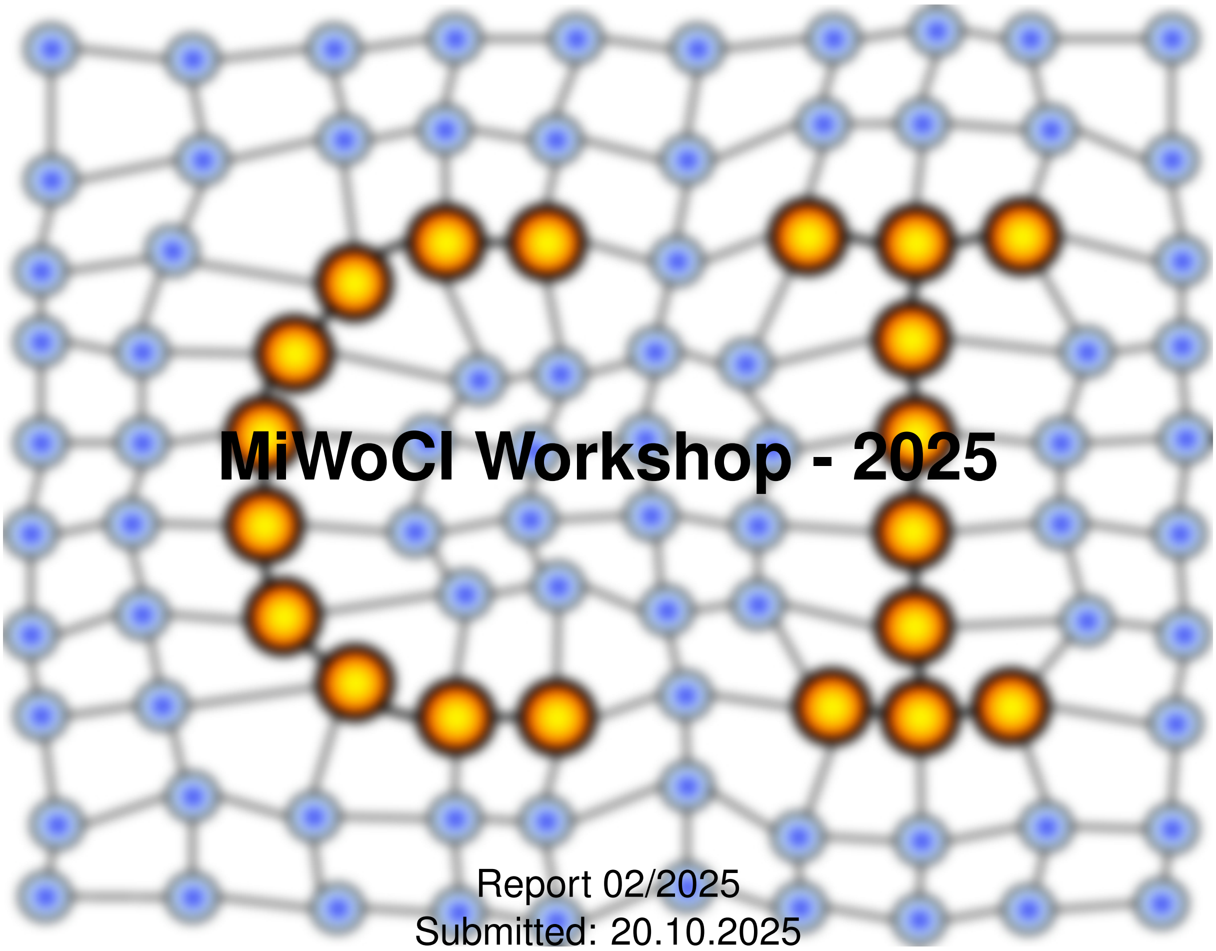
MACHINE LEARNING REPORTS



MiWoCI Workshop - 2025

Report 02/2025
Submitted: 20.10.2025
Published: 21.01.2026

Frank-Michael Schleif[1,2*,3*], Susanne Franke[2], Marika Kaden[2], Thomas Villmann[2]
(Eds.)
(1) Technical University of Applied Sciences Wuerzburg-Schweinfurt,
Sanderheinrichsleitenweg 20, 97074 Wuerzburg, Germany (2) University of Applied Sciences
Mittweida, Technikumplatz 17, 09648 Mittweida, Germany (3) University of Birmingham,
School of Computer Science,
Edgbaston, B15 2TT Birmingham, UK

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Impressum
Publisher: University of Applied Sciences Mittweida
Technikumplatz 17,
09648 Mittweida, Germany
Editor: Prof. Dr. Thomas Villmann
Prof. Dr. Frank-Michael Schleif

Technical-Editor: Prof. Dr. Frank-Michael Schleif
Contact: frank-michael.schleif@fhws.de
URL: http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN: 1865-3960

Abstracts of the 17^{th} Mittweida Workshop on Computational Intelligence and Beyond - MiWoCI 2025 -

Frank-Michael Schleif, Susanne Franke, Marika Kaden, and Thomas Villmann

Machine Learning Report 02/2025

Preface

The 17 th international *Mittweida Workshop on Computational Intelligence* (MiWoCI) gathering together more than 40 scientists from different universities including Bielefeld, Groningen, UAS Mittweida, UAS Würzburg-Schweinfurt, UAS Zwickau, TU Freiberg and University Lübeck. The workshop took place at a new universtity building in Werkbank 32 in Mittweida, a sustainable building designed for meetings and research exchange, and for all who could not attend in person the workshop was hybrid. Thus, from 01.9 - 3.9.2025 the tradition of scientific presentations, vivid discussions, and exchange of novel ideas at the cutting edge of research was continued. They were connected to diverse topics in computer science, life science, and machine learning.

This report is a collection of abstracts and short contributions about the given presentations and discussions, which cover theoretical aspects, applications, as well as strategic developments in the fields.

Contents

Psychological Experiments: Replacing human participants by LLMs???

Valerie Vaquet[1], Sarah Schröder[1], Thekla Morgenroth[2], Ulrike Kuhl[1], and Benjamin Paaßen[1]

[1]Bielefeld University, [2]Purdue University

Abstract

LLMs are increasingly used to automate many tasks in different settings ranging from software development, to content creation and education. They also play a considerable role in research, for instance in scientific writing tasks, literature search, and data annotation[1].

Recently, there has also been work suggesting that one can replace human participants in psychological studies by LLMs. Dillion et al. [2] focused on moral judgement tasks. They considered 464 moral scenarios and compared the ratings obtained by GPT-3.5 to human ones. Based on a high correlation between human and LLM outputs, they argue for the potential to replace human participants by LLMs. Most recently, Binz et al. [1] published the CENTAUR model, "a computational model that can predict and simulate human behaviour in any experiment expressible in natural language". CENTAUR builds opon a Llama-3.1 70b LLM and is fine-tuned on a dataset containing about 10 million human responses collected across 160 psychological experiments.

There also has been considerable criticism focusing on issues connected to biases, hallucinations [4] and the LLMs' inability to accurately represent human diversity [3]. Arging against replacing human participants by LLMs, we rely to the foundational argument of generalization in ML: ML models perform quite well on new tasks that are sufficiently similar to the training data, e.g. experiments that were already conducted and are part of the training corpus. However, there is no reason to believe that LLMs or other ML models generalize to out-of-distribution data, e.g. interesting novel psychological experiments. We showcase this by slightly rewording the moral judgement tasks used in [2] and re-evaluating human responses and LLM outputs.

References

[1] Marcel Binz et al. "A foundation model to predict and capture human cognition". en. In: *Nature* (July 2025). ISSN: 0028-0836, 1476-4687. DOI: `10.1038/s41586-025-09215-4`. URL: `https://www.nature.com/articles/s41586-025-09215-4` (visited on 07/30/2025).

[1] (, and the review process). Let's leave the discussion of whether this is good scientific practice for the coffee break :-)

[2] Danica Dillion et al. "Can AI language models replace human participants?" en. In: *Trends in Cognitive Sciences* 27.7 (July 2023), pp. 597–600. ISSN: 13646613. DOI: 10.1016/j.tics.2023.04.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661323000980 (visited on 07/30/2025).

[3] Jacqueline Harding et al. "AI language models cannot replace human research participants". en. In: *AI & SOCIETY* 39.5 (Oct. 2024), pp. 2603–2605. ISSN: 0951-5666, 1435-5655. DOI: 10.1007/s00146-023-01725-x. URL: https://link.springer.com/10.1007/s00146-023-01725-x (visited on 07/31/2025).

[4] Luca Rossi, Katherine Harrison, and Irina Shklovski. "The Problems of LLM-generated Data in Social Science Research". en. In: *Sociologica* 18.2 (Oct. 2024), pp. 145–168. DOI: 10.6092/ISSN.1971-8853/19576. URL: https://sociologica.unibo.it/article/view/19576 (visited on 07/31/2025).

The Influence of Drift in a Mismatched Student-Teacher Setting

Frederieke Richert, Otavio Citton, Michael Biehl

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Despite the widespread usage of machine learning techniques and especially neural networks, the inner workings of these machines are not yet fully understood. Statistical physics has played an important role in the last decades in increasing the theoretical understanding about the typical behavior of neural networks.

In previous works in the Intelligent Systems group Groningen, the focus was on distinguishing differences between different types of activation functions [1, 2] and techniques were developed to analyse networks with arbitrary activation functions [3, 4].

We study a relatively simple network with only one hidden layer and fixed hidden-to-output weights, the so called Soft Committee Machine (SCM). To control the training task, we assume a second, similarly structured SCM. Thus, the task is given by this so called teacher network, which the trainable student network endeavours to imitate.

Extending previous results on the influence of drift in student-teacher scenarios with either ReLU or sigmoidal activation function [5], we analyse this setting in the mismatched case, where the student and the teacher SCM have different activation functions. This corresponds to a more realistic learning scenario and illuminates interesting dependencies of learning on the activation functions, the drift and weight decay parameters.

References

[1] E. Oostwal, M. Straat, and M. Biehl, Physica A. **Vol. 564**, 125517 (2021)

[2] F. Richert, M. Straat, E.Oostwal, and M. Biehl, Proceedings ESANN 2023, p.435-440 (2023)

[3] O. Citton, F. Richert, M. Biehl, Proceedings ESANN 2024, p.437-442 (2024)

[4] O. Citton, F. Richert, M. Biehl, Physica A. **Vol. 660**, 130356 (2025)

[5] M. Straat, F. Abadi, Z. Kan, C. Göpfert, B. Hammer, M. Biehl, Neural Computing and Applications 34 (1) 101-118 (2022)

Three research directions for more interpretable prototypes

Benjamin Paassen[1]

[1]Faculty of Technology, Bielefeld University

preprint as provided by the authors

One of the key arguments in favor of prototype-based models (namely learning vector quantization or LVQ models) is their interpretability and explainability: Any classification decision is made by assigning the label of the closest prototype. Hence, the decision can be explained by inspecting the closest prototype; and the full model can be interpreted by inspecting the set of prototypes and the metric used Nova and Estévez [2014], Kaden et al. [2022], Lisboa et al. [2023].

More formally, let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ be a metric over some input space $\mathcal{X}$ and let $w_1, \dots, w_K \in \mathcal{X}$ be a set of K (for small K) prototypes with labels $z_1, \dots, z_K \in \{1, \dots, L\}$. Then, any data point $x \in \mathcal{X}$ is assigned the label

$$f(x) = z_k \qquad \text{where} \qquad k = \arg\min_k d(x, w_k). \tag{1}$$

Thus, the model is not only locally explainable, but globally. Even better: Not only the trained model is nicely explainable, but the training is as well, essentially pulling prototypes to data points of the same class and pushing them away from data points of different classes Nova and Estévez [2014], Kaden et al. [2022].

