ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

Geometric-analytical Generation of Counterfactuals for Prototype-based Classifiers

Marika Kaden[1], Lynn Reuss[1], and Thomas Villmann[1,2] *

1 - University of Applied Sciences Mittweida,
Saxon Institute for Comp. Intelligence and Machine Learning,
Mittweida - Germany

2 - Technical University Bergakademie Freiberg, Freiberg - Germany

Abstract. Counterfactuals are useful objects to explain decisions of machine learning classifiers. In the best case, counterfactuals can provide to derive causal inference structure realized by the model. Yet, counterfactual generation in general is known as a constrained optimization problem.
In this contribution we demonstrate that counterfactuals can be determined geometric-analytically in case of prototype based classifiers. For this we only require that nearest prototype classification is based on norms induced by an inner product, which has to be applied for consistency also to evaluate the deviation between a given sample and a desired counterfactual class.

1 Introduction

The behavior of current (deep) machine models often is challenging to interpret due to the huge model complexity. Yet, model explanations frequently are demanded in many applications if crucial decisions have to be made and/or depend on the model outcome. One regarding approach, originally considered in cognitive science to describe contrastive learning , are so-called counterfactuals [10, 14, 21]. In machine learning context, counterfactuals are used to determine model inference limitations in terms of contrastive data samples describing the model behavior for borderline decisions [3, 7, 19]. Yet, in combination with interpretable machine learning approaches like prototype based models, counterfactuals are particularly suited to explore and explain the model inference [12].

Generating counterfactual is understood from machine learning perspective as an optimization process to generate counter-examples regarding model predictions [22], which usually requires efficient optimization schemes due to the generally non-linear optimization scheme [1].

In this contribution we show that the optimization process can be replaced by geometric (analytic) determination of counterfactuals for prototype-based classifiers like learning vector quantization (LVQ) avoiding the utilization of optimization schemes. We give the respective specification of counterfactual definition in contrast to adversarial samples and propose several counterfactual models for prototype-based classifiers like the prominent family of interpretable and shallow learning vector quantization variants [2]. Illustrative numerical examples reflect the behavior of the proposed counterfactual models.

*M.K. and L.R. are supported by grants of the European Social Fund (ESF – 100734114) and of the ILEAS-project *Stiftung Innovation in der Hochschullehre* – (1001-2196).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Counterfactuals as Explorations of Model Limits

As already mentioned, counterfactual explanations should describe model decision limitations. For classification task this can be understood as the determination of a counterfactual $\mathbf{x}_{\text{cf}} \in \mathbb{R}^n$ for a given data sample $\mathbf{x} \in \mathbb{R}^n$ with minimum counterfactual deviation $\delta_{\text{cf}}(\mathbf{x}, \mathbf{x}_{\text{cf}})$ but different class assignment by the classifier model M such that it can be interpreted as a model limit explanation. Accordingly, counterfactual search is formulated as a constrained optimization problem

$$\delta_{\text{cf}}(\mathbf{x}, \mathbf{x}_{\text{cf}}) \stackrel{!}{=} \min \quad \text{subject to} \quad c_M(\mathbf{x}) \neq c_M(\mathbf{x}_{\text{cf}}) \tag{1}$$

as proposed in [22]. Following [6, 15], it has to be carefully distinguished from adversarial examples $\mathbf{x}_{\text{ae}}$, which are designed to fool the model in such a way that the model class assignment $c_M(\mathbf{x}_{\text{ae}})$ differs from the true sample class $c(\mathbf{x})$ with a small deviation $\delta_{\text{ae}}(\mathbf{x}, \mathbf{x}_{\text{ae}}) < \epsilon$ for given threshold $\epsilon > 0$. In particular, the adversarial deviation measure δ_{ae} is usually selected in such a way that in this adversarial aim small disturbances of the original $\mathbf{x}$ contained in the adversarial $\mathbf{x}_{\text{ae}}$ are difficult to detect by human perception. However, the internal, may be implicitly defined, model dissimilarity usually becomes large leading to model misclassification $c_M(\mathbf{x}_{\text{ae}}) \neq c(\mathbf{x})$ [8, 6, 15].

