MUTATION VALIDATION FOR LEARNING VECTOR QUANTIZATION

Nana Abeka Otoo
nanaabekaotoo@authentic.network.com

ABSTRACT

Mutation validation as a complement to existing applied machine learning validation schemes has been explored in recent times. Exploratory work for Learning vector quantization (LVQ) based on this model-validation scheme remains to be discovered. This paper proposes mutation validation as an extension to existing cross-validation and holdout schemes for Generalized LVQ and its advanced variants. The mutation validation scheme provides a responsive, interpretable, intuitive and easily comprehensible score that complements existing validation schemes employed in the performance evaluation of the prototype-based LVQ family of classification algorithms. This paper establishes a relation between the mutation validation scheme and the goodness of fit evaluation for four LVQ models: Generalized LVQ, Generalized Matrix LVQ, Generalized Tangent LVQ and Robust Soft LVQ models. Numerical evaluation regarding these models complexity and effects on test outcomes, pitches mutation validation scheme above cross-validation and holdout schemes.

Keywords Learning vector quantization · Mutation validation

1 Introduction

Machine learning model validation plays an essential role in the practice of applied machine learning (Zhang et al. [2021]). For practitioners, the qualification state of a learned model is based on the metric employed in its evaluation (Brownlee [2020]). An appropriate evaluation metric remains vital for any meaningful learning (Sun et al. [2009]). Nonetheless, selecting the proper evaluation metric must fit in the right framework of a validation scheme in order to ensure the benefits thereof (Ferri et al. [2009]). Cross-validation and holdout schemes are mostly verified and preferred validation schemes used chiefly in machine learning projects (Ma and He [2013], Zhang et al. [2021]). Even though these schemes enjoy popularity in use-case scenarios, utilization has been challenging regarding keeping a small but sizeable fraction of the input data as a test set for validation purposes (Ma and He [2013], Pham et al. [2020]). The problems of the size of a training data set, imbalances in class distribution as well as the selection of instances into the test set for validation may affect the outcome of the evaluation score of a learned model (Brownlee [2020], Piironen and Vehtari [2017]). A validation scheme that uses a whole data set in training and evaluation provides an immediate answer to the compromise faced with cross-validation and holdout schemes (Zhang et al. [2021], Gronau and Wagenmakers [2019]). Mutation validation is based on a single interpretable score which measures the ability of a leaner to withstand perturbation in the target space while the input space remains unadulterated (Zhang et al. [2021]). The mutation validation score provides a measure by which practitioners could further ascertain, complementarily, the goodness of fit state for prototype-based LVQ classification models based on utilizing the entire training set with no credence to a holdout set.

2 Classification Learning with Learning Vector Quantization

Learning vector quantization (LVQ) remains a robust and highly interpretable machine learning algorithm and has a formulation that is mathematically precise and highly comprehensible (Kohonen [1990]). Learning prototypes with generalized class responsibilities is based on a winner take all rule (Kohonen and Kohonen [1995]). The mathematical formulation for LVQ follows that given $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N\} \subseteq \mathbb{R}^n$ as training set with targets

$c(\mathbf{x}) \in \mathcal{C} = \{1, 2, \ldots, C\}$ and atleast per class prototype based on a set of vectors $W = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\} \subseteq \mathbb{R}^n$, the learning goal of attaining a generalised, representative and interpretable set of prototypes is characterized by the winner takes all rule $f : \mathbf{x} \to c(\mathbf{x}) \iff \operatorname{argmin}_j d(\mathbf{x}, \mathbf{w}_j)$ is satisfied $\forall\, \mathbf{x} \in X$ with d being any computable dissimilarity measure usually chosen as the Euclidean distance (Kohonen and Kohonen [1995]). Based on the attraction and repulsion of prototypes, LVQ classifiers can attain representative class distributions (Kohonen [1990]). To minimize errors associated with classification, a cost function approach based on SGD learning is presented by (Sato and Yamada [1995]) as

$$J_{GLVQ}(X, \phi) = \sum_{i=1}^{N} \phi(\mu(\mathbf{x}_i)) \tag{1}$$

where $\mu(\mathbf{x}_i)$ is the discriminant function with a monotonically increasing ϕ as the activation function given respectively as

$$\mu(\mathbf{x}) = \frac{d(\mathbf{x}, \mathbf{w}^+) - d(\mathbf{x}, \mathbf{w}^-)}{d(\mathbf{x}, \mathbf{w}^+) + d(\mathbf{x}, \mathbf{w}^-)} \tag{2}$$

with $d^+(\mathbf{x}) = d(\mathbf{x}, \mathbf{w}^+)$ indicating the correct best distance and $d^-(\mathbf{x}) = d(\mathbf{x}, \mathbf{w}^-)$ incorrect best distance based on $\mathbf{w}^+ \in W$ that best correctly assign $\mathbf{x}$ and $\mathbf{w}^- \in W$ that best incorrectly assign $\mathbf{x}$ based on a differentiable dissimilarity measure d (Sato and Yamada [1995]).

$$\phi_\theta(k) = \left(1 + a.e^{\frac{(k-k_0)}{2\theta^2}}\right)^{-1} \tag{3}$$

The classifier function in (2) attains a hard state for SGD preservation when $\theta \in (0, 1] \searrow 0$ implying $\phi(k) \searrow H(k)$ with

$$H(k) = \begin{cases} 1, & \text{if } k \leq 0. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

(Kästner et al. [2013], Villmann et al. [2018]). For Matrix-GLVQ, the differentiable dissimilarity measure d is expressed along with a positive definite relevant matrix specification as $d_\Omega(\mathbf{x}, \mathbf{w}) = (\Omega(\mathbf{x} - \mathbf{w}))^2$ with $\Omega \in \mathbb{R}^{m \times n}$ utilized for projection purpose, a generalized and numerically stable learner is realized (Hammer and Villmann [2002], Biehl [2006], Villmann et al. [2017]). The projection parameter m accounts for the intended dimension of the relevance matrix Ω and its regularization through dimension reduction allows GMLVQ to attain a robust learner designation whenever $m < n$ (Schneider et al. [2010]). Learning of prototypical representation in GLVQ is based on (5)