Beyond these interpretability and explainability advantages, LVQ models also exhibit favorable computational efficiency with linear training times in stochastic gradient descent, constant-time inference because any classification only requires K distance computations, as well as low memory footprint because only K prototypes need to be stored. Recently, adversarial robustness with respect to the metric d has also been achieved Saralajew et al. [2020]; and LVQ models are easily adaptable to novel data types as long as an appropriate metric d can be defined Paaßen et al. [2018].

All these nice properties beg the question: Why are LVQ models not used more frequently? A modest attempt at an answer has three parts. First, prototype models are, in fact, used quite frequently, just not under the label LVQ but as prototype networks and related concepts in few-shot learning and meta learning Snell et al. [2017]. Second, the metrics used in classic LVQ models are mostly constrained to the Euclidean distance, general quadratic forms, and cosine distances, which insufficiently address the complexities of commonly used data types,

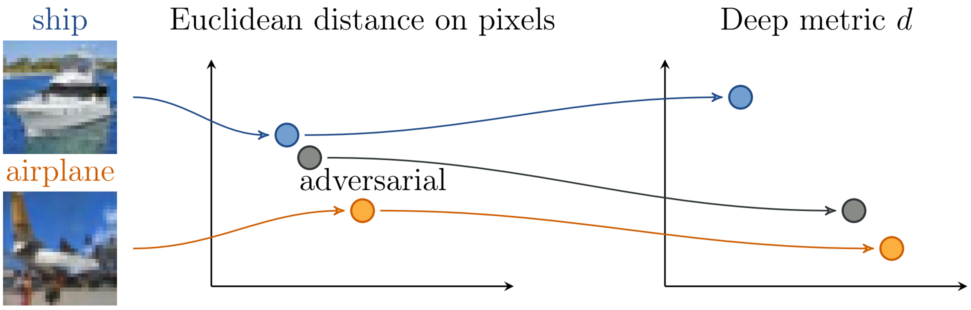


Figure 1: An illustration of adversarial attacks due to representation changes. As soon as points in Euclidean space are pushed apart in the metric used for classification, opportunities for adversarial attacks emerge.

especially images and text; instead, such data is more successfully addressed via representation learning techniques involving multiple layers of neural networks, such as Radford et al. [2021], Hu et al. [2024]. Third, the interpretability and explainability of prototype often breaks down for typical real-world data sets. Consider the examples of images: even in the simple example of the CIFAR10 data set, prototypes typically look like random pixel mush, complicating any interpretation, and we need many of them to achieve reasonable accuracy, also complicating interpretability; or the example of distance datasets, where the prototypes of a relational model are given only in terms of Lagrange coefficients Hammer et al. [2014], Hofmann et al. [2014]. In such cases, we often refrain from interpreting the prototypes directly but move to exemplars Hofmann et al. [2014] or the relevance matrix, instead Schneider et al. [2009], Lövdal and Biehl [2024].

Therefore, we argue that more research is needed to make prototypes more interpretable. In particular, we see three different research directions that may contribute to this overarching goal.

1 From Metric Learning to Deep Representation Learning

A crucial limitation regarding the performance of LVQ models on image or text data so far has been that the Euclidean distance, as a metric, does not conform to human intuitions of similarity. To address this issue, metric learning has been proposed—but metric learning for prototypes has so far mostly been limited to general quadratic forms Schneider et al. [2009] or edit distances Paaßen et al. [2018], while most recent advances in contrastive learning and representation learning with deep networks Hu et al. [2024] have not yet been translated to the world of LVQ models. Technologically, developments like ProtoTorch Ravichandran [2020] make it easy to integrate deep learning methods with prototype-based models. However, one should also acknowledge that this research direction is fraught with other problems: If we increase the representational power of our metric, such that it can bring data points together that humans judge as similar (despite larger Euclidean distance) and pull data points apart that humans judge as dissimilar (despite smaller Euclidean distance), our models will almost inevitably also loose interpretability in the computing process for the distance—and become more susceptible to adversarial attacks Ilyas et al. [2019]. Consider the example in Figure 1: We have two images from different classes in the training data that share a lot of similar pixels (white and blue); hence, their Euclidean distance is relatively low. To classify these images correctly, we need to learn a metric d that pulls these images apart despite their low Euclidean distance. However, as soon as we do so, it is likely that we can find an adversarial, i.e. an image with very low Euclidean distance to one of the training images but high d, such that it is classified incorrectly.

Therefore, we will need to find new representation learning schemes that maintain robustness guarantees Hammer et al. [2005], Saralajew et al. [2020] and "well-behavedness" even if we use highly nonlinear transformations to achieve a better metric d. As a very first, simple starting point, we could try to impose Lipschitz continuity conditions onto our metric d, in the sense that there should exist some constant L, such that for any two points x, x' we have

$$d(x, x') \leq L \cdot \|x - x'\|. \tag{2}$$

If that is true, adversarial robustness with respect to d also translates to adversarial robustness with respect to the Euclidean distance. More precisely, let's assume the LVQ model guarantees that for any point x from the training data, any point x' that would be classified differently is at least ϵ apart from x. Accordingly, the Lipschitz constraint ensures that $\|x - x'\| \geq \frac{\epsilon}{L}$, thus preventing adversarials with a Euclidean distance smaller than that.

2 Median Learning Vector Quantization

Models become substantially more interpretable if the prototypes stem from the training data set as one can directly inspect the prototypes. Such LVQ variants have been dubbed median LVQ Nebel et al. [2015]. A core challenge in median LVQ is that we lose the ability to continuously shift in the input space but, instead, we have to apply discrete optimization methods. Nebel et al. [2015] pushed this forward with an EM approach, which has later been combined with metric learning Paaßen et al. [2018]. But this discrete approach can also be found in the set cover formulation of Bien and Tibshirani [2011].

As one exemplary idea in that direction: While choosing the optimal prototypes in median LVQ is provably NP-hard, we can show that a restricted version of median LVQ can be phrased as a mixed integer linear program (mILP) for which very efficient heuristic solvers exist Huangfu and Hall [2018].

In particular, we will apply two simplifications. First, we replace the GLVQ loss by the hinge loss $[d_i^+ - d_i^- + \gamma]_+$ Paaßen [2019], where d_i^+ is the distance of data point i to the closest correct prototype, d_i^- is the distance to the closest incorrect prototype, and γ is a hyperparameter called *margin*, and we permit only one prototype per class. Now, let $\vec{\alpha} \in \{0, 1\}^N$ be a binary indicator variable, where $\alpha_i = 1$ if and only if data point i is selected as a prototype for class y_i. If we restrict our model to one prototype per class, we know that $\sum_{i:y_i=\ell} x_i = 1$ for all classes ℓ. Accordingly, the distance d_i^+ can be expressed as $d_i^+ = \sum_{j:y_j=\ell} x_j \cdot d_{i,j}$, because exactly one x_j in this sum is one. d_i^- is a bit more challenging to express. For this, we need to introduce a slack variable d_i^-, upper-bounded by the distance of data point i to the prototypes with a different label:

$$d_i^- \leq \sum_{j:y_j=\ell} x_j \cdot d_{i,j} \qquad \forall i \in \{1,\ldots,N\}, \ell \neq y_i.$$

Further, we introduce a slack variable ϵ_i for each data point i that is supposed to express the hinge loss $[d_i^+ - d_i^- + \gamma]_+$ and, thus, is lower-bounded by $\epsilon_i \geq d_i^+ - d_i^- + \gamma$. Overall, we arrive at the following mixed integer linear program:

$$\begin{aligned}
\min_{\vec{x}\in\{0,1\}^N, \vec{\epsilon}\in\mathbb{R}^N, \vec{d}^-\in\mathbb{R}^N} \quad & \sum_{i=1}^N \epsilon_i + \lambda \cdot \sum_{i=1}^N d_i^+ && \qquad (3) \\
\text{such that} \quad & d_i^+ = \sum_{j:y_j=y_i} x_j \cdot d_{i,j} && \forall i \in \{1,\ldots,N\} \\
& d_i^- \leq \sum_{j:y_j=l} x_j \cdot d_{i,j} && \forall i \in \{1,\ldots,N\}, l \neq y_i \\
& \epsilon_i \geq d_i^+ - d_i^- + \gamma && \forall i \in \{1,\ldots,N\} \\
& \epsilon_i \geq 0 && \forall i \in \{1,\ldots,N\} \\
& \sum_{i:y_i=l} x_i = 1 && \forall l \in \{1,\ldots,L\}
\end{aligned}$$

Note that the first equality constraint is not explicitly enforced but is merely a definition. So, we overall have an mILP with $N \cdot (L-1)$ inequality constraints for the d_i^- variables, $2 \cdot N$ inequality constraints for ϵ_i (or, rather, N constraints and N bounds), and L equality constraints to enforce exactly one prototype per class. Also note that we introduce the regularization term $\lambda \cdot d_i^+$ to encourage that each prototype is close to the median of its receptive field.

Even though we do not provide a formal proof at this point, the optimization will really push ϵ_i to represent the hinge loss because the objective enforces ϵ_i to be as low as possible, such that the corresponding inequalities become exact. By contrast, d_i^- may be *smaller* than the true distance to the closest negative prototype because it will only grow until the hinge loss is zero—if that is possible. If that is not possible, d_i^- does represent the distance to the closest negative prototype.

First empirical experiments suggest that this mILP formulation is much slower than prior, heuristic approaches to find good prototypes and may not necessarily improve accuracy, either. As such, this formulation may rather be seen as a starting point for further exploration compared to an actual proposal for a solution. Nonetheless, it shows that further conceptual improvements are possible toward median LVQ models, even though the idea of median LVQ is a decade old.

3 Fewer prototypes in tree-based models

If we need many prototypes to represent our classes, we reduce computational efficiency and make it more difficult to interpret our model. Therefore, it would be beneficial to reduce the number of prototypes needed to achieve good accuracy. One way is metric learning to achieve metric spaces that pull all points in a class together such that a single prototype suffices Snell et al. [2017]. An alternative route is to combine prototypes with other, similarly intuitive decision structures.

A first idea of our lab in that direction is the concept of *prototype decision trees* where each decision node is a protoype-based model. The potential advantages are two-fold: First, decision trees only establish new nodes (and therefore new prototypes) if the decision at a particular part of the space is not clear, yet. In that sense they are parsimoneous in the number of prototypes. Second, they solve a crucial problem of vanilla decision trees, namely that they need excessive numbers of decision nodes for correlated features (where the decision plane should be oblique, not aligned to an axis of the space) Yang et al. [2019].

However, first empirical experiments suggest that the idea is much harder to implement than it may appear on first glance. Prototype decision trees may still overfit and the interpretability problems of classic LVQ models are inherited by prototype decision trees. Hence, further research is needed.

4 Conclusion

We present three ideas to make prototype models more interpretable—all challenging, but all profiting from pairwise synergies. Better metrics will also enhance the accuracy and interpretability of median models or prototype decision trees; median models will also make metric learning more efficient and make prototype decision trees more interpretable; and median models will be much easier to train for one prototype per class, which is sufficient if we embed them in prototype decision trees. However, these are not the only research directions possible. Substantial success has also been achieved in making prototypes more flexible by treating them as combinations of concepts Saralajew et al. [2019]. Further, the final proof of interpretability lies in user studies: can users actually assess, predict, and utilize these prototype-based models in real-world decision contexts Speith et al. [2024]? Performing such studies more frequently is, thus, another direction I recommend.

References

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011. URL `http://www.jstor.org/stable/23069335`.

