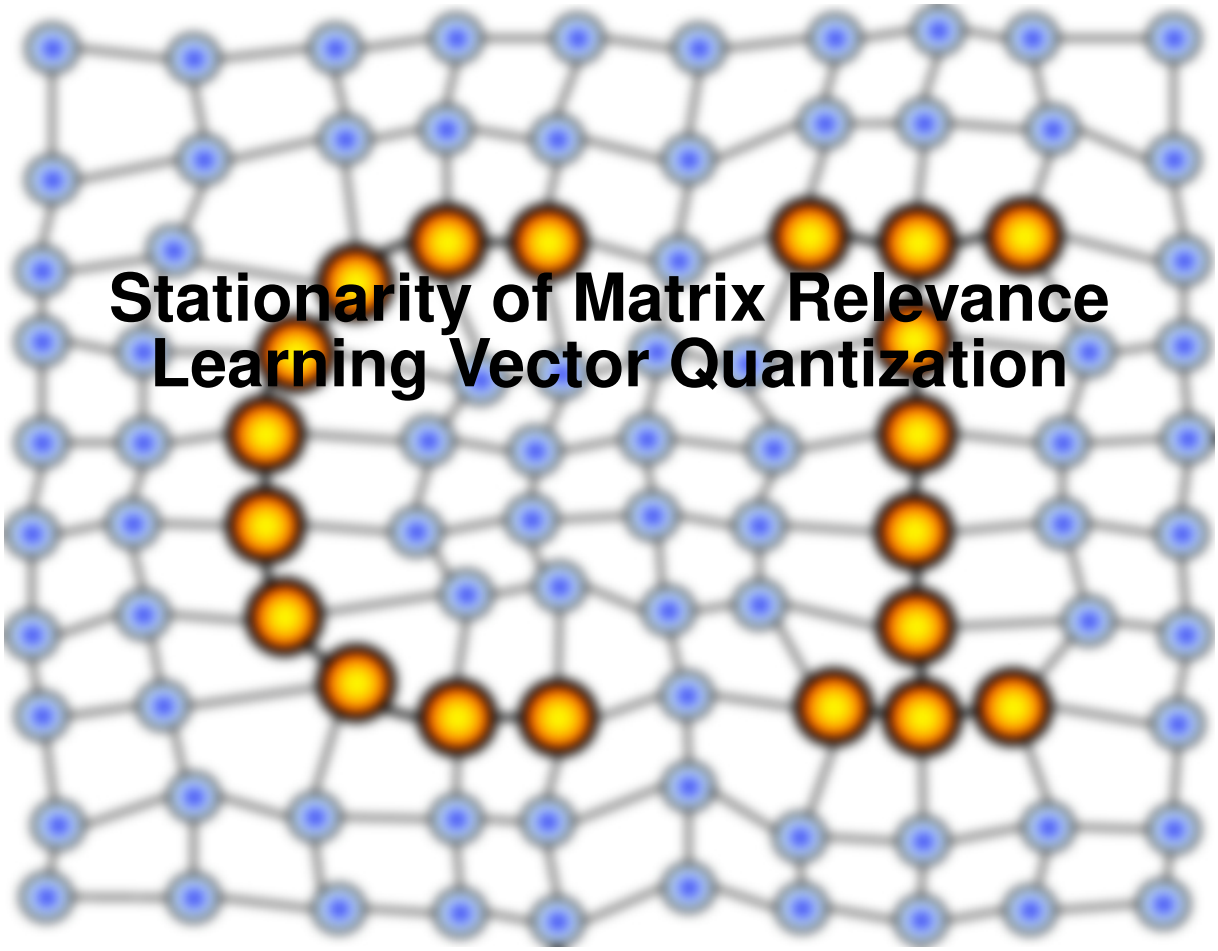
MACHINE LEARNING REPORTS



Stationarity of Matrix Relevance Learning Vector Quantization

Report 01/2009
Submitted: 26.06.2009
Published: 30.06.2009

M. Biehl[1], B. Hammer[2], F.-M. Schleif[3], P. Schneider[1] and Thomas Villmann[4]
(1) University of Groningen, Institute for Mathematics and Computing Science
P.O. Box 407, 9700 AK Groningen - The Netherlands
(2) Clausthal University of Technology, Institute of Computer Science
Julius Albert Strasse 4, 38678 Clausthal-Zellerfeld - Germany
(3) University of Leipzig, Department of Medicine
Semmelweisstrasse 10, 04103 Leipzig - Germany
(4) University of Applied Sciences Mittweida, Department of MPI
Technikumplatz 17, 09648 Mittweida - Germany

Abstract

We investigate the convergence properties of heuristic matrix relevance updates in Learning Vector Quantization. Under mild assumptions on the training process, stationarity conditions can be worked out which characterize the outcome of training in terms of the relevance matrix. It is shown that the original training schemes single out one specific direction in feature space which depends on the statistical properties of the data relative to the approached prototype configuration. Appropriate regularization terms can be used to favor full rank relevance matrices and to prevent oversimplification effects. The structure of the stationary solution is derived, giving insight into the influence of the regularization parameter.