To emphasize the aspect of model limit explanations by counterfactuals, we claim that the counterfactual dissimilarity measure δ_{cf} to evaluate the deviations has to be chosen consistently with the intern model dissimilarity either explicitly or implicitly given e.g. by the respective dissimilarity or its inducing quantity like inner products.

3 Prototype-based Classifiers and Geometric Counterfactual Determination

3.1 Nearest Prototype Classification (NPC)

Prototype-based models for classification are interpretable and robust vector quantizers of data $\mathbf{x} \in X \subseteq \mathbb{R}^n$ based on the nearest-prototype-classification principle (NPC) [17, 11, 2, 16]. Suppose a set $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^n$ of class-responsible prototypes with class labels $c(\mathbf{p}_k) \in \mathcal{C} = \{1, \ldots, C\}$ distributed in the data cloud, for example by training of a LVQ variant like generalized LVQ (GLVQ) [18]. We assume a dissimilarity measure d induced by a norm $\|\cdot\|$. Note that each norm is generated by a semi-inner product $[\cdot, \cdot]$ according to $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ [13], whereas each inner product $\langle \cdot, \cdot \rangle$ also is a semi-inner product and, therefore, induces a norm. Further, as well-known, each norm induces a metric/distance [20] and, hence, the squared Euclidean metric $d_E^2(\mathbf{x}, \mathbf{p}_j) = \langle \mathbf{x} - \mathbf{p}_j, \mathbf{x} - \mathbf{p}_j \rangle_E$ is a inner-product-induced dissimilarity.

In NPC, data classification by the model M is realized by a winner-takes-all (WTA) decision

$$c_M(\mathbf{x}) = c\left(\mathbf{p}_{s(\mathbf{x})}\right) \quad \text{with} \quad s(\mathbf{x}) = \text{argmin}_{j=1}^{M}\left(d(\mathbf{x}, \mathbf{p}_j)\right) \tag{2}$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with respect to the chosen dissimilarity measure d. We take $\mathbf{p}^+ = \mathbf{p}_{s(\mathbf{x})}$ as the best-matching prototype given $\mathbf{x}$, whereas $\mathbf{p}^-$ denotes the best matching prototype regarding another, may be predefined, class $c^- \neq c(\mathbf{p}^+)$. Hence, the local class decision for $\mathbf{x}$ between c^- and $c(\mathbf{p}^+)$ is determined by the hyperplane

$$H\left(\mathbf{p}^+, \mathbf{p}^-\right) = \left\{ \mathbf{x} \in \mathbb{R}^n | \left\langle \mathbf{x}, \mathbf{h} \right\rangle - b_H = 0,\, \mathbf{h} = \frac{\mathbf{p}^+ - \mathbf{p}^-}{\left\| \mathbf{p}^+ - \mathbf{p}^- \right\|},\, b_H = \frac{\left\| \mathbf{p}^+ \right\|^2 - \left\| \mathbf{p}^- \right\|^2}{2 \left\| \mathbf{p}^+ - \mathbf{p}^- \right\|} \right\} \tag{3}$$

where $\mathbf{h}$ is the normal vector determining the hyperplane and b_H is the absolute value shift. We remark that, obviously, $\|\mathbf{h}\| = 1$ is assumed to be valid.