Barbara Hammer, Marc Strickert, and Thomas Villmann. On the generalization ability of grlvq networks. *Neural Processing Letters*, 21:109–120, 2005. doi:10.1007/s11063-004-1547-1.

Barbara Hammer, Daniela Hofmann, Frank-Michael Schleif, and Xibin Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, 131:43–51, 2014. doi:10.1016/j.neucom.2013.05.054.

Daniela Hofmann, Frank-Michael Schleif, Benjamin Paaßen, and Barbara Hammer. Learning interpretable kernelized prototype-based models. *Neurocomputing*, 141:84–96, 2014. doi:10.1016/j.neucom.2014.03.003.

Haigen Hu, Xiaoyuan Wang, Yan Zhang, Qi Chen, and Qiu Guan. A comprehensive survey on contrastive learning. *Neurocomputing*, 610:128645, 2024. doi:10.1016/j.neucom.2024.128645.

Qi Huangfu and JA Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi:10.1007/s12532-017-0130-5.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of the NeurIPS*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf`.

Marika Kaden, Katrin Sophie Bohnsack, Mirko Weber, Mateusz Kudła, Kaja Gutowska, Jacek Blazewicz, and Thomas Villmann. Learning vector quantization as an interpretable classifier for the detection of sars-cov-2 types based on their rna sequences. *Neural Computing and Applications*, 34(1):67–78, 2022. doi:10.1007/s00521-021-06018-2.

P.J.G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann. The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, 535:25–39, 2023. doi:10.1016/j.neucom.2023.02.040.

Sofie Lövdal and Michael Biehl. Iterated relevance matrix analysis (irma) for the identification of class-discriminative subspaces. *Neurocomputing*, 577:127367, 2024. doi:10.1016/j.neucom.2024.127367.

David Nebel, Barbara Hammer, Kathleen Frohberg, and Thomas Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015. doi:10.1016/j.neucom.2014.12.096. Learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis.

David Nova and Pablo A. Estévez. A review of learning vector quantization classifiers. *Neural Computing and Applications*, 25(3):511–524, 2014. doi:10.1007/s00521-013-1535-3. URL `https://arxiv.org/abs/1509.07093`.

Benjamin Paaßen. Large margin learning vector quantization. In Frank-Michael Schleif and Thomas Villmann, editors, *Proceedings of the MiWoCI Workshop 2019*, 2019. URL `https://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2019.pdf`.

Benjamin Paaßen, Claudio Gallicchio, Alessio Micheli, and Barbara Hammer. Tree Edit Distance Learning via Adaptive Symbol Embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3973–3982, 2018. URL `http://proceedings.mlr.press/v80/paassen18a.html`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

J Ravichandran. Prototorch, 2020. URL `https://github.com/si-cim/prototorch`.

Sascha Saralajew, Lars Holdijk, Maike Rees, Ebubekir Asan, and Thomas Villmann. Classification-by-components: Probabilistic modeling of reasoning over a set of components. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/dca5672ff3444c7e997aa9a2c4eb2094-Paper.pdf`.

Sascha Saralajew, Lars Holdijk, and Thomas Villmann. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proceedings of the NeurIPS*, volume 33, pages 13635–13650, 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/9da187a7a191431db943a9a5a6fec6f4-Paper.pdf`.

Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009. doi:10.1162/neco.2009.11-08-908.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf`.

Timo Speith, Barnaby Crook, Sara Mann, Astrid Schomäcker, and Markus Langer. Conceptualizing understanding in explainable artificial intelligence (xai): an abilities-based approach. *Ethics and Information Technology*, 26(2):40, 2024. doi:10.1007/s10676-024-09769-3.

Bin-Bin Yang, Song-Qing Shen, and Wei Gao. Weighted oblique decision trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):5621–5627, 2019. doi:10.1609/aaai.v33i01.33015621.

The Effect of Existential Presuppositions on Hallucinations in Large Language Models

Jonas Vaquet

Bielefeld University

Abstract

With the rapid proliferation of Large Language Models (LLMs), they are increasingly used for information retrieval by end users. Internet search engines provide LLM-generated summaries and users directly turn to LLM-based conversational agents for information retrieval in order to save time and effort [1].

LLMs, however, fail to properly verbalize their uncertainty. This leads to a calibration gap where humans overestimate the accuracy of LLM-generated responses. At the same time, hallucinations where LLMs generate responses that contradict real-world facts are a well-known problem. Their consequences can reach from misinformation to real-world harm.

The main benefit of using LLMs for end users is that they can be interacted with using natural language. In natural language preexisting knowledge, however, is often implicitly encoded using so-called presuppositions [3]. End users can thus inadvertently inject their believes into prompts meant for information retrieval. Presuppositions that imply the existence of an entity are called existential presuppositions.

LLMs have been shown accommodate presuppositions in previous work [2]. If these presuppositions are factually wrong, this mechanism can cause more hallucinations. In this work, we aim to quantify the effect and explain the mechanism by which existential presuppositions included in user prompts influence LLM responses.

References

[1] Thashmee Karunaratne and Adenike Adesina. “Is it the new Google: Impact of ChatGPT on Students’ Information Search Habits”. In: *European Conference on e-Learning* 22 (Oct. 2023), pp. 147–155. DOI: 10.34190/ecel.22.1.1831.

[2] Najoung Kim et al. *$(QA)^2$: Question Answering with Questionable Assumptions.* 2023. arXiv: 2212.10003 [cs.CL]. URL: https://arxiv.org/abs/2212.10003.

[3] Paul Kroeger. *Analyzing meaning. An introduction to semantics and pragmatics. Third edition.* Textbooks in Language Sciences 5. Berlin: Language Science Press, 2022. DOI: 10.5281/zenodo.6855854.

Fast Markov chain Monte Carlo for Bayesian Machine Learning

Björn Sprungk

TU Bergakademie Freiberg

Abstract

In this talk we consider Markov chain Monte Carlo methods for computing the predictive distribution in Bayesian machine learning such as Bayesian neural networks or Gaussian process classification [3]. In Bayesian machine learning the unknown hypothesis or its parameters such as weights and biases are learned by conditioning a chosen prior distribution for the unknowns given the training data. This results in a posterior distribution for the parameters which in practice can only be approximated, e.g., by Markov chain Monte Carlo sampling. A common method to do so, particularly, for Gaussian priors, is the elliptial slice sampler [1]. This method seems particularly suited for high-dimensional settings such as Gaussian process classification. However, a convergence analysis was missing in the literature. In this talk, we present first results which guarantee an exponentially fast convergence of the elliptical slice sampler [2] and comment on recent developments in the field.

References

[1] I. Murray, R. Adams, D. MacKay. Elliptical slice sampling. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 541–548, 2010.

[2] V. Natarovskii, D. Rudolf, B. Sprungk. Geometric Convergence of Elliptical Slice Sampling. In: *Proceedings of the 38th International Conference on Machine Learning*, pp. 7969–7978, 2021.

[3] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for machine learning.* MIT Press, 2006.

Learning Partitions of Dynamical Systems with Stability Guarantees

Lucas Schwarz and Florian Röhrbein

Chemnitz University of Technology

Abstract

Approximating the behaviour of dynamical systems through parametric functions fitted on sampled data when no analytical expression is available constitutes a popular approach in various fields such as control theory, computational neuroscience and robotics [1–3]. A particular method is the partitioning of the state-space into compact subsets and subsequently learning multiple local linear dynamical systems which are combined to model the global non-linear system behaviour. This strategy has been popular in the field of robotics to generate movement policies from kinesthetically demonstrated trajectories [4–6]. However, topological aspects of the available data have been sparsely considered in this domain [7]. In this contribution, we propose to generate a learned explicit topological representation of the underlying system trajectory data using prototypes and to incorporate this topology into the generation of provably stable locally linear dynamical systems with physically consistent mode transitions. We demonstrate the validity of our approach on a synthetic benchmark dataset and on a real robot experiment.

References

[1] A. C. Costa, T. Ahamed, and G. J. Stephens, "Adaptive, locally linear models of complex dynamics," *Proceedings of the National Academy of Sciences*, vol. 116, no. 5, pp. 1501–1510, 2019.

[2] A. Hu, D. Zoltowski, A. Nair, D. Anderson, L. Duncker, and S. Linderman, "Modeling latent neural dynamics with gaussian process switching linear dynamical systems," *Advances in Neural Information Processing Systems*, vol. 37, pp. 33805–33835, 2024.

[3] D. Pfrommer, M. Simchowitz, T. Westenbroek, N. Matni, and S. Tu, "The power of learned locally linear models for nonlinear policy optimization," in *International Conference on Machine Learning*, pp. 27737–27821, PMLR, 2023.

[4] N. Figueroa and A. Billard, "A physically-consistent bayesian non-parametric mixture model for dynamical system learning.," in *CoRL*, pp. 927–946, 2018.

[5] N. Figueroa and A. Billard, "Locally active globally stable dynamical systems: Theory, learning, and experiments," *The International Journal of Robotics Research*, vol. 41, no. 3, pp. 312–347, 2022.

[6] S. Sun and N. Figueroa, “Se (3) linear parameter varying dynamical systems for globally asymptotically stable end-effector control,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5152–5159, IEEE, 2024.

[7] B. Fichera and A. Billard, “Linearization and identification of multiple-attractor dynamical systems through laplacian eigenmaps,” *Journal of Machine Learning Research*, vol. 23, no. 294, pp. 1–35, 2022.

Can Causal Models Learn Robust Representations from Reservoir State Dynamics?

Gengcheng Lyu

Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

Abstract

Reservoir Computing, particularly Echo State Networks [1], excels at modeling complex temporal dynamics but suffers from limited interpretability and poor out-of-distribution generalization. Its rich, high-dimensional reservoir states are complex nonlinear embeddings of the input history, yet their internal structure remains largely opaque. This research proposes a novel approach to enhance the robustness and interpretability of RC by leveraging Causal Representation Learning [2] techniques on the reservoir state dynamics themselves.

References

[1] H. Zhang and D. V. Vargas. A survey on reservoir computing and its interdisciplinary applications beyond traditional machine learning. IEEE Access, 11, 81033-81070, 2023.

[2] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva. Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475, 2022.

GMLVQ for Tracking of Neurodegenerative Disease Progression in Prodromal Stages

Sofie Lövdal[1, 2], Michael Biehl[1], and the REMPET consortium*

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Nijenborgh 9, Groningen, 9747AG, Netherlands
[2]University Medical Center Groningen, Department of Nuclear Medicine and Molecular Imaging, Hanzeplein 1, Groningen, 9713GZ, Netherlands

Abstract

Neuroimaging with [^{18}F]Fluorodeoxyglucose Positron Emission Tomography ([^{18}F]FDG PET) models the glucose metabolism of the brain, and is a biomarker of neurodegeneration by reflecting the functional state of neurons. Isolated REM-sleep behaviour disorder (iRBD) is a strong indicator of prodromal Parkinson's disease (PD), dementia with Lewy bodies (DLB) or multiple system atrophy (MSA). Interpretable machine learning through Generalized Matrix Learning Vector Quantization (GMLVQ) offers a way to robustly model disease progression in this setting [1]. We considered [^{18}F]FDG PET scans of n = 49 iRBD patients, having undergone imaging two or three times with several years in between. We extracted feature vectors from the images using principal component analysis, whereafter a GMLVQ model was trained to classify HC, PD, DLB and MSA based on a set of training patients. Scanner effects in the data space were harmonized with Iterated Relevance Matrix Analysis [2]. We grouped the iRBD patients according to their clinical status, projected them into the trained GMLVQ space, and evaluated their trajectory over time. Subjects who converted during follow-up showed a steady and progressive trajectory from healthy towards the PD and DLB decision space over time. No difference between PD and DLB converters could be observed. A stable metabolic profile in HC space was associated with low risk for short-term conversion.