$$\Delta\mathbf{w}^\pm \propto \frac{-\partial\phi}{\partial\mu} \cdot \frac{\pm 2 d^\mp(\mathbf{x})}{(d^+(\mathbf{x}) + d^-(\mathbf{x}))^2} \cdot \frac{\partial d(\mathbf{x}, \mathbf{w}^\pm)}{\partial \mathbf{w}^\pm} \tag{5}$$

similarly, prototype updates and Ω adaptation in Matrix-GLVQ is based on (7) and (6)

$$\Delta\Omega \propto \frac{-\partial\phi}{\partial\mu} \left(\frac{\partial\mu}{\partial d_\Omega^+(\mathbf{x})} \cdot \frac{\partial d_\Omega^+(\mathbf{x})}{\partial\Omega} + \frac{\partial\mu}{\partial d_\Omega^-(\mathbf{x})} \cdot \frac{\partial d_\Omega^-(\mathbf{x})}{\partial\Omega} \right) \tag{6}$$

$$\Delta\mathbf{w}^\pm \propto \frac{-\partial\phi}{\partial\mu} \cdot \frac{\pm 2 d_\Omega^\mp(\mathbf{x})}{\left(d_\Omega^+(\mathbf{x}) + d_\Omega^-(\mathbf{x})\right)^2} \cdot \frac{\partial d_\Omega(\mathbf{x}, \mathbf{w}^\pm)}{\partial \mathbf{w}^\pm} \tag{7}$$

A tangent distance-based GLVQ where prototypes $W = \{(\mathbf{w}_1, \mathbf{W}_1), (\mathbf{w}_2, \mathbf{W}_2), \ldots, (\mathbf{w}_M, \mathbf{W}_M)\}$ are defined from the affine subspace of the input space for the approximation of class-based variations is introduced (Saralajew and Villmann [2016]). For a given instance $\mathbf{x} \in X$, the translation vector $\mathbf{w}_K$ and the m-dimensional orthonormal basis $\mathbf{W}_M \in \mathbb{R}^{n \times m}$ are used to compute the minimum Euclidean distance to the k-th affine subspace by

$$\min\{d(\mathbf{x}, \mathbf{w}_k + \mathbf{W}_k\theta) \mid \theta \in \mathbb{R}^m\} = d\left(\mathbf{x}, \mathbf{w}_k + \mathbf{W}_k\mathbf{W}_k^T(\mathbf{x} - \mathbf{w_k})\right) \tag{8}$$

where for the k-th affine subspace, the closest prototype is determined by $\mathbf{w}_k + \mathbf{W}_k\mathbf{W}_k^T(\mathbf{x} - \mathbf{w_k})$ (Saralajew and Villmann [2016]). Optimization of the affine-subspaces in W is based on (1) and (5) using SGD learning.

A probabilistic variant of LVQ that utilizes a soft model predictor based on the assumption that the prototypical representation regarding the input space will be the centers of a Gaussian mixture model is introduced by (Seo and Obermayer [2003]) as Robust Soft LVQ. Learning in RSLVQ follows probabilistic approach hence maximizes the mutual information between the predicted probability vector $p_W(\mathbf{x}) = (p_W(1|\mathbf{x}), p_W(2|\mathbf{x}), \ldots, p_W(C|\mathbf{x}))^T$ and actual class target probability vector $p(\mathbf{x}) = (p_1(\mathbf{x}), \ldots, p_C(\mathbf{x}))^T$ by minimizing the cross-entropy loss. The soft model predictor is given by

$$p_{\mathbf{w}}(c|\mathbf{x}) = \frac{\sum_{j:c(\mathbf{w}_j)=c} \exp\left(-\left((\mathbf{x}-\mathbf{w}_j)\right)^2\right)}{\sum_l \exp\left(-\left((\mathbf{x}-\mathbf{w}_l)\right)^2\right)} \tag{9}$$

and for Matrix-RSLVQ, $-\left(\Omega(\mathbf{x}-\mathbf{w}_j)\right)^2$ replaces $-\left((\mathbf{x}-\mathbf{w}_j)\right)^2$ in equation (9). [1]

3 Mutation Validation for Learning Vector Quantization

Consider the training pair $T = \{\mathbf{x}_i, c(\mathbf{x}_i)\}_{i=1}^N \in \{\mathbb{R}^n, \mathcal{C}\}^N$ and $\hat{T} = \{\mathbf{x}_i, \hat{c}(\mathbf{x}_i)\}_{i=1}^N \in \{\mathbb{R}^n, \mathcal{C}\}^N$ as original and mutated sets respectively. We define from the GLVQ discriminant function (2) the following relation $\mu_d : \mathbf{x} \to [-1, 1]$ with $\phi : \mu_d(\mathbf{x}) \to c(\mathbf{x}) \in \mathcal{C}$ and similarly, $\hat{\mu}_d(\mathbf{x}) \to [-1, 1]$ with $\phi : \hat{\mu}_d(\mathbf{x}) \to \hat{c}(\mathbf{x}) \in \mathcal{C}$

The expected risk of the loss function used in GLVQ (1) is given by

$$L_R = E_{\mathbf{x},c(\mathbf{x})}\left[\phi\left(\mu_d(\mathbf{x}_i)\right)\right] \tag{10}$$

and for a given pair of instances with the perturbed target in $\hat{T}$, we have

$$L_{\hat{R}} = E_{\mathbf{x},c(\mathbf{x})}\left[\phi\left(\hat{\mu}_d(\mathbf{x}_i)\right)\right] \tag{11}$$

with

$$\hat{c}(\mathbf{x}) = \begin{cases} c(\mathbf{x}), & p(c(\mathbf{x})) = 1 - \varphi_x. \\ j \backslash c(\mathbf{x}), & p(j) = \bar{\varphi_x}. \end{cases} \tag{12}$$

where the pair $(\mu_d(\mathbf{x}), \hat{\mu}_d(\mathbf{x})) \in H_s^3$ with delta loss $\Delta_L(\mathbf{x}) = \phi(\mu_d(\mathbf{x})) - \phi(\hat{\mu}_d(\mathbf{x}))$ and delta loss rate $\Delta_R(\mathbf{x}) = \frac{\Delta_L(\mathbf{x})}{\varphi}$. The expectation of the delta loss rate follows for symmetric perturbation as