Machine Learning Reports,Research group on Computational Intelligence,
http://www.uni-leipzig.de/~compint

1 Introduction

Similarity based methods play an impportant role in supervised and unsupervised machine learning tasks, such as classification, regression or clustering. For a recent overview see, for instance, [1]. Most methods employ pre-defined measures which are used to evaluate a generalized distance between vectors in feature space. The by far most popular choices are the standard Minkowski measures and generalizations thereof such as weighted Euclidean or Manhattan distance.

The key problem is, of course, the identification of a suitable measure which is appropriate for the problem at hand. Available insights into the scaling of different features or relations between them can be taken advantage of in the construction of a specific weighting scheme, for instance. However, frequently, a priori knowledge is limited and it is unclear which choice would facilitate good performance.

A particular attractive concept is that of adaptive distance measures which change in the course of learning. In this report we discuss adaptive metrics in the context of Learning Vector Quantization (LVQ) [2]. This family of supervised, distance based classification schemes has proven useful in a variety of practical problems. LVQ algorithms are particularly intuitive as they determine a set of *codebook vectors* or *prototypes* in feature space, which are then used to parameterize a *nearest prototype classifier* in the simplest case. LVQ is easy to implement and fast, it is a natural tool for multiclass classification and one can incorporate adaptive metrics in a straightforward way. Training provides at the same time typical representatives of the classes and a discriminative distance measure. Most powerful relevance LVQ (RLVQ) schemes employ different measures locally in different regions of feature space.

Once a particular parametric form of the distance measure is chosen, e.g. weighted Euclidean, relevance learning algorithms allow for data driven parameter adaptation, see [9] for the first realization in the context of LVQ. In cost function based learning, gradient descent or other optimization schemes can be applied to the distance measure as well, see for instance [7]. Here we consider quadratic measures used in LVQ systems for N-dimensional data which are parameterized by an $(N \times N)$ relevance matrix [3, 4, 5]. This generalization of Euclidean distance allows for the weighting of single features and, in addition, takes into account pairwise correlations between the features through off-diagonal matrix entries.

While metrics adaptation has proven useful in many practical applications, a better theoretical understanding is desirable. Essential questions concern, for instance, convergence behavior of the algorithms and uniqueness of the obtained solutions. We investigate properties of stationary relevance matrices in metrics adaptation. We show that generic update rules display a tendency to yield a low rank measure which takes into account a single direction in feature space, asymptotically. On the one hand, this effect can help to avoid overfitting effects in practical situations since it limits the complexity of the distance measure. On the other hand, this tendency can lead to deteriorating performance due to over-simplification and to numerical instabilities as matrices become singular. Regularization methods can be applied to cure these problems, we consider a particular strategy which allows to control the complexity of the relevant eigenspectra continuously [10, 11].

In the next section we introduce a heuristic local metrics adaptation scheme which is based on LVQ1 [2] and serves as an example for the discussion of stationarity. In Sec. 3 the stationarity of unregularized training is discussed while the effect of the

regularization is studied in Section 4. Section 5 summarizes our results and concludes with a brief discussion. In the Appendices details of the analysis are given and we show how our results carry over to global matrix updates.

2 Matrix relevance updates in LVQ

Matrix relevance learning has been introduced in [3, 4] as a modification of Generalized LVQ (GLVQ) [6, 7]. For simplicity, we study here the convergence of matrix learning in the framework of the basic LVQ1 algorithm [2] and its extension by local relevance matrices. It can be viewed as a generalization of the Relevance LVQ (RLVQ) scheme presented in [9] and the formulation is particularly clear-cut in the context of this heuristic prescription.

We will discuss the convergence of the below defined matrix updates under simplifying assumptions in the following section. Note, however, that these considerations should carry over to many similar and related supervised and unsupervised learning algorithms.

2.1 Generalized distance measure

Several similarity based methods of unsupervised or supervised learning use a generalized Euclidean distance measure of the form

$$d(\mathbf{y},\mathbf{z}) = (\mathbf{y}-\mathbf{z})^\top \Lambda\,(\mathbf{y}-\mathbf{z}) \quad \text{for } \mathbf{y},\mathbf{z} \in I\!R^N. \tag{1}$$

We require here that it fulfills the conditions $d(\mathbf{y},\mathbf{y}) = 0$ and $d(\mathbf{y},\mathbf{z}) = d(\mathbf{z},\mathbf{y}) \geq 0$ for all $\mathbf{y},\mathbf{z}$ with $\mathbf{y} \neq \mathbf{z}$. These are conveniently satisfied by assuming the following parameterization:

$$\Lambda = \Omega\Omega^\top \text{ i.e. } d(\mathbf{y},\mathbf{z}) = (\mathbf{y}-\mathbf{z})^\top \Omega\Omega^\top (\mathbf{y}-\mathbf{z}) = \sum_{i,j,l=1}^{N} (y_i - z_i)\,\Omega_{il}\Omega_{jl}\,(y_j - z_j). \tag{2}$$

Here, the $N \times N$-matrix $\Omega^\top$ defines a linear mapping to a space in which standard Euclidean distance can be applied.

It is important to note that, given Λ, the matrix Ω is not uniquely defined by Eq. (2). The distance measure is, for instance, invariant under rotations or reflections in feature space and $\Lambda = \Omega\Omega^\top$ can have many solutions, in general. The unique, positive symmetric root with $\Omega^\top = \Omega$ and $\Lambda = \Omega\Omega$ could be treated as a special case, for instance.