References

[1] R. van Veen, S.K. Meles, R.J. Renken et al. FDG-PET combined with learning vector quantization allows classification of neurodegenerative diseases and reveals the trajectory of idiopathic REM sleep behavior disorder. *Computer Methods and Programs in Biomedicine*, Art. No. 225-107042, 2022.

[2] S.S. Lövdal, R. van Veen, G. Carli, et al. IRMA: Machine learning-based harmonization of ^{18}F-FDG PET brain scans in multi-center studies. *European Journal of Nuclear Medicine and Molecular Imaging*, Art. No. 52-8, 2025.

Feature Relevance and Robustness in Biologically-Informed Classification Learning

Julius Voigt, Marika Kaden, and Thomas Villmann

SICIM, University of Applied Sciences Mittweida, Germany

Microgravity experiments on living cells are very complex and costly. Consequently, Machine Learning practitioners have access to only a small number of data points, which are, however, very high-dimensional. To make matters worse, the inputs are very information-sparse, as they contain many zeros. This almost inevitably leads to overfitting and unstable learning behaviour. Attempting to remedy this, one can reduce the complexity of the model by cutting down on parameters to be learned. This can be achieved by incorporating biological prior knowledge into the model architecture, and further reduced by an informed reduction of the inputs through analysis of feature relevance.

The Importance of Relevance: Combining Boruta with GMLVQ

Roland J. Veen[1,2], Michael Biehl[1]

1- Univ. of Groningen, Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Groningen, The Netherlands
2- Medical Research Council Laboratory of Medical Sciences (MRC LMS) London, United Kingdom

Abstract

Introduction In many areas, but notably medicine, datasets can contain many features, not all of them important or useful. The Boruta algorithm [1, 2] was created as an extension of Random Forests [3] to select the truly important features of a dataset. Random forests calculate the importance score for each feature, but this is not enough to also identify the features that become significantly important in interaction with other features. Boruta, as the spirit of the forest, solves this by adding even more randomness. All features are copied and permuted across all objects to serve as shadow features. Thus, these are not correlated with the decision problem. Subsequently, only features that are more important than their shadow copy can be considered important.

Generalized Matrix Learning Vector Quantisation (GMLVQ), [4], is a classification algorithm that, through the diagonal of its relevance matrix, gives a relevance for each feature. By using the relevances as importances, we can substitute Random Forests.

Methods We re-implemented the Boruta package, which was originally coded in R, later in Python, in MATLAB. Our implementation offers a choice of using Random Forests through the TreeBagger of the Statistics and Machine Learning Toolbox [5], or our implementation of GMLVQ [6]. We applied this method to benchmark and real-world medical data sets.

Results Some modifications were needed to go from relevance to importance, since the sum of the relevances is always normalized to 1. By comparing a feature's relevance only to its shadow feature, as done in [7], and delayed scoring by hit counting, we can successfully integrate GMLVQ as an importance provider for Boruta.

Conclusion GMLVQ can serve as a robust alternative to Random Forests for feature selection. GMLVQ is more computationally expensive than Random Forests, but can achieve comparable results with minor parameter tuning while preserving the robustness of GMLVQ. Further work would include proper cross-validation of the performance and the extension of the GMLVQ importance provider with feature bagging and statistical significance tests. A public release of the MATLAB Boruta toolbox is planned.

References

[1] Kursa MB, Jankowski A, Rudnicki WR. *Boruta - A System for Feature Selection.* Fundamenta Informaticae. 2010;101(4):271-285. https://doi.org/10.3233/FI-2010-288

[2] Kursa MB, Rudnicki WR. (2010). *Feature Selection with the Boruta Package.* Journal of Statistical Software, 36(11), 1-13. https://doi.org/10.18637/jss.v036.i11

[3] Breiman L. "Random Forests." Machine Learning 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324.

[4] Schneider P, Biehl M, Hammer B. *Adaptive Relevance Matrices in Learning Vector Quantization.* Neural Computation 21: 3532-3561 (2009)

[5] The MathWorks, Inc. (2022). *Machine Learning and Statistics Toolbox.* Available: https://www.mathworks.com

[6] Veen RJ, Westerman F, Biehl B. *A no-nonsense beginner's tool for GMLVQ.* https://www.cs.rug.nl/~biehl/gmlvq.html

[7] Manikandan G, Pragadeesh B et al. *Classification models combined with Boruta feature selection for heart disease prediction.* Informatics in Medicine Unlocked, 44, January 2024.

Some Insights on Multi-Perspective Learning with Indefinite Similarity Measures

Peter Maria Preinesberger

Faculty of Computer Science and Business Information Systems
Technical University of Applied Sciences Würzburg-Schweinfurt
peter.preinesberger@study.thws.de

Abstract

In data-scarce scenarios, supervised learning algorithms based on similarity measures play a pivotal role at integrating available domain expert knowledge for the prediction task [6, 7]. Unfortunately, many of these similarity measures violate the properties of valid dot products in one way or the other, obliterating theoretical properties of many methods [6], and requiring strategies for treating indefinite kernels [3]. Another dimension of difficulty is the fact that often, multiple similarity measures capture different aspects of the task, necessitating their fusion with Multiple Kernel Learning approaches [2, 1]. At the intersection of learning with indefinite kernels and Multiple Kernel Learning lies a bouquet of methods designed to deal with some or all of these complications [4, 5, 6, 3, 1], all with different principal approaches at tackling the problem. The presentation focuses on discussing the available approaches, highlighting differences as well as common aspects and pointing out problems and future research directions.

References

[1] Fabio Aiolli and Michele Donini. “EasyMKL: A scalable multiple kernel learning algorithm”. In: *Neurocomputing* 169 (2015).

[2] Mehmet Gönen and Ethem Alpaydin. “Multiple kernel learning algorithms”. In: *JMLR* 12 (2011).

[3] Ronny Luss and Alexandre d’Aspremont. “Support vector machine classification with indefinite kernels”. In: *Math. Program. Comput.* 1.2-3 (2009).

[4] Maximilian Münch et al. “Static and adaptive subspace information fusion for indefinite heterogeneous proximity data”. In: *Neurocomputing* 555 (2023).

[5] Peter Preinesberger, Maximilian Münch, and Frank-Michael Schleif. “Multiclass Adaptive Subspace Learning”. In: *33rd ESANN 2025*. 2025.

[6] Frank-Michael Schleif and Peter Tiño. “Indefinite Proximity Learning: A Review”. In: *Neural Comput.* 27.10 (2015).

[7] Philipp Väth et al. “PROVAL: A framework for comparison of protein sequence embeddings”. In: *Journal of Comp. Math. and Data Science* 3 (2022).

Metric Learning for k-NN

Marc Strickert

Justus Liebig University Gießen

Abstract

Metric learning for k-nearest neighbor classification is used to enhance the classification performance in terms of optimum confusion matrix properties, such as the predictive performance (F-score). In order to achieve that goal, the common discrete error counting is replaced by a mechanism based on continuous soft-ranked distances for class-(mis-)matching nearest neighbors. The derivative of the continuous-valued approximated F-score function w.r.t. adaptive metric parameters is used to improve the F-score by quasi-Newton optimization. For Mahalanobis type metrics, the result can be used for affine transformation of the original feature space. Hence, the proposed method provides an optimum data transformation for k-NN with confusion matrix criteria.

1 Introduction

Feature space transformation can be used to enhance the "clumping" of labeled point clouds in terms of better clustering and classification. By projection on only the "good" dimension, this is obvious for data with one perfectly separating data feature and another with uniform noise. Less extreme data configurations can also profit.

The original k-NN classifier does not require a training phase and it is a very easily interpretable model by relating to the class majority of k-nearest data points of a given pattern. Metric learning keeps the interpretability and may even lead to further simplifications of data configurations in low-dimensional projections. k-NN performance may not compete with state-of-the-art classifiers, but the proposed model allows for easy exchange of classifier behavior regarding versatile performance measures based on the class confusion matrix [1].

2 Methodology

The class confusion matrix with predicted conditions in columns and actual conditions in rows is a powerful tool to express many interesting properties of a classifier. If conditions (class labels) are called positive (P) and negative (N), the intersection cardinality of actual and predicted positives (hits) is the number of so called true positives (TP). Likewise, the true negatives (TN) reflect correct rejections. An actual true positive being predicted negative, a miss, is a false negative (FN), while the actual negative being predicted positive is a false positive (FP) aka false alarm. Using these matrix elements, for example, the well-known accuracy is just expressed as $ACC = (TP + TN)/N$, where $TP = \#(\mathrm{TP})$ is the number of true positives, $TN = \#(\mathrm{TN})$, and N is the number of classified data points. Other common measures are the False Discovery Rate and the Jaccard index. Here the focus is put on the F1-Score [2] which is the harmonic mean of precision $TP/(TP + FP)$ and recall $TP/(TP + FN)$, i.e.:

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \,. \tag{1}$$

Measures related to the confusion matrix involve the integer counting operator $\#(\cdot)$ of class matches and class mismatches and thus impose challenges on optimization procedures. Especially coordinate transformations based on continuous parameters are difficult to tune with discrete jumps in the F_1-response and partially flat cost function domains. High-performance optimization utilizing gradients of the cost function requires continuous formulations along the whole chain of nested computations. Besides counting another discrete entity, the concept of "nearest neighbors" in feature space, has to be replaced by soft ranking for k-NN classification. The metric-optimal k-NN classifier is implemented using these steps:

1. Compute $N \times N$ distance matrix $\mathbf{D}^{\boldsymbol{\lambda}}_{\mathbf{X}}$ of M-dimensional data $\mathbf{X}$ projected by $M \times q$ parameter matrix $\boldsymbol{\lambda}$ to an q-dimensional (sub-)space.
2. Soft ranking of the distance matrix $\mathbf{D}^{\boldsymbol{\lambda}}_{\mathbf{X}}$.
3. Sigmoidal transformation of soft ranks: $\mathrm{sgd}_{k,\iota}(r) = 1 - (1 + e^{-\iota(r-k-1/2)})^{-1}$ which for $\iota \geq 4$ is ≈ 1 for $r \leq k$ and ≈ 0 for $r > k$, where k refers to k-NN neighborhood size. The "logistic" result indicates k-neighborhood (specifically, soft neighborliness).
4. Conditioned summation of these indices for each reference data point: if the reference data point belongs to class P and the class of the comparison point is also P the summation of its neighborliness index contributes to TP. Two points of class N sum up TN. A reference point class P with a comparison point class N leads to summation of FN, and vice versa for FP.
5. For each point each of its four above sums s is squashed sigmoidal by $-\mathrm{sgd}_{k/2,\iota}(s)$ with turning point at $k/2$ and typical squashing values of $\iota = 100$. This leads to exclusive contribution of ≈ 1 to either TP, FN, FP, or TN, i.e. to specific point characterization with respect to the type of class match or mismatch in its neighborhood context.
6. Based on summation of these ratings for all points create the total soft confusion matrix $\left(\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right)$ as soft counts of TP, TN, FP, and FN.
7. Calculate the soft F1-Score using Eqn. 1 and take the negative as cost function value.

For two input vectors $\mathbf{x}^i$ and $\mathbf{x}^j \in \mathbf{X}$ being column vectors their adaptive distance is defined as

$$(\mathbf{D}^{\boldsymbol{\lambda}}_{\mathbf{X}})^{1/2}_{i,j} = \sqrt{(\mathbf{x}^i - \mathbf{x}^j)^{\mathsf{T}} \cdot \boldsymbol{\Lambda} \cdot (\mathbf{x}^i - \mathbf{x}^j)} \quad \text{with} \quad \boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}} \,. \tag{2}$$

This expression looks like the Mahalanobis distance. Since the square root does not affect the nearest neighbor configurations, this operation can be omitted and $\mathbf{X}^{\mathsf{T}} \cdot \boldsymbol{\lambda}$ be considered as data projection for which the pairwise squared Euclidean distance can be used.