$$E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R(\mathbf{x})\right] = E_{\mathbf{x},c(\mathbf{x})}\left[\frac{(1-\varphi)\,\phi_i(\hat{\mu}_d(\mathbf{x})) + \frac{\varphi}{C-1}\sum_{j\neq i}\phi_j(\hat{\mu}_d(\mathbf{x})) - \phi_i(\mu_d(\mathbf{x}))}{\varphi}\right] \tag{13}$$

$$E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R(\mathbf{x})\right] = E_{\mathbf{x},c(\mathbf{x})}\left[\frac{(1-\varphi)}{\varphi}\phi_i(\hat{\mu}_d(\mathbf{x})) + \frac{1}{C-1}\sum_{j\neq i}\phi_j(\hat{\mu}_d(\mathbf{x})) - \frac{1}{\varphi}\phi_i(\mu_d(\mathbf{x}))\right] \tag{14}$$

$$E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R(\mathbf{x})\right] = \frac{(1-\varphi)}{\varphi}L_{\hat{R}} + \frac{J - L_{\hat{R}}}{C-1} - \frac{L_R}{\varphi} \tag{15}$$

$$E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R(\mathbf{x})\right] = \left(\frac{1}{\varphi} - \frac{C}{C-1}\right)L_{\hat{R}} - \frac{1}{\varphi}L_R + \frac{J}{C-1} \tag{16}$$

[1] In applications $a = 1$, $k_0 = 0$ and $\theta > 0$ (Villmann et al. [2018]). The magnitude of the orientation of the discriminant function in (2) determines the extent of the classification decision with $\mu \in [-1, 1]$ meaning for incorrect classification, $d\left(\mathbf{x}, \mathbf{w}^+\right) > d\left(\mathbf{x}, \mathbf{w}^-\right)$ and vice-versa. When $m = n$, the positive definite matrix $\Lambda = \Omega^T\Omega \in \mathbb{R}^{n\times n}$ and H is the Heaviside function herein obtained by transitioning through the Sigmoid function in (3). Robustness in Matrix-GLVQ can be improved by a regularization technique which involves a penalized dimension (Schneider et al. [2010], Bunte et al. [2012]).

since GLVQ loss function (1) minimizes classification error (Sato and Yamada [1995]), J^1 is taken as 1.

$$E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R\left(\mathbf{x}\right)\right] = \left(\frac{1}{\varphi} - 2\right) L_{\hat{R}} - \frac{1}{\varphi} L_R + 1 \tag{17}$$

$$L_R = \left(1 - 2\varphi\right) L_{\hat{R}} - \varphi E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R\left(\mathbf{x}\right)\right] + \varphi \tag{18}$$

for a specified evaluation metric ξ, we have
$\xi\left(\phi_i\left(\mu_d\left(\mathbf{x}\right)\right)\right) = 1 - E_{\mathbf{x},c(\mathbf{x})}\left[\phi\left(\mu_d\left(\mathbf{x}_i\right)\right)\right]$ and $\xi\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) = 1 - E_{\mathbf{x},c(\mathbf{x})}\left[\phi\left(\hat{\mu}_d\left(\mathbf{x}_i\right)\right)\right]$ based on (10) and (11)

$$mv = \left(1 - 2\varphi\right) \xi\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) + \varphi E_{\mathbf{x},c(\mathbf{x})}\left[\Delta_R\left(\mathbf{x}\right)\right] + \varphi \tag{19}$$

$$mv = \left(1 - 2\varphi\right) \xi_T\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) + \varphi \frac{\xi_T\left(\phi_i\left(\mu_d\left(\mathbf{x}\right)\right)\right) - \xi_{\hat{T}}\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right)}{\varphi} + \varphi \tag{20}$$

$$mv = \left(1 - 2\varphi\right) \xi_T\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) + \xi_T\left(\phi_i\left(\mu_d\left(\mathbf{x}\right)\right)\right) - \xi_{\hat{T}}\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) + \varphi \tag{21}$$

where $\xi_T\left(\phi_i\left(\mu_d\left(\mathbf{x}\right)\right)\right)$, $\xi_T\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right)$ and $\xi_{\hat{T}}\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right)$ are specified evaluation metric scores with respect to $\{T, \hat{T}\}$ and to reflect the disposition to LVQ with regards to the accuracy metric,

$$\xi_T\left(\phi_i\left(\mu_d\left(\mathbf{x}\right)\right)\right) = \frac{\#\left\{\left(\mathbf{x}, c\left(\mathbf{x}\right)\right) \in T |\ c(\mathbf{x}) = c(\mathbf{w}_{s(\mathbf{x})})\right\}}{\#T} \tag{22}$$

$$\xi_T\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) = \frac{\#\left\{\left(\mathbf{x}, \hat{c}\left(\mathbf{x}\right)\right) \in \hat{T} |\ c(\mathbf{x}) = \hat{c}(\mathbf{w}_{s(\mathbf{x})})\right\}}{\#T} \tag{23}$$

$$\xi_{\hat{T}}\left(\phi_j\left(\hat{\mu}_d\left(\mathbf{x}\right)\right)\right) = \frac{\#\left\{\left(\mathbf{x}, \hat{c}\left(\mathbf{x}\right)\right) \in \hat{T} |\ \hat{c}(\mathbf{x}) = \hat{c}(\mathbf{w}_{s(\mathbf{x})})\right\}}{\#\hat{T}} \tag{24}$$

and hence referred to as *training accuracy* (TA), *robust mutation training accuracy* (RMTA) and *mutation training accuracy* (MTA) respectively for equations (22, 23 and 24).

$$mv = \left(1 - 2\varphi\right) RMTA + TA - MTA + \varphi \tag{25}$$

It follows from equation (25), for the best-case scenario, the LVQ learners are expected to possess two properties, namely (1) good generalization and (2) numerical stability and hence equation (23) is almost not susceptible to target space perturbation implying RMTA $\searrow 1$ and a magnified difference for $(TA - MTA) \searrow \varphi$ and with a constant φ in a limited target space perturbation, $mv \searrow 1$ and this follows same conclusions derived by (Zhang et al. [2021]). The mv scores are reminiscent of the interpretability of existing model validation scores such as CV but with extra information depicting the goodness of fit of LVQ learners.