Mainly, we will consider unrestricted matrices Ω without imposing symmetry or other constraints on the structure. In this case, the elements of Ω can be varied independently. The derivative with respect to an arbitrary entry is

$$\frac{\partial d(\mathbf{y},\mathbf{z})}{\partial \Omega_{km}} = \sum_j (y_k - z_k)\Omega_{jm}(y_j - z_j) + \sum_i (y_i - z_i)\Omega_{im}(y_k - z_k) = 2(y_k - z_k)\left[\Omega^\top(\mathbf{y}-\mathbf{z})\right]_m \tag{3}$$

In matrix notation this reads

$$\frac{\partial d(\mathbf{y},\mathbf{z})}{\partial \Omega} = 2\,(\mathbf{y}-\mathbf{z})\left[\Omega^\top(\mathbf{y}-\mathbf{z})\right]^\top = 2\,(\mathbf{y}-\mathbf{z})\,(\mathbf{y}-\mathbf{z})^\top\,\Omega \tag{4}$$

which will be the basis of the analysed adaptation schemes.

Note that the derivative with respect to a symmetric matrix would read

$$\frac{\partial d(\mathbf{y},\mathbf{z})}{\partial \Omega} = 2\,(\mathbf{y}-\mathbf{z})(\mathbf{y}-\mathbf{z})^\top \Omega + 2\,\Omega(\mathbf{y}-\mathbf{z})(\mathbf{y}-\mathbf{z})^\top - 2\,I \circ \big((\mathbf{y}-\mathbf{z})(\mathbf{y}-\mathbf{z})^\top \Omega\big) \tag{5}$$

which takes into account that off-diagonal elements occur twice in a symmetric quadratic form.[1] Here I is the N-dim. identity matrix and $A \circ B$ is the element-wise or Hadamard product of matrices A and B, see [12] for a compilation of derivatives with respect to generic and structured matrices.

2.2 Local Matrix LVQ1 (LMLVQ1)

The following summarizes a heuristic extension of LVQ1 along the lines of relevance matrix learning for a set of prototypes $\{\mathbf{w}^k\}_{k=1}^M$ carrying class labels $\{s^k\}_{k=1}^M$. The distance of ξ from a particular prototype $\mathbf{w}_k$ is evaluated using a local matrix $\Lambda^k = \Omega^k {\Omega^k}^\top$:

$$d(\mathbf{w}^k, \xi) = (\xi - \mathbf{w}^k)^\top \Omega^k {\Omega^k}^\top (\xi - \mathbf{w}^k). \tag{6}$$

Upon presentation of a single example ξ from class σ, the update rule reads

(1) determine the winning prototype $\mathbf{w}^*$ with $\quad d(\mathbf{w}^*, \xi) = \min_k \{d(\mathbf{w}_k, \xi)\}$

(2) update $\mathbf{w}^*$ as in LVQ1: $\quad \mathbf{w}^* \to \mathbf{w}^* + \eta_w\, \psi(s^*, \sigma)\, \Omega^* {\Omega^*}^\top\, (\xi - \mathbf{w}^*)$
where the pre-factor $\psi(s, \sigma) = 1$ if $s = \sigma$ and $\psi(s, \sigma) = -1$ else. The learning rate η_w controls the magnitude of the prototype update which is along the gradient of $d(\mathbf{w}, \xi)$ w.r.t. $\mathbf{w}$.

(3) update the unrestricted, local matrix Ω^* attached to the winner according to

$$\Omega^* \to \Omega^* - \eta\, \psi(s^*, \sigma)\, \frac{1}{2}\, \frac{\partial d(\mathbf{w}^*, \xi)}{\partial \Omega^*} = \Omega^* - \eta\, \psi(s^*, \sigma)\, \mathbf{x}\, \mathbf{x}^\top\, \Omega^*, \tag{7}$$

where we have introduced the shorthand

$$\mathbf{x} = (\xi - \mathbf{w}^*). \tag{8}$$

In addition, we normalize the matrix after each step such that $\sum_i \Lambda^*_{ii} = \sum_{ij} (\Omega^*{}_{ij})^2 = 1$ remains satisfied. Note that the learning rate η for matrix updates is frequently chosen different from (smaller than) η_w.

The effect of this heuristic update is that, if the same or a similar feature vector is presented thereafter, a smaller (larger) distance will be observed if the class label σ agrees (disagrees) with that of the winning prototype, respectively.

As an alternative, we can consider the update of a symmetric Ω as a special case. While the corresponding gradient step could be derived from Eq. (5) we resort to a somewhat simpler, symmetrized version of Eq. (7):

($3'$) update the local symmetric matrix Ω^* attached to the winner according to

$$\Omega^* \to \Omega^* - \eta\, \psi(s^*, \sigma)\, \frac{1}{2}\, \left[\Omega^*\, \mathbf{x}\, \mathbf{x}^\top + \mathbf{x}\, \mathbf{x}^\top\, \Omega^*\right] \tag{9}$$

[1] We would like to thank Sander Land for drawing our attention to this point.

Again, Ω^* is normalized after each training step.