Instead of utilizing a sorting operation, for nearest neighbor identification, the ranking of distances in column vectors $\boldsymbol{u}$ of $\mathbf{D}_{\mathbf{X}}^{\boldsymbol{\lambda}}$ can be alternatively achieved by summing up rows of the indicator matrix $\boldsymbol{R}$:

$$\mathsf{rnk}(\boldsymbol{u}) = \begin{pmatrix} \sum_{i=1}^{N} \mathsf{R}(\boldsymbol{u}_1, \boldsymbol{u}_i) \\ \dots \\ \sum_{i=1}^{N} \mathsf{R}(\boldsymbol{u}_N, \boldsymbol{u}_i) \end{pmatrix} \quad \text{for} \quad \boldsymbol{R}(\boldsymbol{u}) = \begin{pmatrix} \mathsf{R}(\boldsymbol{u}_1, \boldsymbol{u}_1) & \dots & \mathsf{R}(\boldsymbol{u}_1, \boldsymbol{u}_N) \\ & \dots & \\ \mathsf{R}(\boldsymbol{u}_N, \boldsymbol{u}_1) & \dots & \mathsf{R}(\boldsymbol{u}_N, \boldsymbol{u}_N) \end{pmatrix} . \tag{3}$$

For the Heaviside step function $\mathsf{R}(\boldsymbol{u}_k, \boldsymbol{u}_l) = H(\boldsymbol{u}_k - \boldsymbol{u}_l)$, providing zero for negative arguments and else one, correct ranks are obtained for vector elements $\boldsymbol{u}_k$ in the absence of ties. Using the standard deviation $\sigma_{\boldsymbol{u}}$

$$\sigma_{\boldsymbol{u}} = \Big(\frac{1}{N-1} \cdot \sum_{i=1}^{N} (\boldsymbol{u}_i - \mu_{\boldsymbol{u}})^2 \Big)^{1/2} \tag{4}$$

the step function $H(\boldsymbol{u}_k - \boldsymbol{u}_l)$ can be replaced by a differentiable sigmoid

$$\mathsf{R}(\boldsymbol{u}_k, \boldsymbol{u}_l) = \mathsf{sgd}_{\kappa}^{kl} + \frac{1}{2} = \mathsf{sgd}_{\kappa}\Big(\frac{\boldsymbol{u}_k - \boldsymbol{u}_l}{\sigma_{\boldsymbol{u}}} \Big) + \frac{1}{2} = \frac{1}{1 + e^{\kappa \cdot (\boldsymbol{u}_k - \boldsymbol{u}_l)/\sigma_{\boldsymbol{u}}}} + \frac{1}{2} \tag{5}$$

with mid-tied ranks being approximated for $\kappa \to \infty$. In practice, $5 < \kappa < 100$ is numerically adequate. Since each column distance matrix triggers an $N \times N$ indicator matrix, the overall computational costs are $\mathcal{O}(N^3)$; this is certainly huge, considering that only $k \times N$ ranks are effectively needed for the nearest neighbors.

Optimization is done with the Broyden-Fletcher-Goldfarb-Shanno BFGS-algorithm [3]. The derivatives of the nested cost function (see seven steps above) employed therein as product of several Jacobian matrices would exceed the scope of the present text. Interested readers can refer to previous work [4] and to the MATLAB/GNU-Octave code [5]. Validations were carried out by numeric gradient approximations using DERIVESTsuite [6].

3 Matrix Initialization

The simplified expression $\boldsymbol{x}^{\mathsf{T}} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{x}$ in Eqn. 2 describes the mixing of the k-th and m-th attribute of $\boldsymbol{x}$ by the matrix components $(\boldsymbol{\Lambda})_{km}$. The identity matrix $\boldsymbol{\Lambda} = \boldsymbol{E}$ implements the Euclidean norm of $\boldsymbol{x}$ with independent attribute contributions, while $\boldsymbol{\Lambda} = \frac{1}{2} \cdot (\boldsymbol{E} + \mathbf{1})$ leads to equal contributions of all attribute pairs.

Without prior knowledge both choices are good options for starting the matrix adaptation in an unbiased way. Yet, $\boldsymbol{\lambda}$ gets adapted, not $\boldsymbol{\Lambda}$, which, leads to low ranks of $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ and inability to express the identity matrix. This discrepancy can be minimized by an optimized initialization matrix, where gradient descent is used on an initially random matrix $\boldsymbol{\lambda}_0$ with the cost function

$$S = \Big\| \boldsymbol{\lambda}_0 \cdot \boldsymbol{\lambda}_0^{\mathsf{T}} - \frac{1}{2} \cdot (\boldsymbol{E} + \mathbf{1}) \Big\|_{\mathsf{F}}^2 . \tag{6}$$

Therein, the squared Frobenius norm $\| \cdot \|_{\mathsf{F}}^2$ is used to express a minimum least square approach for optimizing the initial matrix elements. The converged matrix is used to initialize $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ for metric adaptation. This way the rank mismatch is distributed equally over all data attributes and, consequently, the sum of all mixing coefficients per attribute.

4 Matrix Interpretation

For the interpretation of the finally obtained parameter matrix $\boldsymbol{\lambda}$ it is more natural to look at the mixing matrix elements in $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top}$. Basically, large absolute values $|\Lambda_{ij}|$ denote "important" contributions of attribute pairs i, j to the given association task. Since the covariance values of the attributes affect the magnitude of the mixing factors, a rescaled mixing matrix without covariance structure is obtained by inserting $\boldsymbol{K}$ into the central expression of Eqn. 2 to $\boldsymbol{x}^{\top} \cdot [\boldsymbol{K} \cdot (\boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top}) \cdot \boldsymbol{K}^{\top}] \cdot \boldsymbol{x}$, where $\boldsymbol{K} \cdot \boldsymbol{K}^{\top} = \mathsf{cov}(\mathbf{X})$, i.e. $\boldsymbol{K} = \mathsf{cov}(\mathbf{X})^{1/2}$. Equivalently, $\mathsf{cov}(\mathbf{X})^{1/2}$ can be multiplied to $\boldsymbol{\lambda}$ prior to calculating $\boldsymbol{\Lambda}$.

Alternatively, the influence of independent attribute variances, disregarding further covariance structure, can be removed by multiplying the k-th row of $\boldsymbol{\lambda}$ by the standard deviation of the k-th data attribute prior to calculating $\boldsymbol{\Lambda}$.

5 Experiments

5.1 Toy Example

First, an artificial dataset is generated with the following properties: two classes in five normal distributed data clusters of 7, 10, 9, 4, and 12 points around the coordinates $(-1, 0), (1, -1), (0, 0), (1, 1), (2, -1)$ with class labels P, P, N, N, N, respectively. Three more normal distributed attributes are added to create a 5-dimensional dataset of which the first two are meaningful. The cluster configuration features an XOR-like arrangement for which a linear class separation is impossible. 3-NN is employed.

The initial configuration, a simple projection to the X-Y-plane, and a typical optimization result are shown in figure 1. The inset whitened parameter matrix $\boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top}$ highlights the first two attributes as important, which is in accordance with the original dataset design. The top-left 2x2 matrix elements indicate not only original relevances, though, but also a rotation of the first two dimensions by about 45° (plus meaningless horizontal flip) that allows for a perfectly class-separating projection onto a single axis; thereby, some useful features of noise dimensions are utilized as well, revealed as weak mosaic pattern of the first two columns and rows.

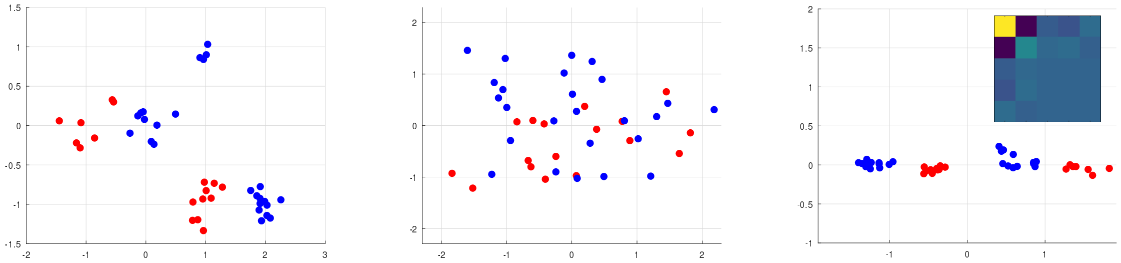


Figure 1: Left: Toy example, generating first two dimensions. Center: random projection. Right: optimized projection subspace with the final parameter matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top}$ as inset with nothing hidden behind.

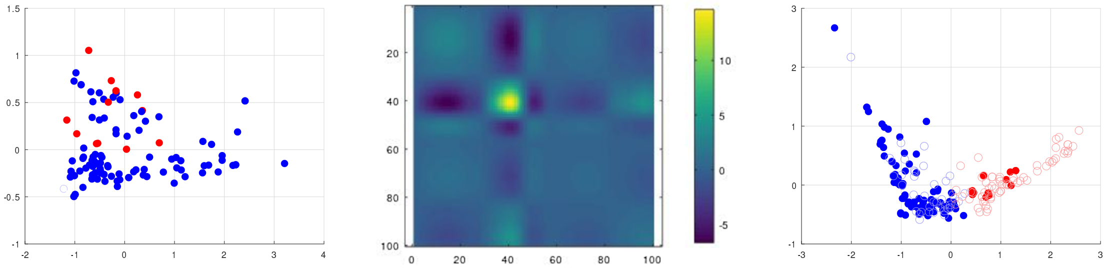
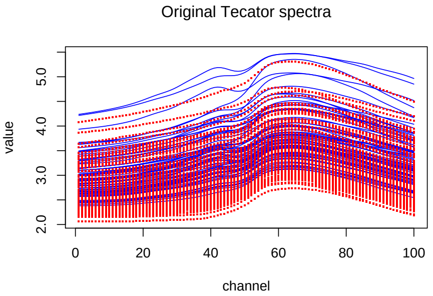


Figure 2: Left: Tecator training data projection by initial $\boldsymbol{\lambda}$. Center: final parameter matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ with 100 dimensions. Right: projection of training data (filled circles) and test data (open circles) by final $\boldsymbol{\lambda}$.

5.2 Tecator



The Tecator spectral data set is taken from the UCI repository of machine learning. It contains 215 samples of 100-dimensional infrared absorbance spectra recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050nm by the Near Infrared Transmission (NIT) principle. 138 analysed meat samples have a low (red label P) fat and 77 high (blue label N) fat content for being classified based on their spectra, displayed in the right figure. Low fat content is reflected by red dashed lines, high fat content by blue solid lines. For illustration purposes, 13 P-samples and 100 N-samples were used for metric-tweaking a 5-NN classifier in 2 dimensions.

Figure 2 show the projection results. Configuration for initial $\boldsymbol{\lambda}$ is show in the left, the optimized metric $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ in the center and the data projection by the final $\boldsymbol{\lambda}$ in the right panel. The training confusion matrix for an ordinary (discrete) k-NN classifier applied to the transformed 2D-data space matrix is $\left(\begin{smallmatrix}13 & 0\\ 0 & 100\end{smallmatrix}\right)$. This corresponds to a perfect F1-score of 1 and its soft cost function counterpart of -0.9999. For the test data (open circles) the confusion matrix $\left(\begin{smallmatrix}63 & 1\\ 3 & 35\end{smallmatrix}\right)$ yields the accuracy value of 0.961 and an F1-score of 0.969. The trained metric highlights channel numbers around 40 as important, corresponding to the "bump" of the blue spectral curves.