The set $V_k = \{\mathbf{x} \in \mathbb{R}^n | k = s(\mathbf{x})\}$ is the Voronoï-cell of the prototype $\mathbf{p}_k$ and $\mathscr{G}_{\mathrm{D}}(\mathcal{P})$ being the corresponding Delaunay-graph of $\mathcal{P}$ [5]. For a given prototype $\mathbf{p}_k \in \mathcal{P}$ with class assignment $c(\mathbf{p}_k)$, $\mathcal{N}(\mathbf{p}_k) \subseteq \mathcal{P}$ denotes the set of prototype neighbors of the prototype $\mathbf{p}_k$ in the graph $\mathscr{G}_{\mathrm{D}}(\mathcal{P})$. This neighbor set can be decomposed into $\mathcal{N}(\mathbf{p}_k) = \cup_{c=1}^{C} \mathcal{N}_c(\mathbf{p}_k)$, where $\mathcal{N}_c(\mathbf{p}_k)$ is the subset of only those neighbors $\mathbf{p}_j$ of $\mathbf{p}_k$ belonging to class c. Additionally, $\mathcal{N}^-(\mathbf{p}_k) = \mathcal{N}(\mathbf{p}_k) \setminus \mathcal{N}_{c(\mathbf{p}_k)}(\mathbf{p}_k)$ denotes the set of counterfactual prototype neighbor of $\mathbf{p}_k$.

3.2 Geometric Generation of Counterfactuals

In the following we will present three options to generate counterfactuals by geometrically analytic solutions. Here we make explicitly use of the assumption claimed previously that the counterfactual dissimilarity measure δ_{cf} has to consistent with the model dissimilarity. Thus, we assume a norm $\|\cdot\|$ generated by an inner product $\langle \cdot, \cdot \rangle$ for the data space to be used for the model dissimilarity d in the WTA-rule (2) of NPC and set $\delta_{\mathrm{cf}} = d$.

Let $\mathbf{p}^+, \mathbf{p}^- \in \mathcal{P}$ be prototypes of a classifier model M in the sense of the previous section such that the separating hyperplane $H(\mathbf{p}^+, \mathbf{p}^-)$ between them is given by (3). The distance of an arbitrary sample $\mathbf{x} \in \mathbb{R}^n$ to this hyperplane is defined by the solution of the minimum problem $d(H(\mathbf{p}^+, \mathbf{p}^-), \mathbf{x}) = \min_{\mathbf{u} \in H} \|\mathbf{u} - \mathbf{x}\|$ which can be calculated analytically. Using the property $\|\mathbf{h}\| = 1$, we result

$$d\left(H\left(\mathbf{p}^+, \mathbf{p}^-\right), \mathbf{x}\right) = \left| b_H - \langle \mathbf{h}, \mathbf{x} \rangle \right|$$

based on the orthogonal projection point $\mathbf{x}_H \in H$ of $\mathbf{x}$ by

$$\mathbf{x}_H = \mathbf{x} + \underbrace{\left(\frac{\left\| \mathbf{p}^+ \right\|^2 - \left\| \mathbf{p}^- \right\|^2}{2 \left\| \mathbf{p}^+ - \mathbf{p}^- \right\|} - \left\langle \frac{\mathbf{p}^+ - \mathbf{p}^-}{\left\| \mathbf{p}^+ - \mathbf{p}^- \right\|}, \mathbf{x} \right\rangle \right)}_{\lambda} \cdot \mathbf{h} \tag{4}$$

which, in fact, is the best approximation of $\mathbf{x}$ within the hyperplane. Because $c(\mathbf{p}^+) \neq c(\mathbf{p}^-)$ and $d(\mathbf{x}_H, \mathbf{p}^+) = d(\mathbf{x}_H, \mathbf{p}^-)$ is valid, $\mathbf{x}_H$ constitutes a counterfactual sample $\mathbf{x}_{\mathrm{cf}}$ for the original sample $\mathbf{x}$ with model class assignment $c_M(\mathbf{x})$ with respect to the class $c(\mathbf{p}^-)$.