Note that Eq. (9) performs larger updates of diagonal elements than the gradient, Eq. (5), would suggest for symmetric matrices. However, (9) still constitutes a descent step with respect to $d(\mathbf{w}^*, \xi)$ and is easily analysed in the same framework as the generic update (7). Note that one could also modify (9) by using a learning rate $\eta/2$ for the diagonal elements Ω_{ii} in order to recover true gradient step in d.

2.3 Regularized updates

As we will discuss in the following sections, the updates introduced above result in singular matrices Ω and Λ, generically, which can lead to oversimplification effects [10, 11]. One can force Λ to maintain full rank, i.e. remain non-singular, by means of a regularization technique which was introduced in [10, 11].

Note that, under the normalization constraint $\mathrm{Tr}\Lambda = 1$, $\det\Lambda$ is maximized if all eigenvalues of Λ are equal. Therefore, we extend the matrix update by a contribution in the direction of the gradient of $\ln\det\Lambda$. For unrestricted matrices Ω without symmetry constraints the latter reads in matrix form [12]

$$\frac{\partial \ln\det\Omega\Omega^\top}{\partial\Omega} = 2\,\Omega^{-\top} \tag{10}$$

with the shorthand $\Omega^{-\top} = (\Omega^{-1})^\top$. Its contribution to the training prescription is controlled by an additional parameter μ. Before normalization the training step for the matrix attached to the winning prototype reads

$$\Omega^* \to \Omega^* \;-\; \eta\,\psi(s,\sigma)\,\mathbf{x}\mathbf{x}^\top\,\Omega^* \;-\; \mu\,\Omega^{*-T}. \tag{11}$$

A similar, cost function based algorithm is presented and discussed in greater detail in [10, 11]. Here we study the stationarity of the simpler regularized LMRLVQ1 update as a first example.

In case of symmetric matrices Ω one could base the regularization term on the gradient

$$\frac{\partial \ln\det\Omega\Omega}{\partial\Omega} = 4\,\Omega^{-1} - 2\,I\circ\Omega^{-1}. \tag{12}$$

For simplicity and in analogy with the previous section, we will consider the naïvely symmetrized version of Eq. (11),

$$\Omega^* \to \Omega^* \;-\; \frac{1}{2}\,\eta\,\psi(s,\sigma)\left(\mathbf{x}\mathbf{x}^\top\Omega^* + \Omega^*\mathbf{x}\mathbf{x}^\top\right) - \mu\,\Omega^{*-1}, \tag{13}$$

instead.

For high-dimensional data, costly matrix inversions might constitute a problem in practice. It is, however, possible to devise schemes that iteratively update an approximation of the inverse. We will address this point in a forthcoming publication.

3 Stationarity of unregularized matrix updates

For the following arguments, we make the simplifying assumption that the prototypes can be considered stationary, either because we set $\eta_w = 0$ explicitly, or because they have converged after an initial training phase.

In addition, we assume that the assignment of feature vectors to the prototypes does not change even though the distance measure is modified. One scenario in which this condition is satisfied is that of data drawn from well separated clusters – which may contain examples from different classes. Strictly speaking, the following applies only to cases where finite gaps separate the clusters of data which are assigned to different prototypes. Generically, this should be the case for finite training sets under mild assumptions on the positions of the input vectors in feature space.

Assuming that the assignment of inputs to prototypes is fixed, we can consider a single prototype $\mathbf{w}^* = const.$, its local relevance matrix Ω^*, and the fixed set of vectors ξ assigned to $\mathbf{w}^*$. In an update step (which concerns $\mathbf{w}^*$), one example is selected randomly and the matrix is updated as in Eq. (7). or (9).

In the following section, we analyse the above defined matrix updates for a given set of examples and derive the corresponding stationary state.

3.1 Unrestricted matrix update

In the following, we omit the index that identifies the winning prototype $\mathbf{w}$ with label s. It is implicitly understood that the local matrix is updated only from the corresponding subset of data. In time step t of this training (sub-) process an example ξ is presented. Before normalization, the update of the corresponding matrix in time step t of the training reads

$$\Omega(t+1) \;\sim\; \Omega(t) - \eta\,\psi(s,\sigma)\,\mathbf{x}\,\mathbf{x}^\top\,\Omega(t)$$

with the abbreviation $\mathbf{x}$ defined in Eq. (8).

We will consider the update on average over all examples in the relevant subset of data. This corresponds to interpreting the single example update as a stochastic gradient descent with respect to the sum of the distances $d(\xi^\mu, \mathbf{w})$ over the subset of data assigned to $\mathbf{w}$. We introduce the corresponding shorthand notation

$$\langle \cdots \rangle \;=\; \sum_\mu (\cdots)\phi(\xi^\mu,\mathbf{w}) \Big/ \sum_\mu \phi(\xi^\mu,\mathbf{w})$$

where the sum is formally over all examples and the indicator function $\phi(\xi,\mathbf{w}) = 1$ if $\mathbf{w}$ is the prototype closest to ξ and $\phi(\xi,\mathbf{w}) = 0$ else. For the averaged update step we obtain

$$\Omega(t+1) \;\sim\; \Omega(t) - \eta\,\left\langle\, \psi(s,\sigma)\;\mathbf{x}\,\mathbf{x}^\top \right\rangle \Omega(t) \;=\; [I - \eta\, C\,]\;\Omega(t), \tag{14}$$

where the symmetric matrix C is defined as

$$C \;=\; \left\langle\, \psi(s,\sigma^\mu)\;\mathbf{x}\,\mathbf{x}^\top \right\rangle. \tag{15}$$

Under the assumption of constant prototype posiontions and assignments $\phi(\xi,\mathbf{w})$, C itself remains constant in the iteration. Note that t in Eq. (14) counts only non-zero updates of the matrix Ω, i.e. only updates from examples assigned to $\mathbf{w}$.