5.3 Ring Classification

Particle identification is one of the ultimate goals in high-energy physics and detector design. Many methods use indirect phenomena such as the Cherenkov effect: charged particles in transparent media (such as gas, aerogel or fused silica) faster than the light propagation therein create a cone-like wave front with characteristic opening angle (cf. sonic boom). Usually, only few light photons (≈100) are created. Quantum efficiency leaves sparse elliptic patterns on a matrix of single photon detectors with around 7–50 dots to be used for ring finding [7]. The ring-imaging Cherenkov detector (RICH) for compressed baryonic matter (CBM), currently being developed at the Universities of Gießen and Wup-

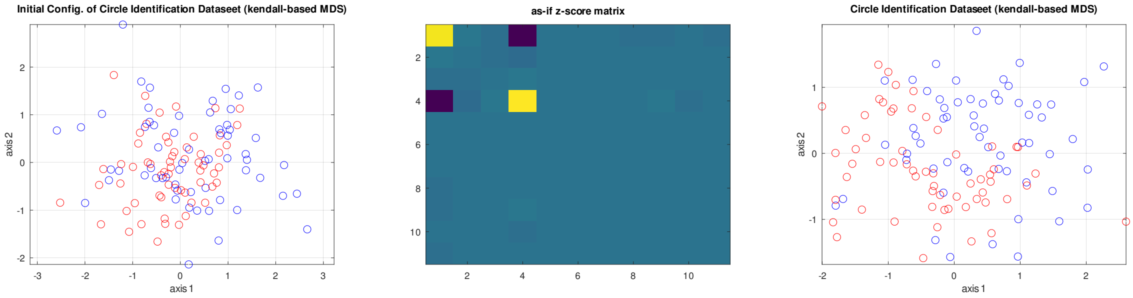


Figure 3: Left: Circle training data projection by initial $\boldsymbol{\lambda}$. Center: final parameter matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ with 11 dimensions. Right: projection of training data by $\boldsymbol{\lambda}_{\text{final}}$.

pertal, highlights a sophisticated simulation and particle detection analysis pipeline. An annotated ring dataset with $6 \cdot 10^5$ 11-dimensional feature vectors is available. Besides intuitive data features like "radius" and "center coordinate", less intuitive values like ellipse fitting parameters are included to predict whether a point configuration is part of a particle-induced ring structure. 5-NN with 8-dimensional subspace projection is used.

Figure 3 show the projection results. Configuration of training data projected by initial $\boldsymbol{\lambda}$ is show in the left panel as neighborhood-preserving 2D-embedding by correlation-based multidimensional scaling using Kendall correlation (cbMDS,[4]). The optimized parameter matrix in the center indicates that only two out of eleven features are relevant for k-NN classification. Since implicit whitening is used, this specific attribute emphasis is not related to scale. The 2D-cbMDS embedding of final projection of training data in the right shows how the initial central density and class confusion is counteracted. Note that cbMDS is much better in preserving data neighborhood than PCA, but the true clustering is only found in the 8-dimensional projection subspace. Thus, false data overlaps and false separations might still be displayed.

The metric learning was based on only 120 randomly drawn samples (0.2 ‰ of the data). The initial accuracy of 0.767 and F1-score of 0.788 increases to 0.942 and 0.945, respectively. However the test set reaches values of only 0.690 and 0.707, respectively, while boosted decision tree reaches 0.79 for both values. Test values so much lower than for the training data indicate that the random sub-sampling destroyed characteristic neighborhood configurations. Without further data "summarization" such as vector quantization, this is a general limitation of the proposed method: on the one hand, a small number of training data suffices to characterize a global feature space transformation; on the other hand, random sampling might not represent data structure and outliers well enough. And, as expected, the proposed data projection method cannot compete with modern classifiers.

6 Considerations and Conclusions

The inherently discrete task of error counting in nearest neighbor configurations was approximated by nested soft formulations based on the sigmoid ("logistic") function. Like in artificial neural network designs, a cost function gradient was used to optimize the desired parameters, here, the linear projection parameters of the feature space in order to enhance the clustering of class-consistent nearest neighbors.

To get things working, double precision floating point numbers are recommended, because of the small but necessary gradients of sigmoids. Due to accuracy problems and also because of time demands a numeric approximation of the gradient is no valid option for the described optimization task. Analytic gradient calculations impose some memory limitations though: the largest Jacobian matrix for the soft ranking matrix requires $\mathcal{O}(N^3)$ bytes which adds up to about 8 gigabytes in double floating point representation for only 1000 data points – and this is only one factor in the total cost function derivative. Also, even on modern computers such matrix multiplications take some time.

The optimization result clearly differs from linear discriminant analysis or from learning vector quantization. It shares their idea of simplicity though, because linear data projections and nearest neighbors allow for intuitive model interpretation. Also, the use of cost function with whatever differentiable can be computed from the values of the confusion matrix, such as accuracy and F1-score, is a very appealing property of the proposed method.

Acknowledgments

The author would like to thank Prof. Dr. Annika Thiel, II Physics Institute at Justus Liebig University, for this freedom of work and Martin Beyer (CBM-RICH group) at II Physics Institute, for the circle data set, discussion, and results. No AI was used for manuscript preparation.

References

[1] Wikipedia (English), Confusion matrix. https://en.wikipedia.org/wiki/Confusion_matrix

[2] Wikipedia (English), F-Score. https://en.wikipedia.org/wiki/F-score

[3] Dirk-Jan Kroon, FMINLBFGS: Fast Limited Memory Optimizer. MATLAB Central File Exchange (last seen 31 Juli 2025). https://www.mathworks.com/matlabcentral/fileexchange/23245-fminlbfgs-fast-limited-memory-optimizer

[4] M. Strickert, K. Bunte, F.-M. Schleif, and E. Hüllermeier, Correlation-based embedding of pairwise score data. *Neurocomputing*, 141:97–1009, 2014. https://doi.org/10.1016/j.neucom.2014.01.049

[5] M. Strickert, Metric learning k-NN implementation in MATLAB/GNU-Octave. https://jlubox.uni-giessen.de/getlink/fiT43d8Yb7sQAFb33357x16d/knnmetric.zip

[6] John D'Errico, Adaptive Robust Numerical Differentiation. MATLAB Central File Exchange (last seen 31 July 2025). https://www.mathworks.com/matlabcentral/fileexchange/13490-adaptive-robust-numerical-differentiation

[7] S. Lebedev, C. Hoehne, G. Ososkov, for the Cbm Collaboration et al., Ring recognition and electron identification in the RICH detector of the CBM experiment at FAIR. Journal of Physics: Conference Series (IOP Publishing, 2010), Vol. 219, p. 032015. https://doi.org/10.1088/1742-6596/219/3/032015

Fairness in the Flow - Building Better Benchmarks for Fair Stream Learning

Kathrin Lammers

Bielefeld University

31.07.2025

Abstract

When machine learning algorithms make decisions about human beings, such as when they are used for pre-selecting job applications or assessing risk of fraud in social security administration, fairness is a crucial aspect that demands consideration [5]. The EU AI Act requires that algorithmic decision-making be free from discrimination and unfair biases [1], for example. This does not only hold true for the standard batch set-up of machine learning, but also and especially when working with non-stationary data streams.

Fairness in stream learning, however, is currently an under-explored topic. While multiple algorithms have been proposed in recent years, some claiming to address issues of fairness, concept drift, and class imbalance simultaneously, their evaluation is currently somewhat limited due to a lack of suitable benchmark data streams [4] and a focus on individual fairness metrics, which were also directly optimized during training.

Our research tries to address these issues by proposing a novel framework for generating suitable benchmark streams based on currently used, real-world fairness benchmark datasets. To this end, we first extract the causal relations and feature codependencies from the original data by learning its Bayesian network. In an additional step, this Bayesian network is then manipulated to introduce concept drift - both with and without directly targeting fairness-relevant connections between features. Both specific biases and class imbalance can be either introduced or potentially mitigated during this step, which allows for the creation of realistic biased and streaming data to be used as benchmarks when evaluating fair stream learning classifiers.

Additionally, we provide a structured overview of the currently available fair stream-learning algorithms, amongst others, 2FAHT [6], FABBOO [2], and CFSMOTE [3]. We briefly characterize their core mechanisms - such as base models, addressed notions of fairness, and stage of fairness-related interventions (i.e., pre-, in-, or post-processing), and whether they are suitable for imbalanced data.

Finally, we generate a few example streams using this method - based on the *Adult* and *Portuguese Student* datasets [4] - and employ them to evaluate a select number of state-of-the-art fairness-aware stream learning algorithms on five key fairness metrics to investigate their performance and potential fairness trade-offs.

References

[1] Council of European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), June 2024. Legislative Body: CONSIL, EP.

[2] V. Iosifidis and E. Ntoutsi. *FABBOO* - Online Fairness-Aware Learning Under Class Imbalance. In A. Appice, G. Tsoumakas, Y. Manolopoulos, and S. Matwin, editors, *Discovery Science*, volume 12323, pages 159–174. Springer International Publishing, Cham, 2020.

[3] K. Lammers, V. Vaquet, and B. Hammer. Continuous Fair SMOTE – Fairness-Aware Stream Learning from Imbalanced Data, 2025.

[4] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairnessâaware machine learning. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1452, May 2022.

[5] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022.

[6] W. Zhang, M. Zhang, J. Zhang, Z. Liu, Z. Chen, J. Wang, E. Raff, and E. Messina. Flexible and Adaptive Fairness-aware Learning in Non-stationary Data Streams. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 399–406, Baltimore, MD, USA, Nov. 2020. IEEE.

There's a Bias in There

Marika Kaden, Ronny Schubert, Julius Voigt, Lynn Reuss, Alexander Engelsberger, Sofie Lövdal, Elina L. van den Brandhof, Michael Biehl, Thomas Villmann

SICIM, University of Applied Sciences Mittweida, Germany

Fairness in machine learning classification is a critical research area. Biased training data can lead to discriminatory outcomes. In the talk, we present a workflow for detecting and mitigating data bias using the interpretable, shallow machine learning model: Generalized Learning Vector Quantization (GLVQ). We use the latest extension of the Matrix GLVQ, the Iterated Relevance Matrix Analysis, to mitigate the influence of bias in the data for classification. We demonstrate the effectiveness of this approach in some examples, showing how the workflow can be used to detect and reduce bias, promoting fairer decision-making.

Partially Interpretable Embeddings enable Fairness through Awareness

Sarah Schröder

Bielefeld University

Abstract

Large language models are widely used in modern AI applications, yet concerns about fairness and bias persist. Recently, several works utilize concept erasure to mitigate bias, where sensitive attributes are entirely scrubbed from the language model or its representations to achieve fairness [1, 2, 3]. While there might exist cases where this improves fairness with a performance tradeoff, other works discussed conceptual issues like concept entanglement [4] or how blindness to sensitive attributes can actively harm the groups that were supposed to be protected from algorithmic bias [5].

Contrary to these approaches, we aim to enhance fairness through awareness of sensitive attributes. Instead of removing them, we investigate methods to identify such sensitive concepts in embeddings of large language models. These concepts are incorporated into a learning pipeline, where we leverage LLMs as language preprocessors, transform their embeddings into a partially interpretable feature space, where sensitive attributes are encoded by specific features, and finally stack a task specific model on top.

