag{38}$$

Because of the divide & conquer paradigm, we would have to try all possible target classes  $i \in \mathcal{Y}$  and finally select the one yielding the lowest objective – i.e. the closest counterfactual. While one could do this using constraint Eq. (36), the optimization would be rather complicated because the constraint is not convex and rather “ugly” – although it could be tackled by an evolutionary optimization method. However, the number of optimization problems we have to solve is rather small – it is equal to the number of classes.

We therefore, additionally, propose a surrogate constraint which captures the same “meaning/intuition” as Eq. (36) does, but is easier to optimize over – however, note that by using a surrogate instead of the original constraint Eq. (36) we give up closeness which, in our opinion, would be acceptable if the solutions stay somewhat close to each other<sup>6</sup>. We try out and compare both approaches in the experiments (see Section 4).

<sup>6</sup> Furthermore, in case of additional plausibility & actionability constraints, closeness becomes even less important.

First, we apply the natural logarithm to Eq. (36) and then bound it by using the maximum:

$$\begin{aligned} & \log \left( \sum_{j \neq i} \exp \left( \alpha \frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} + \beta \right) \right) \\ & \leq \max_{j \neq i} \left( \alpha \frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} + \beta \right) + \log(|\mathcal{Y}| - 1) \end{aligned} \quad (39)$$

We therefore approximate the constraint Eq. (36) by using Eq. (39), which yields the following constraint:

$$\max_{j \neq i} \left( \alpha \frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} + \beta \right) \leq \log(-c') - \log(|\mathcal{Y}| - 1) \quad (40)$$

Note that, in theory it could happen that  $-c' \leq 0$  – we fix this by simply taking  $\max(-c', \epsilon)$ , which results in a feasible solution but the approximation gets a bit worse.

Assuming that the maximum  $j$  is fixed, we get the following constraint:

$$\frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} \leq \gamma \quad (41)$$

where

$$\gamma = \frac{\log(-c') - \log(|\mathcal{Y}| - 1) - \beta}{\alpha} \quad (42)$$

Further simplifications reveal that:

$$\begin{aligned} & \frac{d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x})}{d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x})} \leq \gamma \\ & \Leftrightarrow d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x}) \leq (d_{i,j}^-(\mathbf{x}) + d_{i,j}^+(\mathbf{x}))\gamma \\ & \Leftrightarrow d_{i,j}^-(\mathbf{x}) - d_{i,j}^+(\mathbf{x}) - \gamma d_{i,j}^-(\mathbf{x}) - \gamma d_{i,j}^+(\mathbf{x}) \leq 0 \\ & \Leftrightarrow (1 - \gamma)d_{i,j}^-(\mathbf{x}) - (1 + \gamma)d_{i,j}^+(\mathbf{x}) \leq 0 \\ & \Leftrightarrow (\gamma - 1)d_{i,j}^-(\mathbf{x}) + (1 + \gamma)d_{i,j}^+(\mathbf{x}) \geq 0 \end{aligned} \quad (43)$$

Next, we assume that the closest prototype with the correct label is fixed and denote it by  $\mathbf{p}_i$  – we denote prototypes from the other class as  $\mathbf{p}_j$ . In the end, we iterate over all possible closest prototypes  $\mathbf{p}_i$  and select the one that minimizes the objective (i.e. closeness to the original sample) – note that this approximation drastically increases the number of optimization problems that must be solved and thus the overall complexity of the final algorithm. We then can rewrite Eq. (43) as follows – we make sure that Eq. (43) is satisfied for every possible  $\mathbf{p}_-$ :

$$\begin{aligned} & (\gamma - 1)d_{i,j}^-(\mathbf{x}) + (1 + \gamma)d_{i,j}^+(\mathbf{x}) \geq 0 \\ & \Leftrightarrow (\gamma - 1)d(\mathbf{x}, \mathbf{p}_j) + (1 + \gamma)d(\mathbf{x}, \mathbf{p}_i) \geq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_i) \end{aligned} \quad (44)$$