Another option for counterfactual generation would be the search for a counterfactual starting at the given sample $\mathbf{x}$ along the direction vector $\mathbf{z} = \mathbf{p}^- - \mathbf{x}$ pointing to $\mathbf{p}^-$. In this case we get $\mathbf{x}_{\mathrm{cf}} = \mathbf{x} + \alpha \cdot \mathbf{z}$ where α is determined by

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

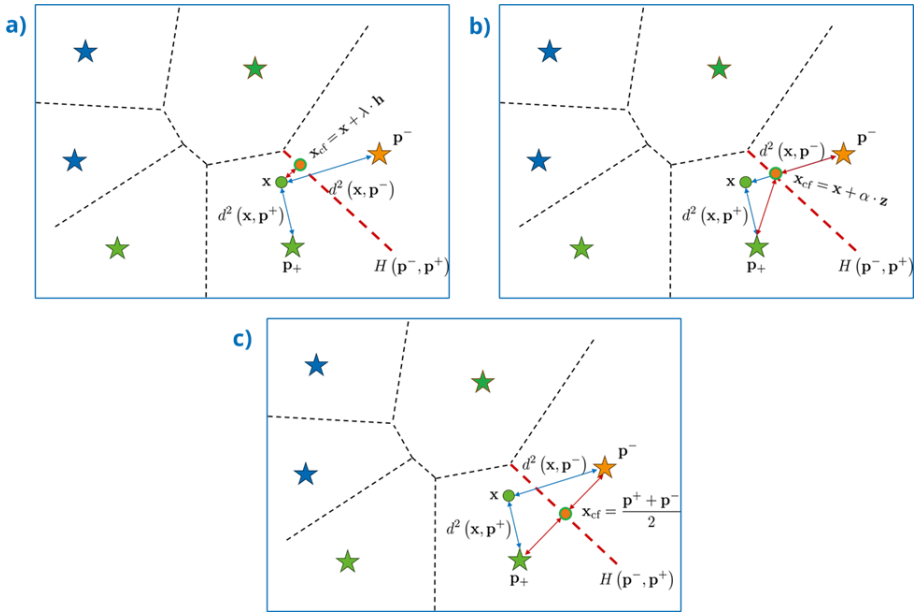

Fig. 1: Illustration of the proposed three options to generate counterfactuals $\mathbf{x}_{\text{cf}}$: **a)** determination by means of the orthogonal projection, **b)** using the shift into the direction of the best matching prototype $\mathbf{p}^-$ of the desired counterfactual class and **c)** midpoint between best matching prototype $\mathbf{p}^+$ and $\mathbf{p}^-$.

the constraint $\|\mathbf{x}_{\text{cf}} - \mathbf{p}^+\| = \|\mathbf{x}_{\text{cf}} - \mathbf{p}^-\|$ ensuring $\mathbf{x}_{\text{cf}} \in H(\mathbf{p}^+, \mathbf{p}^-)$. One easily calculates

$$\alpha = \frac{\|\mathbf{x} - \mathbf{p}^+\|^2 - \|\mathbf{x} - \mathbf{p}^-\|^2}{2\langle \mathbf{p}^- - \mathbf{x}, \mathbf{p}^+ - \mathbf{p}^- \rangle}$$

as the resulting scaling factor. This directed counterfactual version particularly appreciates the information acquired by the prototype $\mathbf{p}^-$ during the learning.

If for the given sample must be assumed noisy uncertainty, it might be an option to neglect the exact value of $\mathbf{x}$ and the counterfactual is determined using only the prototypes $\mathbf{p}^+$ and $\mathbf{p}^-$. This motivation gives the natural choice

$$\mathbf{x}_{\text{cf}} = \frac{\mathbf{p}^+ + \mathbf{p}^-}{2} + \mathbf{p}^+ \tag{5}$$

constituting a noise robust setting.

Finally, one has to verify for the generated counterfactual that the class prediction by the model is in agreement with the desired couterfactual class, i.e. one has to validate $c_M(\mathbf{x}_{\text{cf}}) = c(\mathbf{p}^-)$ because this restriction is not explicitly used in the generating process.