It is important to point out that, in general, C cannot be interpreted as a positive (semi-) definite covariance matrix, since it takes into account label information: Examples can contribute with positive or negative sign ψ.

We assume an ordering of the eigenvalues of C: $\lambda_1 < \lambda_2 < \lambda_3 \ldots < \lambda_N$ and exploit the fact that the set of eigenvectors forms an orthonormal basis $\{\mathbf{v}_j\}_{j=1}^N$ of $I\!R^N$. In the

presence of degeneracies $\lambda_i = \lambda_j$ the following arguments have to be modified slightly, considering the corresponding eigenspaces.

An unnormalized update of the form (14) will be dominated by the largest eigenvalue and corresponding eigenvector of the matrix $[I - \eta C]$. For sufficiently small η this eigenvalue is $(1 - \eta\lambda_1) > 0$, where λ_1 is the smallest eigenvalue of C. However, the iteration of Eq. (14) would yield either divergent behavior for $\lambda_1 < 0$ or the trivial stationary solution $\Omega \to 0$ with $t \to \infty$ for $\lambda_1 > 0$. In order to take the normalization into account explicitly, we rewrite Eq. (14) element-wise (omitting the time step t)

$$\Omega_{ij} \;\to\; \Omega_{ij} - \eta\,(C\,\Omega)_{ij} \tag{16}$$

and assume that the previous Ω was normalized. We obtain

$$\sum_{ij} \Omega_{ij}^2 \;\to\; \underbrace{\sum_{ij} \Omega_{ij}^2}_{=1} - 2\eta \sum_{ij} \Omega_{ij}(C\Omega)_{ij} + \eta^2 \sum_{ij} (C\Omega)_{ij}^2. \tag{17}$$

For small learning rate η the last term in Eq. (17) can be neglected. With the abbreviation $\kappa \equiv \sum_{ij} \Omega_{ij}(C\Omega)_{ij} = \sum_{ij} \Omega_{ji}^T(C\Omega)_{ij} = \mathrm{Tr}(\,\Omega^T C\Omega\,)$, the normalized update of Ω becomes

$$\Omega \;\to\; \frac{\Omega - \eta\, C\,\Omega}{\sqrt{1 - 2\eta\kappa}} \;\approx\; (\Omega - \eta\, C\Omega)\;(1 + \eta\,\kappa) \;\approx\; \Omega - \eta\,[C\,\Omega - \kappa\,\Omega] \tag{18}$$

where we neglect terms of order $\mathcal{O}(\eta^2)$. Note that the matrix C appears only linearly in (18). Hence, applying the normalization $\sum_{ij} \Omega_{ij}^2$ before taking the average $\langle \ldots \rangle$ over the subset of examples would have lead to the same result in the limit $\eta \to 0$.

Eq. (18) shows that the stationarity condition for the matrix update is of the form $C\,\Omega = \kappa\,\Omega$, implying that each column of the stationary matrix Ω is a multiple of some eigenvector $\mathbf{v}_j$ of C. Note that, apart from the specific normalization, the update is equivalent to a *von Mises* iteration [13] of the columns of Ω with respect to the matrix $[I - \eta\, C]$. Hence, the update will display linear convergence and the actual convergence speed depends on the gap between the two largest eigenvalues of $I - \eta\, C$ [13].

As we detail in Appendix A, one can show that all columns of the stationary matrix Ω are multiples of the eigenvector $\mathbf{v}_1$ corresponding to λ_1:

$$\Omega = \left[A^1\,\mathbf{v}_1, A^2\,\mathbf{v}_1, \ldots, A^N\,\mathbf{v}_1\right] \quad \text{with} \quad \sum_k (A^k)^2 = 1. \tag{19}$$

The normalization of the coefficients $A^k \in I\!R$ reflects that $\sum_{ij} \Omega_{ij}^2 = 1$.

Finally, we can work out the resulting matrix Λ obtained from the above solution:

$$\Lambda_{ij} \;=\; \left(\Omega\,\Omega^\top\right)_{ij} \;=\; \sum_k \Omega_{ik}\,\Omega_{jk} \;=\; \sum_k (A^k)^2\, v_1^{(i)} v_1^{(j)} \;=\; \left(\mathbf{v}_1\,\mathbf{v}_1^\top\right)_{ij}. \tag{20}$$

Hence, the resulting relevance matrix is simply $\Lambda = \Omega\Omega^\top = \mathbf{v}_1\mathbf{v}_1^\top$, given by the eigenvector of C that corresponds to its smallest eigenvalue. Note that, under the assumptions made, Λ is indeed unique, despite the freedom in selecting the coefficents A_1^k in Ω. Note also that the only non-zero eigenvalue of the resulting Λ is 1.

3.2 Symmetrized matrix updates

The above considerations have to be modified slightly for the symmetrized update rule, Eq. (9). The treatment is along the lines of the previous subsection and details are given in the Appendix B.1.