The evaluation metric employed here must be selected informatively based on a specific variant of LVQ with regard to its associated cost function. The GLVQ cost function minimizes the classification error (Sato and Yamada [1995]) and hence may not be overly appropriate to learn prototypes for imbalanced training sets. In this regard, users must opt for a variant of GLVQ that uses a threshold parameter γ in the classifier function in (2) and is referred to as ROC-LVQ (Villmann et al. [2018]). However, a bi-directional approach focusing on a learner's susceptibility to perturbation in the input space and corresponding mutation in the target space may require using the relu (triplet loss) and the GLVQ-loss for optimization. Further investigation is needed in this regard. The mutation of labels is executed based on a fixed user-defined rate φ. [2]

[2] H_s is the hypothesis space and is identical. The perturbation method must be chosen cautiously based on prior knowledge of the target space distribution. This paper recommends global or uniform target space mutations for randomly selected instances when the target space is balanced. The balanced class-aware swap scheme should be used for an imbalanced target space. The mutation degree must be fixed ([Zhang et al., 2021]). The feature space is not affected thereof.

Label Perturbation

Algorithm 1 Target space perturbation

Require: Training set $T = \{\mathbf{x}_n, c(\mathbf{x}_n)\}_{n=1}^{N} \in \{\mathbb{R}^n, \mathcal{C}\}^N$
1: **Initialize** a fixed $\varphi \in [0, 0.5]$
2: **Select** randomly φ labels of the target space by global, uniformly class-aware or balance class-aware schemes.
3: **Perturb** the selected labels either by swap-next or swap randomly but uniquely method
4: **Return** the perturbed data set $\hat{T} = \{\mathbf{x}_n, \hat{c}(\mathbf{x}_n)\}_{n=1}^{N} \in \{\mathbb{R}^n, \mathcal{C}\}^N$

Mutation Validation

Algorithm 2 MV Algorithm for LVQ

Require: Training set $T = \{\mathbf{x}_n, c(\mathbf{x}_n)\}_{n=1}^{N} \in \{\mathbb{R}^n, \mathcal{C}\}^N$
1: **Perturb** T using the target space perturbation algorithm in (1)
2: **Setup** specified LVQ model with fixed hyper-parameters and learn prototypes using T
3: **Evaluate** and record the model performance $\xi_T(\phi_i(\mu_d(\mathbf{x})))$
4: **Train** a reinitialized but same model-type on same hyper-parameters used in step 3. based on $\hat{T}$
5: **Evaluate** and record the model performance $\xi_T(\phi_j(\hat{\mu}_d(\mathbf{x})))$ and $\xi_{\hat{T}}(\phi_j(\hat{\mu}_d(\mathbf{x})))$
6: **Compute** MV score using eqn. (21)

The evaluation metric must be cautiously chosen based on model cost function type, target space distribution and learning goals.

4 Experimentation

This section illustrates the aforementioned approach with two artificial sets S_1 and S_2 and two real-world data sets, namely WDBC from the UCI data set ([Blake, 1998]) and MNIST handwritten data ([Deng, 2012]).

Since the complexity of LVQ models scales with a high cardinality of prototypes, care must be taken in choosing the optimal number of prototypes per class to avoid over-fitting (Villmann et al. [2018]). However, further research has revealed that increasing the number of prototypes for GLVQ and GTLVQ leads to improved generalization ability and robustness ([Saralajew et al., 2020a]). This paper investigates the relationship between the mutation validation score and the behavior of the decision boundary of an LVQ learner regarding the model complexity. This paper also investigates, for a fixed perturbation ratio φ, the relation to the size of the data set and learner complexity bearing that, performance of LVQ learners may be affected for extensive data (Villmann et al. [2017]).Furthermore, this paper investigates the relationship between existing model validation schemes (CV and hold-out) and the proposed MV scheme for evaluating and selecting LVQ classification algorithms before deployment.

To reflect the behavior of mutation validation score concerning the decision boundary of a learner herein considered, a signed delta significance for decision-making is introduced and indicated as

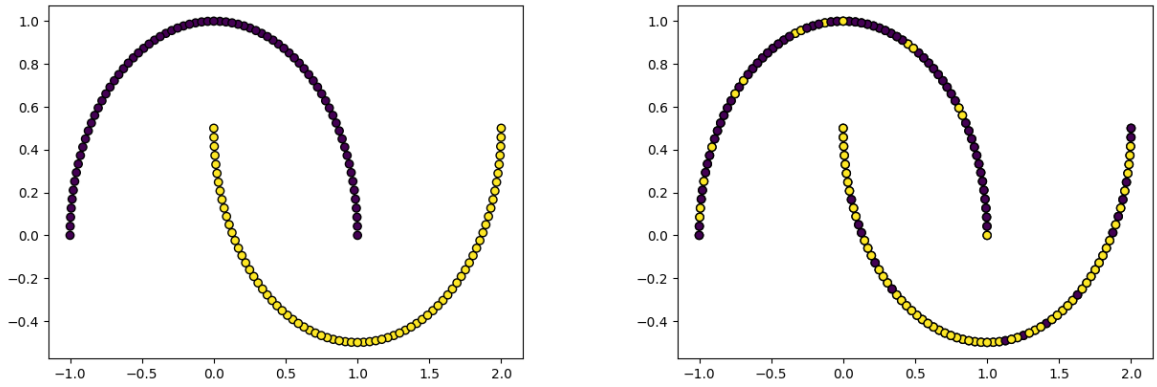
$$\delta_d = \begin{cases} +v, & \text{if } MV \geq CV. \\ -v, & \text{otherwise.} \end{cases} \tag{26}$$

with $v = |MV - CV|$, when $\delta_d > 0$ means the decision boundary will be insensitive to mutant labels in the target space and for $\delta_d < 0$ the sensitivity of the decision boundary to mutant labels is brought to bear. In comparison, v accounts for the extent of the observed sensitivity. In use case scenarios, any model selected by the mutation validation scheme is only deployed based on steps (2) and (3) in the MV Algorithm for LVQ (2) since by this way, practitioners can make adequate use of the entire data set for training without compromise. The implementation of the algorithms (1) and (2) in Python is made accessible by the author at (Otoo [2023])

4.1 Artificial data sets

The toy sets S_1 and S_2 with two attributes each where$|S_1| < |S_2|$ for binary classes $\{0:75, 1:75\}$ and $\{0:120, 1:80\}$ were respectively generated. The prototypes were selected by random initialization from the feature space, with focus on unique labels.The proven robust learners GLVQ and GTLVQ (Bunte et al. [2012], Kaden et al. [2014],

Saralajew et al. [2020a,b], Voráček and Hein [2022]), some-what robust[1] Matrix-GLVQ learner([Schneider et al., 2010, Saralajew et al., 2020a]) and the yet to be proven robust RSLVQ leaner were used for the experiments ([Saralajew et al., 2020a]). The mutation validation, cross-validation and holdout results are investigated for an equal prototype distribution with a learning rate chosen as 0.01 and fixed for both learning models using the SGD optimizer. Presented below are the results and learner behavior for a fixed φ with a five-fold cross validation and 0.2 holdout scheme using the accuracy metric.