The proposed variants of counterfactuals are visualized for a two-dimensional setting in Fig. 1

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

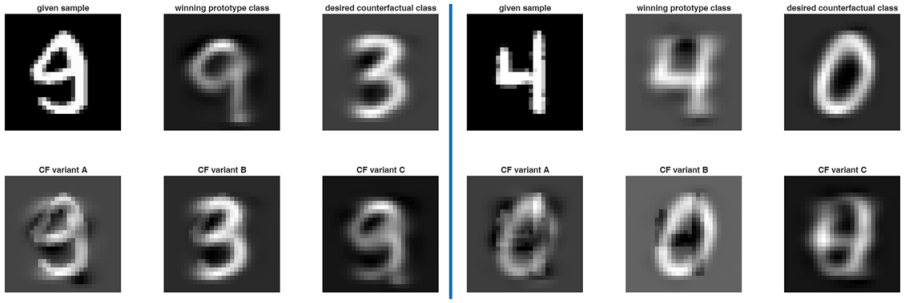

Fig. 2: Examples of counterfactual generation for MNIST-data: **Left block** The sample $\mathbf{x}$ (uppper left corner) is classified by a prototype '9' ($\mathbf{p}^+$) and the desired counterfactual class is '3' with prototype $\mathbf{p}^-$ being depicted in the upper right corner. The lower row shows the three types of geometric counterfactual generation proposed in this paper see Fig. 1. **Right block:** Same as for the left block but here for $\mathbf{x}$ as '4' and the counterfactual should be '0'.

4 Illustrative Numerical Experiments and Summary

For illustration of the proposed counterfactual variants we have chosen the well-known MNIST-database [9]. We trained a standard GLVQ with 3 prototypes per class and determined for a given sample the best matching prototype $\mathbf{p}^+$ together with the best matching prototype $\mathbf{p}^-$ of a given desired class. Using these prototypes, the introduced types of counterfacuals weere generated - see Fig. 2. Obviously, we detect substantial variations between the generated counterfactual depending on the type of generation, i.e. depending on the perspective/aim of the counterfactual definition. This, however, is application and user dependent. The approach can be extended to use in GMLVQ if applied in the embedding space. An disadvantage of the current geometric counter factual generation is the sensitivity regarding the separating hyperplane between the protototypes $\mathbf{p}^+$ and $\mathbf{p}^-$.

Therefore, future developments should take into account GLVQ hypothesis margin information [4], which allows to generate 'more safe' counterfactuals. Another optioon for future research is to extend this approach for norms induced by semi-inner products like genera l_p norms [13].

References

[1] A. Artelt and B. Hammer. Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. *Neurocomputing*, 470:304–317, 2022.

[2] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[3] R. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19, Macao)*, pages 6276–6282, 2019.

[4] K. Crammer, R. Gilad-Bachrach, A. Navot, and N.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural*

Information Processing (Proc. NIPS 2002), volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

[5] B. Delaunay. Sur la sp re vide. *Bull. Acad. Sci. USSR (VII), Classe Sci. Mat. Nat.*, pages 793–800, 1934.

[6] T. Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32:77–109, 2021.

[7] M. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680, San Diego, 2014. Curran Associates, Inc.

[9] Y. LeCun, C. Cortes, and C. Burges. The MNIST database, 1998.

[10] D. Lewis. *Counterfactuals*. Blackwell Publishers, Oxford, 1973.

[11] O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 3531–3537. Association for the Advancement of Artificial Intelligence, AAAI Press, 2018.

[12] A. V. Looveren and J. Klaise. Interpretable counterfactual explanations guided by prototypes. In N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano, editors, *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2021)*, pages 650–665, Cham, 2021. Springer International Publishing.

[13] G. Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[14] S. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition, 2014.

[15] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, and H. Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In G. Camps-Valls, F. Ruiz, and I. Valera, editors, *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pages 4574–4594, 2022.

[16] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Survey*, 16:1–85, 2022.

[17] S. Saralajew, L. Holdijk, and T. Villmann. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 13635–13650. Curran Associates, Inc., 2020.

[18] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[19] I. Stepin, J. Alonso, A. Catala, and M. Pereira-Farina. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[20] H. Triebel. *Analysis und mathematische Physik*. BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 3rd, revised edition, 1989.

[21] N. van Hoeck, P. Watson, and A. Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9(420):1–18, 2015.

[22] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Havard Journal of Law and Technology*, 31(2):841–887, 2018.