Applying even more simplifications yield:

$$\begin{aligned}
& (\gamma - 1)d(\mathbf{x}, \mathbf{p}_j) + (1 + \gamma)d(\mathbf{x}, \mathbf{p}_i) \geq 0 \\
& \Leftrightarrow (\gamma - 1)(\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - 2\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j) + \\
& \quad (1 + \gamma)(\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - 2\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{p}_i) \geq 0 \\
& \Leftrightarrow (\gamma - 1)\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} + (1 + \gamma)\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - 2(\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{x} \\
& \quad - 2(1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{x} + (\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j + (1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{p}_i \geq 0 \\
& \Leftrightarrow 2\gamma\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} - 2(\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{x} - 2(1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{x} + \\
& \quad (\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j + (1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{p}_i \geq 0 \\
& \Leftrightarrow \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{q}_j^\top \mathbf{x} + c_j \leq 0
\end{aligned} \tag{45}$$

where

$$\begin{aligned}
\mathbf{q}_j^\top &= \frac{1}{-2\gamma} (-2(\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} - 2(1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega}) \\
c_j &= \frac{1}{-2\gamma} ((\gamma - 1)\mathbf{p}_j^\top \boldsymbol{\Omega} \mathbf{p}_j + (1 + \gamma)\mathbf{p}_i^\top \boldsymbol{\Omega} \mathbf{p}_i)
\end{aligned} \tag{46}$$

Finally, we get the following convex quadratic optimization problem:

$$\begin{aligned}
& \min_{\mathbf{x}_{\text{cf}} \in \mathbb{R}^d} \|\mathbf{x}_{\text{orig}} - \mathbf{x}_{\text{cf}}\|_1 \\
& \text{s.t. } \mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x} + \mathbf{q}_j^\top \mathbf{x} + c_j \leq 0 \quad \forall \mathbf{p}_j \in \mathcal{P}(o_i)
\end{aligned} \tag{47}$$

Note that we have to solve Eq. (47) for every possible closest prototype, every possible class different from the  $i$ -th class and finally for every possible class. Thus, we get the following number of optimization problems (quadratic in the number of classes):

$$|\mathcal{Y}| \cdot (|\mathcal{Y}| - 1) \cdot P_N = |\mathcal{Y}|^2 \cdot P_N - |\mathcal{Y}| \cdot P_N \tag{48}$$

where  $P_N$  denotes the number of prototypes per class used in the pair-wise classifiers.

Note that this number is much larger than  $|\mathcal{Y}|$  which we got without introducing any surrogate or approximation. However, in contrast to Eq. (38), the surrogate optimization problem Eq. (47) is much easier to solve because it is a convex quadratic program (convex QP) which are known to be solved very fast [5].

## B Additional Empirical Results

### B.1 Probabilistic Reject Option

The results for the different methods for computing counterfactual explanations of the probabilistic reject option Eq. (8) are given in Table 4.



Table 4: Algorithmic properties – Mean sparsity (incl. variance) of different counterfactuals, smaller values are better, for feasibility, larger values are better for feasibility.

<i>DataSet</i>	Bb Eq. (12) Feasibility	Algo. 2 Feasibility	BlackBox Eq. (12)	Algo. 2	Algo. 3
Wine	$0.45 \pm 0.15$	$0.45 \pm 0.21$	$11.1 \pm 2.09$	$13.0 \pm 0.0$	$12.67 \pm 0.22$
Breast Cancer	$0.69 \pm 0.12$	$0.49 \pm 0.02$	$26.43 \pm 3.45$	$30.0 \pm 0.0$	$26.04 \pm 60.5$
t21	$0.82 \pm 0.03$	$0.4 \pm 0.24$	$15.94 \pm 2.34$	$17.39 \pm 0.64$	$12.55 \pm 7.41$
Flip	$0.5 \pm 0.2$	$0.4 \pm 0.24$	$9.5 \pm 2.25$	$11.75 \pm 0.23$	$11.5 \pm 1.47$