(a) S_1 (b) S_1 mutated with $\varphi = 0.2$ based on 1

Figure 1: Visualization of the artificial training data set S_1 and class distributions generated for the experiments.

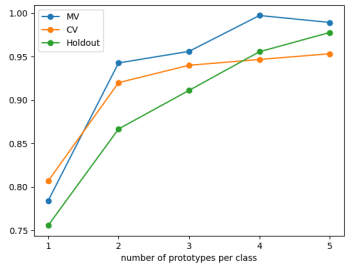
Table 1: classification accuracies with delta significance for the artificial data set S_1

	GLVQ		GMLVQ		GTLVQ		RSLVQ	
Method	$1\|1$	$5\|5$	$1\|1$	$5\|5$	$1\|1$	$5\|5$	$1\|1$	$5\|5$
MV($\varphi = 0.2$)	78.40%	98.93%	91.20%	100.00%	89.33%	94.80%	$87.33,\%$	88.00%
CV	80.67%	95.33%	86.67%	91.33%	86.00%	90.00%	86.00%	86.00%
Holdout	75.56%	97.78%	84.40%	84.40%	80.00%	93.3%	80.00%	$84.44\ \%$
δ_d	-2.27%	$+3.60\%$	$+4.53\%$	$+8.67\%$	$+3.33\%$	$+4.80\%$	$+1.33\%$	$+2.00\%$
# param.	4	20	8	40	8	40	4	20

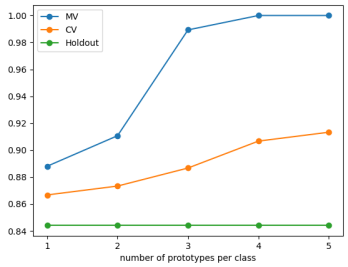
The results in Table 1 indicate a true reflection of the decision boundary behavior by the mutation validation scores. The CV and Holdout scores do not depict extra information regarding the response of the discriminant function when faced with mutants during learning. In Figures (2a,2b,2c), MV and CV scores increase along with the model complexity with slight exception in the RSLVQ learner, where fair but steady undulating validation scores are recorded. The MV scores capture the behavior of the learning dynamic and offer a complementary insight into numerical stability based on mutant training.

The obtained MV scores show that the models in Figure (2) can resist perturbation despite increased complexity. From Figure (2c), the Holdout evaluation scores stabilize and tend to be higher than the MV scores for increased learner complexity. The observed behavior of the Holdout scores in Figure (2c) is an attestation of the existence of a much less complex LVQ learner (few prototypes per class) with an appreciably close and similar performance. Moreover, whenever the CV scores rise, it implies a much more complex LVQ learner (more prototypes per class) with a higher MV score exist. In consequence, the corresponding δ_d for the observed cases indicates a good fit for all the models considered.

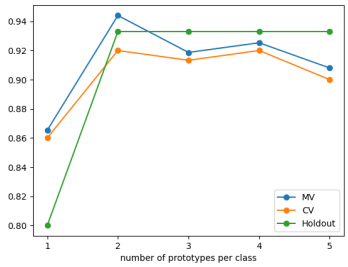
From Table 1, MV scores for the GMLVQ model rise more rapidly than CV scores, indicating the ability to capture the goodness of fit of the learner concerning increasing learner complexity. The MV scores for the GLVQ and GTLVQ models are more significant than the CV scores. Regarding RSLVQ, which is not a proven robust learner against adversarial attacks (Saralajew et al. [2020b]), Figure (2d) indicates how MV scores capture the RSLVQ learner behavior with regards to mutant instances.



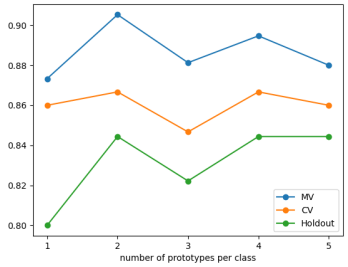
(a) GLVQ



(b) GMLVQ



(c) GTLVQ

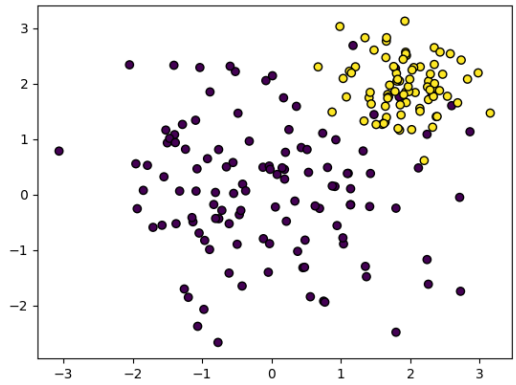


(d) RSLVQ

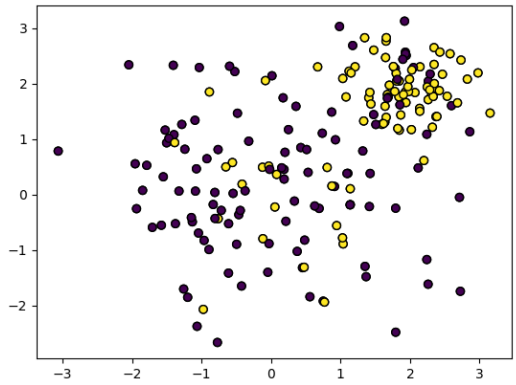
Figure 2: MV, CV and Holdout accuracy scores against model complexity for the S_1 artificial data set

This behavior suggests how responsive the training by mutation validation reflects the response of the decision boundary regarding model complexity and susceptibility to mutant labels in the training data. The results in Table 1 indicate the extent of a learner's susceptibility to fit perturbations in the training data as the model complexity increases. Thus, a practitioner can select the best model according to training by mutation validation, which compliments the corresponding CV scores.