The stationary symmetric matrix is shown to be $\Omega = \mathbf{v}_1 \mathbf{v}_1^\top$ and the resulting relevance matrix coincides with the result (20) of the non-symmetrized update:

$$\Lambda = \Omega\,\Omega = \mathbf{v}_1\mathbf{v}_1^\top.$$

This confirms that the restriction to symmetric Ω does not constitute a limitation of matrix relevance learning [3]. Here, however, even the stationary Ω is unique due to the additional symmetry requirement.

4 Regularized matrix update

The reduction to a single, relevant direction can lead to over-simplified classifiers with poor performance in data sets that do require multi-dimensional representations [10, 11]. Furthermore, singular or nearly singular relevance matrices can lead to numerical instabilities in practical implementations of the algorithm.

A regularized version of matrix learning which circumvents these problems is addressed in the following. We extend the analysis to a regularized version of LMRLVQ1 which favors non-singular matrices Λ. Again, unrestricted and explicitly symmetrized updates can be considered.

4.1 Non-symmetric, regularized matrices

In analogy to the previous section we investigate the dynamics of a single, local matrix taking an average over the relevant subset of data. The regularized, average update without explicit symmetrization reads

$$\Omega \sim \Omega - \eta\, C\, \Omega + \eta\, \mu\, \Omega^{-T} \tag{21}$$

with the regularization parameter μ. Rewriting the update element-wise we obtain

$$\sum_{ij} \Omega_{ij}^2 \to \underbrace{\sum_{ij} \Omega_{ij}^2}_{=1} - 2\eta \underbrace{\sum_{ij} \Omega_{ij} (C\Omega)_{ij}}_{\sum_j (\Omega^\top C \Omega)_{jj}} + 2\mu\eta \underbrace{\sum_{ij} \Omega_{ij} (\Omega^{-\top})_{ij}}_{\sum_j (\Omega\Omega^{-1})_{jj} = Tr(I) = N} + \mathcal{O}(\eta^2).$$

Neglecting terms quadratic in η, the normalized update of Ω reads

$$\begin{aligned}\Omega &\to \frac{\Omega - \eta\, C\, \Omega + \mu\, \eta\, \Omega^{-\top}}{\sqrt{1 - 2\eta \sum_j (\Omega^T C \Omega)_{jj} + 2\mu\eta\, N}} \\ &\approx \Omega - \eta \Big[C\Omega - \Big(\sum_j (\Omega^\top C \Omega)_{jj} - \mu N \Big) \Omega - \mu \Omega^{-\top} \Big].\end{aligned} \tag{22}$$

With the abbreviation $\kappa = \sum_j (\Omega^\top C \Omega)_{jj} - \mu N$ the stationarity condition reads

$$C\,\Omega - \kappa\,\Omega = \mu\,\Omega^{-\top}.$$

Obviously, for $\mu = 0$, this reduces to the same eigenvalue problem for each column of Ω as discussed above. For $\mu \neq 0$, multiplication with $\Omega^\top$ yields the modified stationarity condition

$$(C - \kappa\, I)\, \Omega\Omega^\top = \mu\, I, \quad \text{with the formal solution } \Lambda = \Omega\Omega^\top = \mu\, (C - \kappa\, I)^{-1}. \tag{23}$$

4.2 Symmetric regularized matrices

The convergence of regularized training in the case of symmetric matrices $\Omega = \Omega^\top$ can be analysed analogously. In Appendix B.2 we show that the corresponding stationarity condition reads in this case

$$-\kappa\Omega^2 + \frac{1}{2}\,(C\Omega^2 + \Omega C\Omega) = \frac{1}{2}\,\big((C - \kappa I)\Omega^2 + \Omega(C - \kappa I)\Omega)\big) = \mu\, I, \tag{24}$$

where $\kappa = \sum_j (\Omega C \Omega)_{jj} - \mu N$. The formal solution is given by the symmetric

$$\Omega = \sqrt{\mu}\,(C - \kappa I)^{-1/2}. \quad \text{or, equivalently} \quad \Lambda = \Omega^2 = \mu\,(C - \kappa I)^{-1}. \tag{25}$$

In terms of Λ we have obtained the same stationary solution as for the unrestricted matrix update which confirms, once more, that imposing symmetry does not limit the flexibility of the adaptative metrics.

4.3 Stationary solution

From both stationarity conditions, Eqs. (23) and (25), we obtain the same solution:

$$\Lambda = \mu\,(C - \kappa\, I)^{-1}. \tag{26}$$

The symmetric matrix $(C - \kappa\, I)$ is invertible as long as κ does not coincide with one of the eigenvalues λ_i of C. Since Λ is strictly positive definite for $\mu > 0$, we can even restrict the discussion to $\kappa < \lambda_1$, where λ_1 is the smallest eigenvalue of C.