The premise is to actively use the sensitive attributes to (i) detect biases and investigate sources for biases (e.g. feature interactions, stereotypes), (ii) mitigate those biases under task-specific requirements and (iii) explain model decisions while highlighting how sensitive attributes are used.

Building upon [6], we investigate how sensitive attributes can be reliable predicted, considering challenges such as data availability and underrepresented groups. Furthermore, we demonstrate how the pipeline can be used to discover and mitigate bias in real world datasets.

References

[1] Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., & Biderman, S. (2023). Leace: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems

[2] Ravfogel, S., Twiton, M., Goldberg, Y., & Cotterell, R. D. (2022, June). Linear adversarial concept erasure. In International Conference on Machine Learning (pp. 18400-18421). PMLR.

[3] Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. 2023. Shielded Representations: Protecting Sensitive Attributes Through Iterative Gradient-Based Projection. In Findings of the Association for Computational Linguistics: ACL 2023

[4] Amara, I., Humayun, A. I., Kajic, I., Parekh, Z., Harris, N., Young, S., ... & Rostamzadeh, N. (2025). Erasebench: Understanding the ripple effects of concept erasure techniques. arXiv preprint arXiv:2501.09833.

[5] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The measure and mismeasure of fairness. JMLR

[6] Schroeder, S., Schulz, A., & Hammer, B. (2025). Evaluating Concept Discovery Methods for Sensitive Attributes in Language Models. ESANN 2025 proceedings.

Drift Explanations for System Monitoring: From Descriptive to Causal

Fabian Hinder

Faculty of Technology, Bielefeld University

Abstract

Real-world environments – whether industrial manufacturing, critical infrastructure, or online services – are anything but static. Here, the data-generating process tends to constantly evolve and change over time, a phenomenon known as *concept drift* (or drift, for short). Most commonly, drift is studied in the context of learning models, where it degrades performance over time. While there is a large body of literature on detecting and adapting to drift [1], the more human-centered question of *what exactly is changing, and why?* is usually left unanswered.

Drift explanations [2] aim to fill this gap by describing the often high-dimensional distributional changes in a human-understandable fashion. This way, drift explanations provide supporting information for model maintenance, system supervision, and monitoring even outside the learning domain.

In our recent work [3], we integrated drift explanation with *computational causality* [4]. By moving beyond purely descriptive explanations to causal ones, we gain explanations that are not only interpretable but actionable by pointing to interventions that mitigate the drift – guiding understanding or even uncovering upstream faults in physical systems.

References

[1] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[2] F. Hinder, V. Vaquet, and B. Hammer, "One or two things we know about concept drift—a survey on monitoring in evolving environments. part b: locating and explaining concept drift," *Frontiers in Artificial Intelligence*, vol. 7, p. 1330258, 2024.

[3] D. Komnick, K. Lammers, , B. Hammer, V. Vaquet, and F. Hinder, "Causal explanation of concept drift – a truly actionable approach," *arXiv preprint*, 2025.

[4] J. Pearl, *Causality.* Cambridge university press, 2009.

On Competitive Networks

Ronny Schubert

University of Applied Sciences Mittweida - Hochschule Mittweida,
Saxon Institute of Computational Intelligence and Machine Learning,
Germany

September 1, 2025

Abstract

Competitive networks have been investigated at least since the 1970's and are tightly linked to the origins of mathematical formulations of neural networks [RZ85]. The neurons of such a network inhibit the activity of entities within their level or layer which creates a competition and gives this kind of network its name. The biological paragon for the design and dynamic are synaptic inhibition and the presence of interneurons, which are then modeled either as lateral connections or mathematical interneurons [KK94]. In a more recent context, biological inhibition has been studied more extensively and research suggests, that such phenomenons play a key role in homeostatic plasticity and learning regulation [Bar21; GK25]. An example of a formal realization are *Winner-takes-All* (*WTA*) networks, e.g., [MEA88; KK94], which, due to their formulation and recurrent nature, are in relation to Hopfield networks. Another example are *ART* networks based on *Adaptive Resonance Theory* [Gro76a; Gro76b]. In this regard, we will elaborate in more depth about these networks and the potential when considering other networks in the perspective of competitive networks.

References

[Bar21] Helen C Barron. “Neural inhibition for continual learning and memory”. In: *Current Opinion in Neurobiology.* Neurobiology of Learning and Plasticity 67 (Apr. 2021), pp. 85–94. ISSN: 0959-4388. DOI: 10.1016/j.conb.2020.09.007.

[GK25] Elisa Galliano and Tara Keck. “Interactions between homeostatic plasticity and statistical learning: A role for inhibition”. In: *Current Opinion in Neurobiology* 93 (Aug. 2025), p. 103065. ISSN: 0959-4388. DOI: 10.1016/j.conb.2025.103065.

[Gro76a] S. Grossberg. "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors". en. In: *Biological Cybernetics* 23.3 (Sept. 1976), pp. 121–134. ISSN: 1432-0770. DOI: 10.1007/BF00344744. URL: https://doi.org/10.1007/BF00344744.

[Gro76b] Stephen Grossberg. "Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions". en. In: *Biological Cybernetics* 23.4 (Dec. 1976), pp. 187–202. ISSN: 1432-0770. DOI: 10.1007/BF00340335. URL: https://doi.org/10.1007/BF00340335.

[KK94] Samuel Kaski and Teuvo Kohonen. "Winner-take-all networks for physiological models of competitive learning". In: *Neural Networks*. Models of Neurodynamics and Behavior 7.6 (Jan. 1994), pp. 973–984. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(05)80154-6. URL: https://www.sciencedirect.com/science/article/pii/S0893608005801546.

[MEA88] E. Majani, Ruth Erlanson, and Yaser Abu-Mostafa. "On the K-Winners-Take-All Network". In: *Advances in Neural Information Processing Systems*. Vol. 1. Morgan-Kaufmann, 1988. URL: https://papers.nips.cc/paper/1988/hash/6c4b761a28b734fe93831e3fb400ce87-Abstract.html.

[RZ85] David E. Rumelhart and David Zipser. "Feature discovery by competitive learning". In: *Cognitive Science* 9.1 (Jan. 1985), pp. 75–112. ISSN: 0364-0213. DOI: 10.1016/S0364-0213(85)80010-0. URL: https://www.sciencedirect.com/science/article/pii/S0364021385800100.

Dynamic Mode Decomposition meets Prototype-based Learning

Janis Norden

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Dynamic Mode Decomposition (DMD) is a tool to analyse multi-variate time series data. Originally developed in the fluid dynamics community, it was designed to extract physically meaningful flow patterns from experiment and simulation data [3, 4]. Outside of fluid dynamics, DMD has been a great success too: since its conception in 2008, DMD has been applied to problems ranging from neuroscience all the way to oceanography and robotics, to only name a few [1]. In principle, DMD is an unsupervised technique. However, *supervised DMD* [2] and *discriminant DMD* [5] are two recent variants which incorporate label information and have shown great promise to extract class-specific patterns, which may be used in subsequent classification tasks. In this talk, I will briefly introduce the basics of DMD, supervised DMD and discriminant DMD. Then, taking inspiration from these two supervised methods, I will outline an approach that brings together prototype-based learning and DMD, which may be able to alleviate some of the limitations inherent to supervised DMD and discriminant DMD.

References

[1] Steven L. Brunton and J. Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control.* Cambridge University Press, 2022.

[2] Keisuke Fujii and Yoshinobu Kawahara. Supervised dynamic mode decomposition via multitask learning. *Pattern Recognition Letters*, 122:7–13, 2019.

[3] Clarence W. Rowley, Igor Mezic, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[4] Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[5] Naoya Takeishi, Keisuke Fujii, Koh Takeuchi, and Yoshinobu Kawahara. Discriminant dynamic mode decomposition for labeled spatiotemporal data collections. *SIAM Journal on Applied Dynamical Systems*, 21(2):1030–1058, 2022.

Good by Default? Generalization in Highly Over-Parameterized Neural Networks

Thomas Martinetz

University of Lübeck

May 25, 2025

Abstract

Modern deep neural networks are often highly over-parameterized, containing many more parameters than available training samples. Surprisingly, even in this regime, such models tend to generalize well, achieving low test errors despite being trained to zero training loss and without explicit regularization. This behavior contradicts traditional expectations from statistical learning theory, which warn against overfitting in these situations. While previous explanations focused on *implicit regularization* effects (e.g., from stochastic gradient descent), we study an alternative hypothesis: good generalization arises because bad solutions become intrinsically rare in the solution space [1].

Main result

We show that under certain conditions, the fraction of global minima with poor generalization (that is, large true error) among all zero-training-error solutions vanishes exponentially with the number of training samples n. This offers an alternative perspective: even without regularization, over-parameterized models are likely to land on a good solution, simply because bad ones are scarce.

We look at classification settings. Classifiers h from an hypothesis set $\mathcal{H}$ are evaluated based on their true error $E(h)$ and empirical error $E_{\mathcal{S}}(h)$.

- $\mathcal{H}(\mathcal{S}) \subseteq \mathcal{H}$: Subset of classifiers achieving $E_{\mathcal{S}}(h) = 0$ on training set $\mathcal{S}$.
- $\mathcal{H}_\varepsilon(\mathcal{S}) \subseteq \mathcal{H}(\mathcal{S})$: Subset of $\mathcal{H}(\mathcal{S})$ with $E(h) > E_{\min} + \varepsilon$ (bad solutions).
- $D(E)$: *Density of classifiers* (DOC), i.e., the distribution of true errors over the hypothesis space.

Under certain conditions, the expected fraction of "bad" zero training error solutions within the zero training error solution set decays exponentially:

$$\begin{aligned} \mathbb{E}_S\left[\frac{|\mathcal{H}_\varepsilon(\mathcal{S})|}{|\mathcal{H}(\mathcal{S})|}\right] &\leq \frac{\int_{E_{min}+\varepsilon}^{1}(1-E)^n D(E)\, dE}{\int_0^1 (1-E)^n D(E)\, dE} \\ &\leq \frac{1}{g_{\varepsilon/2}}\, e^{-\frac{\varepsilon}{2}n}. \end{aligned}$$

This bound depends on the shape of $D(E)$, not on the model size or parameter count. $g_{\varepsilon/2}$ denotes the overall fraction of good solutions within the hypotheses set.

Empirical confirmations

The theoretical results are validated by several experiments:

1. Synthetic Data: Two overlapping Gaussian distributions in 10-dimensional space are classified using neural networks with 120 and 1,200 weight parameters. Despite a ten-fold difference in parameter count, the corresponding Density of Classifiers $D(E)$ are remarkably similar. As predicted by the theory, the proportion of bad solutions decreases at nearly identical rates as the number of training samples n increases. Moreover, the empirically observed mean true errors closely match the theoretically derived values.
2. MNIST (Digits 1 vs. 2): Even for a network with 7,860 weights, bad solutions become rare with just a few dozen training samples. Once again, the empirically observed mean true errors align closely with the theoretical predictions.
3. VGG19 and ResNet18 on Caltech101: Randomly sampled zero-training-error solutions of VGG19 (140M parameters) and ResNet18 (11M) begin to generalize with as few as 10 to 20 training examples per class. Remarkably, VGG19 demonstrates better generalization despite its significantly larger size.

In all the experiments, the theoretical bounds match closely with empirical measurements. The shape of $D(E)$ determines how likely a bad solution is.

Conclusion

The results challenge the conventional view that generalization requires explicit or implicit regularization. Instead, it argues that in over-parameterized settings, bad solutions are statistically rare. As a result:

- The geometry of the solution space, as captured by $D(E)$, plays a crucial role.
- Over-parameterization may *help* generalization by increasing the proportion of good solutions.
- Stochastic gradient descent does not need a strong bias towards good solutions if bad ones are rare to begin with.