In order to gather further insights regarding the behavior of the proposed mutation validation scheme for learning vector quantization, the second generated artificial set S_2 with a much less even distribution (skewed) and considerably imbalanced target size (3) is used for further experiments. The observed numerical evaluation for the MV, CV and Holdout schemes is presented below.



(a) S_2

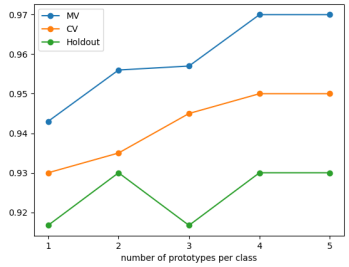


(b) S_2 mutated with $\varphi = 0.2$ (1)

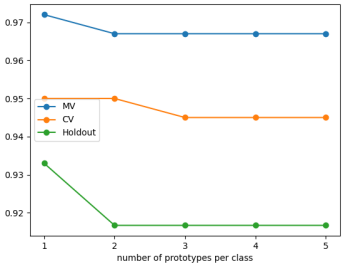
Figure 3: Visualization of artificial training data set S_2 and class distributions generated for the experiments.

Table 2: classification accuracies with delta significance for the artificial data set S_2

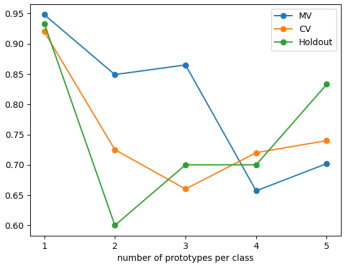
	GLVQ		GMLVQ		GTLVQ		RSLVQ	
Method	1|1	5|5	1|1	5|5	1|1	5|5	1|1	5|5
MV($\varphi = 0.2$)	94.30%	97.00%	97.20, %	96.70%	94.80%	70.20%	94.50%	97.00%
CV	93.00%	95.00%	95.00%	94.50%	92.00%	74.00%	95.00%	95.00%
Holdout	91.67%	93.00%	93.30%	91.67%	93.33%	83.33%	90.00%	91.67 %
δ_d	+1.30%	+2.00%	+2.20%	+2.20%	+2.80%	−3.80%	−0.50%	+2.00%
# param.	4	20	8	40	8	40	4	20



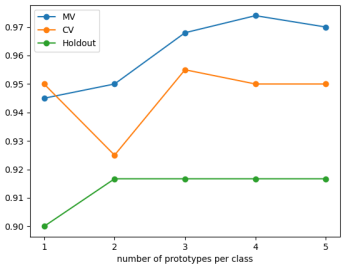
(a) GLVQ



(b) GMLVQ



(c) GTLVQ



(d) RSLVQ

Figure 4: MV, CV and Holdout accuracy scores against model complexity for the S_2 artificial data set

From Table 2 and Figures (4a,4b,4d), the MV scores pitch above CV scores for all the learners used for the training. However, the rising (4a, 4d), rising but fairly steady (4b) and decreasing (4c) behaviors observed indicates how well mutation validation captures informatively the faithful response of the decision boundary regarding the goodness of fit of the learners. Interpreting the MV scores indicates that the goodness of fit of GLVQ and RSLVQ rises with an increasing number of prototypes per class. GMLVQ has a reasonably stable goodness of fit and GTLVQ has a decreasing goodness of fit for an increasing number of prototypes. The analysis of learner(s) behavior in Figure (4) is an indication of the fact that whenever the Holdout scores decrease, stabilize and are less than the CV scores, there exists a less complex learner with better MV performance. Moreover, when the Holdout scores rise and stabilize, a much more complex learner with better MV performance exists. Hence, for the analyzed cases, the corresponding δ_d of the LVQ learner(s) indicates good fit.

In Figure (4c), the Holdout evaluation metric scores get more significant for an increased number of prototypes per class as compared to MV scores and hence notifies the presence of a learner with few prototypes per class but having appreciably equivalent or better performance. Thus, the MV scores are very responsive and can capture the learning dynamics and numerical stability in a single interpretable score.

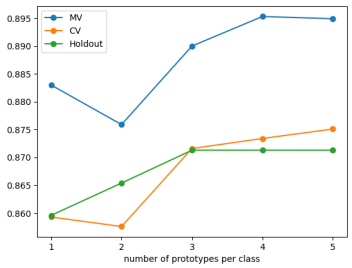
4.2 Real data sets

The WDBC data set from the UCI data repository was used to investigate the process for real-world cases. WDBC is a binary class data set with the non-infectious and infectious classes, with a total cardinality of 562 instances (Blake [1998]).

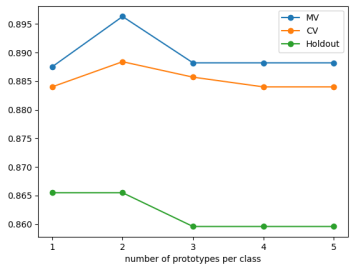
The GLVQ, GTLVQ, GMLVQ and RSLVQ models were employed for the training using a uniform prototype distribution with a learning rate of 0.01 fixed for all prototype-based models considered. The goodness of fit relation is explored using the mutation validation algorithm for LVQ described in (2). Target space perturbation is executed accordingly using (1) based on a fixed φ along with five-fold CV and 0.2 holdout evaluated with the accuracy metric.

Table 3: classification accuracies with delta significance for the artificial data set WDBC

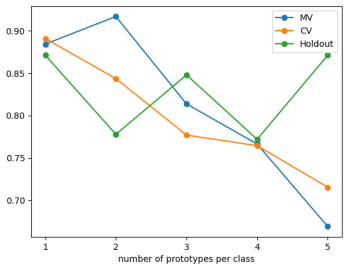
	GLVQ		GMLVQ		GTLVQ		RSLVQ	
Method	$1\|1$	$5\|5$	$1\|1$	$5\|5$	$1\|1$	$5\|5$	$1\|1$	$5\|5$
MV($\varphi = 0.2$)	88.30%	89.49%	88.75%	88.82%	88.44%	66.89%	37.08%	91.60%
CV	85.93%	87.51%	88.40%	88.40%	89.10%	71.53%	56.81%	88.93%
Holdout	85.96%	87.13%	86.55%	85.96%	87.13%	87.13%	31.58%	$88.89\ \%$
δ_d	$+2.37\%$	$+1.98\%$	$+0.35\%$	$+0.42\%$	-0.66%	-4.64%	-19.73%	$+2.67\%$
# param.	4	20	8	40	8	40	4	20