Note that $(C - \kappa\, I)$ has the same eigenvectors $\mathbf{v}_i$ as C and its eigenvalues are $(\lambda_i - \kappa)$. Hence, we can construct the inverse explicitly:

$$\Lambda = \sum_k \frac{\mu}{\lambda_k - \kappa}\, \mathbf{v}_k\, \mathbf{v}_k^\top \qquad \text{with } \mathrm{Tr}\,\Lambda = \sum_i \frac{\mu}{\lambda_i - \kappa} \stackrel{!}{=} 1, \tag{27}$$

where the normalization constraint determines κ, given the parameter μ and the set of $\{\lambda_k\}$. Alternatively, Eq. (27) can be solved for μ, given κ, and we obtain

$$\mu = \left(\sum_i \frac{1}{\lambda_i - \kappa}\right)^{-1} \text{ and } \Lambda = \sum_k c_k\, \mathbf{v}_k \mathbf{v}_k^\top \text{ with } c_k = \left(\sum_i \frac{\lambda_k - \kappa}{\lambda_i - \kappa}\right)^{-1}. \tag{28}$$

For every $\kappa < \lambda_1$, the corresponding solution is obtained by using the appropriate regularization parameter μ as given above.

In order to obtain further insights, we investigate several cases in greater detail: For positive definite C, i.e. $\lambda_1 > 0$, we can consider the special solution with $\kappa = 0$. We obtain the corresponding μ and Λ from the normalization constraint as

$$\mu = \frac{1}{\mathrm{Tr}\, C^{-1}} \text{ and } \Lambda = \frac{C^{-1}}{\mathrm{Tr}\, C^{-1}}. \tag{29}$$

Note that this special case can be interpreted as a Mahalonobis distance only for a well separated cluster which exclusively contains data from one class. Only in such an extreme setting with $\mathbf{w}^*$ in the cluster mean, C is the corresponding covariance matrix, see Sec. 3.1.

The case $\mu = 0$ has been treated separately in Sec. 3. There, $\kappa = \lambda_1$ and Eq. (27) can only be considered in the simultaneous limit $\kappa \to \lambda_1^-, \mu \to 0^+$ to recover the results.

For weak regularization, i.e. small but non-zero $\mu > 0$, we expect $\kappa \approx \lambda_1$, resulting in a dominating contribution of $\mathbf{v}_1\mathbf{v}_1^\top$, whereas the coefficients of $\mathbf{v}_j\mathbf{v}_j^\top$ for $j > 1$ should be relatively small. If we make the ansatz $\kappa = \lambda_1 - \epsilon$ with small $\epsilon > 0$ for $\mu \to 0$ we obtain from Eq. (28) the relation $\epsilon = \mu + \mathcal{O}(\mu^2)$ and coefficients

$$\begin{aligned} c_1 &= 1 - \sum_{i \geq 2} \frac{\mu}{\lambda_i - \lambda_1} + \mathcal{O}(\mu^2), \\ c_k &= \frac{\mu}{\lambda_k - \lambda_1} + \mathcal{O}(\mu^2) \quad \text{for} \quad k \geq 2. \end{aligned} \tag{30}$$

Neglecting terms of order $\mathcal{O}(\mu^2)$ the normalized stationary matrix is given by

$$\Lambda = \left(1 - \sum_{i \geq 2} \frac{\mu}{\lambda_i - \lambda_1}\right) \mathbf{v}_1\mathbf{v}_1^\top + \sum_{i \geq 2} \frac{\mu}{\lambda_i - \lambda_1} \mathbf{v}_i\mathbf{v}_i^\top. \tag{31}$$

As expected, the eigendirection corresponding to λ_1 dominates the distance measure. However, the influence of other eigenvectors $\mathbf{v}_k$ increases with μ and is inversely proportional to $(\lambda_k - \lambda_1)$.

Finally, we consider the limit of strong regularization, $\mu \to \infty$. Eq. (27) implies that $\kappa \sim -\mu N$ which yields

$$\Lambda = \frac{1}{N} \sum_k \mathbf{v}_k\mathbf{v}_k^\top = I/N.$$

Obviously, all eigenvectors contribute equally in this limit and the distance reduces to the simple Euclidean measure apart from the normalization factor $1/N$.

We would like to point out again that in an actual training process, the stationary Ω will not be unique. Some symmetric or non-symmetric solution of Eqs. (23) or (24) will be approached, depending on the variant of the algorithm and on initilization. The emerging distance measures, however, and the contributions of eigendirections $\mathbf{v}_k$ are uniquely described by the above results.

5 Summary and Conclusion

We have analysed the stationary behavior of a class of matrix updates in Relevance Learning Vector Quantization. We have exemplified the treatment in terms of the extension of basic LVQ1 by local relevance matrices. Our considerations show that unregularized matrix updates tend to select a single relevant direction in feature space.

The matrix C which governs the convergence properties depends, of course, on the positioning of prototypes and implicitly on the distance measures emerging from the learning process. As a consequence, our results do not imply the possibility of obtaining the distance measure beforehand and independent of the LVQ training. In general, the relevant matrix C and its eigenvectors cannot be directly constructed from

the data set as they emerge from the complex interplay between prototype positions and distance measure in the course of training.

For the same reason, the stationary distance measure, as specified by the matrix Λ, can depend on the initialization of the LVQ system in practical learning. While, given the matrix C and the learning rate η, the obtained Λ is unique, the stationary winner configuration and, thus, the matrix C can vary between randomized learning processe in the same data set.