These insights may open new directions for the theoretical understanding of deep learning and encourages further investigation into what structural properties of network models and data lead to favorable $D(E)$ distributions.

References

[1] J. Martinetz, T. Martinetz, Do Highly Over-Parameterized Neural Networks Generalize Since Bad Solutions Are Rare?, IEEE Transactions on Neural Networks and Learning Systems, 2025, 10.1109/TNNLS.2025.3529297.

The impact of over-parameterization in shallow feed-forward neural networks

Otavio Citton[1,2], Frederieke Richert[1], and Michael Biehl[1]

[1]Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands
[2]Groningen Cognitive Systems and Materials Center (CogniGron), University of Groningen, The Netherlands

August 1, 2025

Abstract

Neural Network (NN) architectures are fundamental pieces in the ongoing Artificial Intelligence (AI) revolution and, despite the huge advancements in the realm of applications, the theoretical understanding behind these models lags behind. Here we present our work that tries to shed some light on the behavior of over-parameterized shallow feed forward neural networks using techniques borrowed from Statistical Mechanics [1].

We consider a special type of two-layer NN known as Soft Committee Machines (SCM) which allows us to obtain analytical results for the equilibrium distribution over the parameter space after training for a long time [2, 3]. By considering a teacher network responsible for correct output of the examples, we are able to define a measure of the degree of over-parameterization and analyze how it impacts the performance of the student network. Moreover, our results hold for any activation function[4], allowing us to compare how different functions perform in this over-parameterized scenario.

References

[1] M. Mezard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications.* World Scientific, 1986.

[2] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45:6056–6091, Apr 1992.

[3] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, Apr 1993.

[4] O. Citton, F. Richert, and M. Biehl. Phase transition analysis for shallow neural networks with arbitrary activation functions. *Physica A: Statistical Mechanics and its Applications*, 660:130356, 2025.

Analysing the Stability of Feature Importance

S. Panda, M. Kaden, M. Karimi, and T. Villmann

SICIM, University of Applied Sciences Mittweida, Germany

Learning Vector Quantisation (LVQ) is an interpretable classification method. Despite being a relatively simple, shallow approach, LVQ often achieves performance comparable to that of more complex deep learning models. While many papers have demonstrated that GLVQ provides robust classification, this cannot be transferred to the interpretation of feature importance. We demonstrate through a number of examples that explanations derived from an LVQ model are often specific to that particular model and do not generalise well. This is a crucial consideration when using cross-validation to assess generalisation ability, as the resulting interpretations may not be universally valid.

Counterfactuals in Generalized Learning Vector Quantization

Thomas Villmann

SICIM, University of Applied Sciences Mittweida, Germany

Counterfactual explanations are valuable for interpreting machine learning classifiers. While often requiring constrained optimization, we demonstrate that prototype-based classifiers allow for the analytical determination of counterfactuals. This is possible when nearest prototype classification and counterfactual distance are evaluated using norms induced by an inner product. Our approach avoids optimization, offering a more efficient solution.

It AI-n't what you think – Let's drop the term[*]

Michael Biehl
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence
University of Groningen, Groningen, The Netherlands

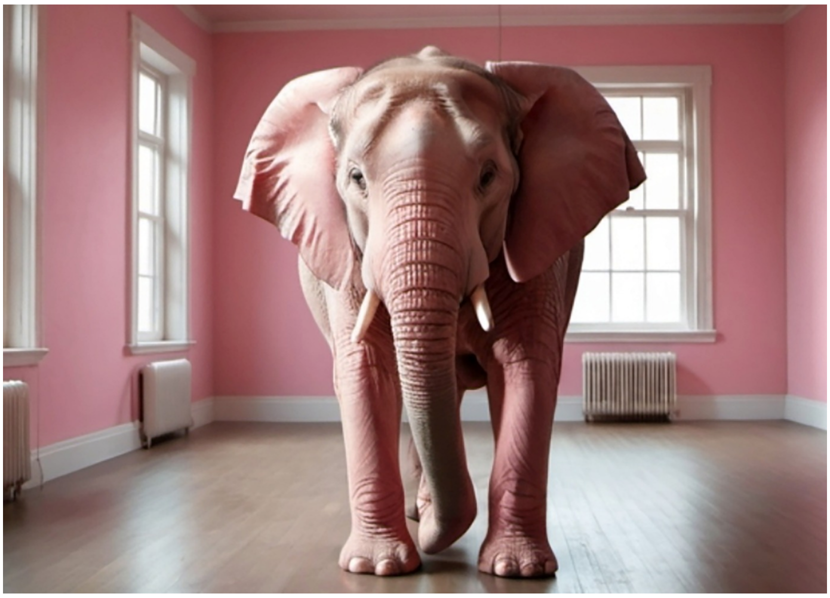
Abstract

In this short rant I argue that we, being serious machine learning researchers and data scientists, should avoid the term *Artificial Intelligence (AI)* as much as possible. At the very least, we have the responsibility to provide precise definitions of our methods and models and explain how they differ from the heavily over-hyped Large Language Models and other *Generative AI*. This obviously concerns publications and, perhaps more importantly, the communication with collaboration partners and the media. In the talk, a brief discussion of the most important shortcomings and risks associated with these currently very popular tools is provided. Given the current, very one-sided perception and limited "understanding" of *AI* in the public and in the media, it is essential to clearly distance ourselves and to avoid the term *AI* as much as possible. In view of the imminent collapse of the *GenAI* bubble it is expected that the disillusionment will hit hard. Eventually, it will turn against all forms of machine learning and data analysis as well. The community should be prepared for the worst.

Unfortunately, the public perception of *Artificial Intelligence* is dominated by the naive identification of this longstanding, very broad field of research with the currently heavily over-hyped *generative AI*. In my opinion, it is too late to correct this counterproductive, purposefully created and maintained misconception. Hence my plea to drop the term after all.

Large Language Models (LLM) and other generative systems do what they are designed too, often displaying stunning performance. Without a doubt, there are some (few) reasonable use cases for LLM or text-to-image generators. They generate eloquent texts, beautiful images or impressive videos. However, big tech companies, uncritical media, and an ever-growing army of self-proclaimed experts keep promoting the believe that *generative AI* provides universal tools with unlimited abilities. Exaggerated praise predicts billion-dollar-markets, overwhelming fears range from the loss of countless jobs to the end of humanity. *AI* systems are sold as being just one step away from achieving *artificial general intelligence (AGI)*, whatever that means precisely. It is claimed that they truly comprehend texts, really analyse problems and are able

[*] Title suggested by ChatGPT



The elephant in the room. Image generated by OpenArtSDXL, using the prompt “A photorealistic image of a totally empty room without any elephant. In particular, there should be no pink elephant.”

to reason [1]. Overstated claims of LLMs passing the Turing test [2] or achieving high scores in academic exams are frequently based on shady evaluation methods and generally obscured by the intransparent training.

However, slowly but surely, people begin to realize, for instance, that even the largest language models are what it says on the tin: language models. They put together phrases, imitate or blatantly reproduce and remix pieces of text from their training data. So-called *hallucinations* (a humanising misnomer) in text, image or video generation may be reduced or mildened by labor-intensive fixes or by resorting to the dubious *art of prompt engineering*. But the basic problem remains: factually wrong texts, implausible images, and physically impossible videos are intrinsic features of these systems, not bugs that will go away with the next version [3]. This renders *generative AI* an extremely risky technique that is not suitable for critical applications [4].

The many ethical issues of *generative AI* begin to be recognized as well: their enormous waste of resources [5], the exploitation of cheap labor worldwide in order to correct bugs [6], the unauthorized misuse of copyrighted materials [7], the abusive collection of user data [8], and the uncritical amplification of biases in the training data [9]. Last not least, there is a growing awareness and fear of possible manipulations by authoritarian regimes or the big tech companies that release and control these systems [10].

As an obvious consequence, we should be very hesitant - to put it mildly - to use *genAI* in any scientific or educational context. Moreover, as a researcher with main interests in methods, models and serious applications of old school machine learning, I appeal to my colleagues: Stop using the term *Artificial Intelligence, AI* as much as you can. At the very least, be very transparent and specific about its precise intended meaning in your publications, press releases or when communicating with collaborators. When the *genAI* bubble bursts – and it will – the media and the public mood will turn against everything labelled *AI*. Do not fall for the hype, do not fuel it any further, and refrain from trying to take advantage of it. It will backfire.

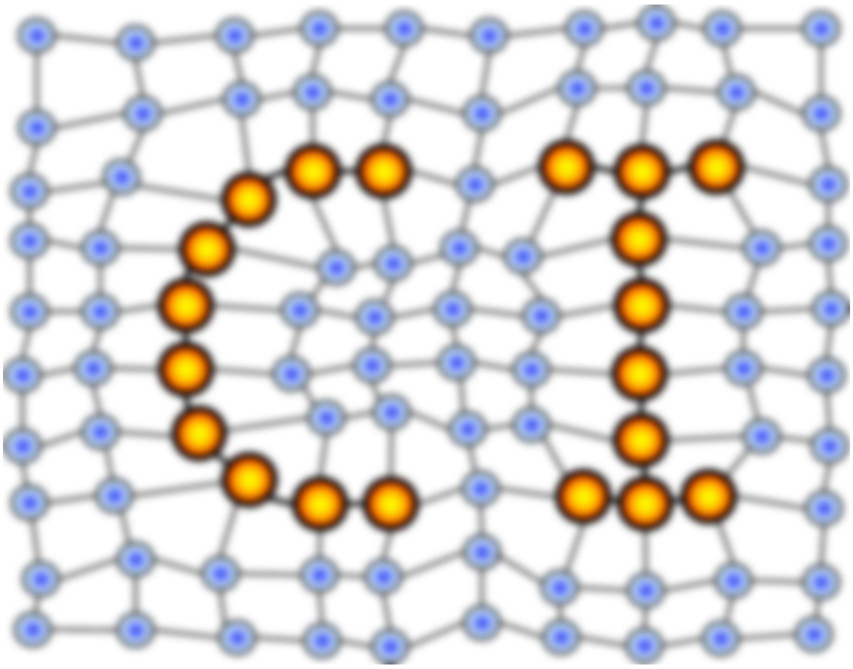
Example links

[1] https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf
[2] https://garymarcus.substack.com/p/ai-has-sort-of-passed-the-turing
[3] https://www.computerworld.com/video/2099753/why-ai-hallucinations-are-here-to-stay.html
[4] https://www.monettdiaz.com/, https//garymarcus.substack.com
[5] https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117
[6] https://www.theguardian.com/technology/2025/sep/11/google-gemini-ai-training-humans
[7] https://sites.usc.edu/iptls/2025/02/04/ai-copyright-and-the-law-the-ongoing-battle-over-intellectual-property-rights/
[8] https://www.sciencenewstoday.org/the-dark-side-of-ai-cybersecurity-threats-and-privacy-concerns
[9] https://studyfinds.org/ai-systems-amplify-human-bias/
[10] https://cointelegraph.com/news/elon-musk-grok-ai-rewrite-the-entire-corpus-human-knowledge

Related opinion piece by Charlotte Vlek, University of Groningen:
https://www.rug.nl/fse/news/digital-society/the-genai-bubble-will-burst-but-dont-give-up-on-ai-altogether

Machine Learning Reports

Report 02/2025



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Prof. Dr. rer. nat. Frank-Michael Schleif
Technical University of Applied Sciences Wuerzburg-Schweinfurt
Sanderheinrichsleitenweg 20, 97074 Wuerzburg, Germany
• https://fiw.thws.de/

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.