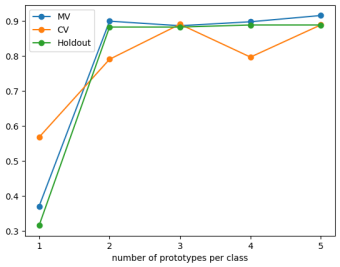
(a) GLVQ



(b) GMLVQ



(c) GTLVQ



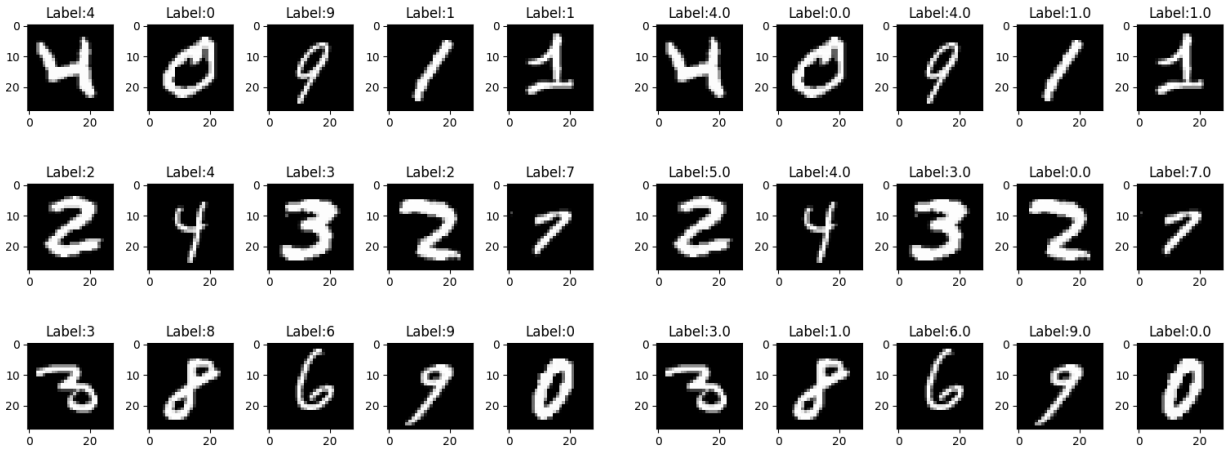
(d) RSLVQ

Figure 5: MV, CV and Holdout accuracy scores against model complexity for the WDBC data set

From Table 3 and Figure 5, the MV scores (5a,5b,5d) pitch above the CV scores for all the learners. It is also worth noting that the MV scores tend to increase with model complexity for the GLVQ and RSLVQ models and stabilize with increased complexity for GMLVQ. In contrast, for GTLVQ (5c), the MV scores decrease with regard to increased model complexity. This behavior is a notifier regarding the sensitivity of MV scores to capture the true reflection of decision boundary behavior. Regarding increasing learner complexity, the Holdout evaluation metric scores tend to be higher than the matching MV scores for GTLVQ (5c). The implication of the observation in (5c) is the existence of a less complex LVQ learner (few prototypes per class) with an appreciably close and almost similar performance. In such cases, practitioners must opt for learners with less complexity. Moreover, analysis of learner(s) behavior in Figures (5) shows that whenever the Holdout scores fall, stabilize and are less than the CV scores, there exists a less complex LVQ

learner with better MV performance. Conversely, when the Holdout scores rise and stabilize, a complex LVQ learner with a better MV score exists. Consequently, for the above-observed cases, using (26), the matching δ_d, indicates no susceptibility of the learners to mutants in the target space and therefore MV scores provide complement information needed for model selection.

In order to investigate the suggested validation process for non-tabular data with large size, the MNIST handwritten database ([Deng, 2012]) with the pair cardinalities $\{train : 60000,\ test : 10000\}$ is used for the holdout validation scheme while the pair union of 70000 images used for the CV and MV schemes. A uniform prototype distribution with a target space mutation degree $\varphi = 0.2$ was utilized.



(a) Extract of MNIST data set

(b) Extract of MNIST data set mutated with $\varphi = 0.2$ (1)

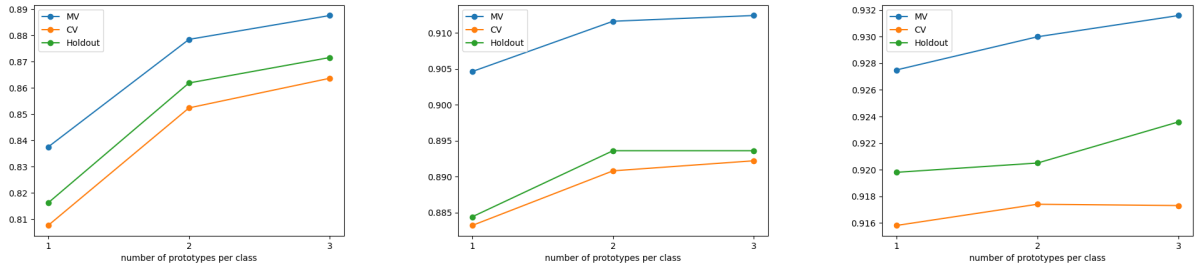
Figure 6: Visualization of MNIST handwritten training data set.

The results from Table (4) and Figure (7) indicate stable behavior for GLVQ, GMLVQ and GTLVQ. The MV scheme pitch ahead of the CV and Holdout model evaluation schemes and is indicative of the faithful response of the decision boundary of the learner(s) regarding the goodness of fit to the training set. Numerical evaluation of learner(s) behavior in Figures (7) shows that whenever the Holdout scores rise and stabilize, there exists a complex LVQ learner with a better MV score. Using (26) and results from (7), the learners exhibit a good fit regarding the behavior of the decision boundary to mutant labels in the target space.

The mutation validation scheme for LVQ provides an informative and responsive alternative to existing machine learning validation schemes. In most cases, it complements or replaces the CV and holdout model validation schemes. The MV scheme for LVQ is an additional but complementary model selection tool for applied prototype-based machine learning practitioners.

Table 4: classification accuracies with delta significance for the artificial data set MNIST

	GLVQ		GMLVQ		GTLVQ	
[1]Method	$1\|1$	$3\|3$	$1\|1$	$3\|3$	$1\|1$	$3\|3$
MV($\varphi = 0.2$)	83.75%	$88.76,\%$	$90.46,\%$	91.24%	92.75%	93.16%
CV	80.76%	86.37%	88.32%	89.22%	91.58%	91.73%
Holdout	81.62%	87.16%	88.44%	89.36%	91.98%	92.36%
δ_d	$+2.99\%$	$+2.39\%$	$+2.14\%$	$+2.02\%$	$+1.17\%$	$+1.43\%$
# param.	7.8 K	23.5 K	622 K	638 K	227 K	682 K



(a) GLVQ (b) GMLVQ (c) GTLVQ

Figure 7: MV, CV and Holdout accuracy scores against model complexity for the MNIST data set

5 Discussion

The analysis of outcomes in this paper shows that the MV scheme offers a faithful evaluation irrespective of the inherent distributions within the data set, the training data set size and the complexity of the prototype learners used for the experiments. Also, by the aforementioned MV scheme, a careful study of the outcomes presents an interpretable score by which practitioners can select learners based on a range of learnable parameters influencing model complexity. This study establishes a relationship between the MV scheme and existing model validation schemes(CV and Holdout), with recommendations for practitioners regarding their complementary use during the model selection process. Hence, a vital note would be for prototype-based machine learning practitioners to opt for the MV scheme in research areas where training data is diminutive and inadequate for the CV and Holdout evaluation schemes. The MV scheme is a useful tool in applied machine learning problems where critical evaluation of model selection is required before deployment.

6 Conclusion

A mutation validation scheme (MV) for prototype-based learning vector quantization classifiers has been presented in this paper. This work establishes a relationship between this model validation scheme and existing facts relating to the goodness of fit of the highly interpretable prototype-based LVQ classification algorithms. The mathematical formulation for this complimentary model validation scheme for LVQ classifiers has been introduced in this paper. The numerical evaluation of experimental results indicates that this model validation scheme captures well the behavior of the decision boundary regarding the model complexity of the LVQ classifiers. The MV score offers an interpretable reflection of the goodness of fit measurement during the model selection process for LVQ classifiers. The mutation validation scheme introduced can be implemented in parallel with existing validation schemes in machine learning pipelines and executed during runtime. Posited and confirmed is a new model validation for LVQs. Future work will investigate the application of the MV scheme as a replacement for the existing holdout evaluation of feature selection tasks for LVQs.

References

Jie M. Zhang, Mark Harman, Benjamin Guedj, Earl T. Barr, and John Shawe-Taylor. Model validation using mutated training labels: An exploratory study, 2021.

Jason Brownlee. *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery, 2020.

Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

César Ferri, José Hernández-Orallo, and R Modroiu. An experimental comparison of performance measures for classification. *Pattern recognition letters*, 30(1):27–38, 2009.

Yunqian Ma and Haibo He. Imbalanced learning: foundations, algorithms, and applications. 2013.

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. Problems and opportunities in training deep learning software systems: An analysis of variance. In *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*, pages 771–783, 2020.

Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27:711–735, 2017.

Quentin F Gronau and Eric-Jan Wagenmakers. Limitations of bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior*, 2:1–11, 2019.

Teuvo Kohonen. Improved versions of learning vector quantization. In *1990 ijcnn international joint conference on Neural networks*, pages 545–550. IEEE, 1990.

Teuvo Kohonen and Teuvo Kohonen. Learning vector quantization. *Self-organizing maps*, pages 175–189, 1995.

Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. *Advances in neural information processing systems*, 8, 1995.

Marika Kästner, Martin Riedel, Marc Strickert, Wieland Hermann, and Thomas Villmann. Border-sensitive learning in kernelized learning vector quantization. In *Advances in Computational Intelligence: 12th International Work-Conference on Artificial Neural Networks, IWANN 2013, Puerto de la Cruz, Tenerife, Spain, June 12-14, 2013, Proceedings, Part I 12*, pages 357–366. Springer, 2013.

Thomas Villmann, Marika Kaden, Wieland Hermann, and Michael Biehl. Learning vector quantization classifiers for roc-optimization. *Computational statistics*, 33:1173–1194, 2018.

Barbara Hammer and Thomas Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9): 1059–1068, 2002.

Michael Biehl. Matrix learning in learning vector quantization. 2006.

Thomas Villmann, Andrea Bohnsack, and Marika Kaden. Can learning vector quantization be an alternative to svm and deep learning?-recent trends and advanced variants of learning vector quantization for classification learning. *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017.

Petra Schneider, Kerstin Bunte, Han Stiekema, Barbara Hammer, Thomas Villmann, and Michael Biehl. Regularization in matrix relevance learning. *IEEE Transactions on Neural Networks*, 21(5):831–840, 2010.

Sascha Saralajew and Thomas Villmann. Adaptive tangent distances in generalized learning vector quantization for transformation and distortion invariant classification learning. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2672–2679. IEEE, 2016.

Sambu Seo and Klaus Obermayer. Soft learning vector quantization. *Neural computation*, 15(7):1589–1604, 2003.

Kerstin Bunte, Petra Schneider, Barbara Hammer, Frank-Michael Schleif, Thomas Villmann, and Michael Biehl. Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26:159–173, 2012.

Catherine Blake. Uci repository of machine learning databases. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Sascha Saralajew, Lars Holdijk, Maike Rees, and Thomas Villmann. Robustness of generalized learning vector quantization models against adversarial attacks. In *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization: Proceedings of the 13th International Workshop, WSOM+ 2019, Barcelona, Spain, June 26-28, 2019 13*, pages 189–199. Springer, 2020a.

Nana Abeka Otoo. Mutation-validation. `https://github.com/naotoo1/Mutation-Validation`, 2023.

Marika Kaden, Mandy Lange, David Nebel, Martin Riedel, Tina Geweniger, and Thomas Villmann. Aspects in classification learning-review of recent developments in learning vector quantization. *Foundations of Computing and Decision Sciences*, 39(2):79–105, 2014.

Sascha Saralajew, Lars Holdijk, and Thomas Villmann. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. *Advances in Neural Information Processing Systems*, 33:13635–13650, 2020b.

Václav Voráček and Matthias Hein. Provably adversarially robust nearest prototype classifiers. In *International Conference on Machine Learning*, pages 22361–22383. PMLR, 2022.