In the extreme case of a well-separated, pure cluster of data from class s only, C is the positive definite covariance matrix with respect to $\mathbf{w}$ which will be approximately in the center of the cluster. In such a setting, training singles out the direction of minimal variance. Clearly, the smallest distances are measured along this direction, which favors correct classification. In the more general case of separated but mixed clusters, the eigendirection selected in the training will not directly correspond to its geometric properties but reflects the cluster composition. Then, the aim of obtaining small distances for data from class $\sigma = s$ competes with the objective of achieving large distances for examples from the other classes.

One can argue that, after successful training of the prototypes, the majority of data assigned to prototype $\mathbf{w}$ should belong to the same class $\sigma = s$. For *well-behaved* data sets and if only a few examples contribute with $\psi(s, \sigma) = -1$, the relevant matrix C will still be positive definite, typically. Note, however, that it is not necessary to make this assumption for the above arguments.

In summary, the results indicate that unregularized matrix updates yield matrices Ω and Λ of small rank. Asymptotically, the distance measure takes into account a single eigendirection of C, only. This implies that the effective number of degrees of free parameters reduces drastically from $\mathcal{O}(N^2)$ in the full matrix to $\mathcal{O}(N)$ in the stationary solution.

On the one hand, the rank reduction explains the observation that over-fitting does not seem to play a critical role in matrix relevance learning from real world data sets [3, 4]. On the other hand, over-simplification effects can be due to the selection of a single relevant direction in feature space.

Within the same formal framework, we have also analysed a regularized version of matrix updates which overcomes this potential problem. It is shown that a proper regularization enforces non-singularity of the stationary relevance matrix. For small regularization parameter μ the resulting matrix is still dominated by the eigenvector $\mathbf{v}_1$ corresponding to the smallest eigenvalue of C. By choice of μ one can control the influence of the other eigendirections continuously.

The fact that the rank of Λ remains N prevents numerical instabilities and over-simplified classification. Which choice of μ gives the best learning and generalization behavior depends, of course, on the problem at hand and properties of the data set. Standard validation procedures could be applied to optimize the parameter in practice.

In forthcoming studies we will investigate the precise stationarity conditions for several prototype based algorithms, including cost-function oriented variants of LVQ and schemes for unsupervised learning. Since the mathematical structure of various prescriptions is similar, low rank stationary distance measures constitute a wide-spread phenomenon in similarity based learning. Also, the regularization strategy discussed here is not specific to LVQ relevance learning and provides a universal tool for the prevention of over-simplification effects.

A Stationarity of unrestricted matrix updates

In order to obtain the stationary solution of the matrix update given in Eq. (18) we use the following ansatz for the columns of Ω:

$$\Omega = [z_1, z_2, \ldots, z_N] \quad \text{with} \quad z_k = \sum_j a_j^k \mathbf{v}_j. \tag{32}$$

This is possible because the $\mathbf{v}_j$ provide an orthonormal basis of $I\!R^N$. Inserting this ansatz into Eq. (18) allows us to read off an update scheme for the column $z_k(t) = (\Omega(t))_{(k)}$:

$$\underbrace{\sum_j a_j^k(t+1)\mathbf{v}_j}_{(\Omega(t+1))_k} = \underbrace{\sum_j a_j^k(t)\mathbf{v}_j}_{(\Omega(t))_{(k)}} - \eta\Big[\underbrace{\sum_j a_j^k(t)\lambda_j\mathbf{v}_j}_{(C\Omega(t))_{(k)}} - \kappa(t)\underbrace{\sum_j a_j^k(t)\mathbf{v}_j}_{(\Omega(t))_{(k)}}\Big]. \tag{33}$$

Due to the orthogonality of eigenvectors, this corresponds to the following evolution of coefficients a_j^k:

$$a_j^k(t+1) \;=\; a_j^k(t)\;\left[1 - \eta\;(\lambda_j - \;\kappa(t))\right]. \tag{34}$$

Now let $\tilde{\kappa}$ and A_j^k be the stationary values of κ and a_j^k, respectively. Since a unique value of $\tilde{\kappa}$ has to satisfy $0 = A_j^k(\lambda_j - \tilde{\kappa})$ for all j, the only non-trivial solutions correspond to $\tilde{\kappa} = \lambda_m$ for one particular m with $A_m^k \neq 0$, but $A_j^k = 0$ for all $j \neq m$.
Then, Eq. (34) reads close to stationarity

$$a_j^k(t+1) \approx a_j^k(t) \cdot [1 - \eta(\lambda_j - \lambda_m)]\,. \tag{35}$$

Now let us assume that $m > 1$. In this case, the factor $[1 - \eta(\lambda_j - \lambda_m)] > 1$ for all $j < m$. Small deviations of the corresponding a_j^k from $A_j^k = 0$ would grow in the iteration, indicating an instability. Hence, the only stable, consistent solution is

$$m = 1, \qquad \tilde{\kappa} = \lambda_1, \qquad A_1^k \neq 0, \qquad A_j^k = 0 \quad \text{for all} \;\; j > 1.$$

This leads immediately to Eq. (19) where we omit the subscript 1 of the A_1^k for brevity.

B Stationarity of symmetrized updates

B.1 Unregularized matrix learning

For the symmetrized update (9) we obtain, instead of Eqs. (16) and (17),

$$\Omega_{ij} \sim \Omega_{ij} - \frac{1}{2}\eta\Big((C\Omega)_{ij} + (C\Omega)_{ji}\Big) \tag{36}$$

and

$$\sum_{ij}\Omega_{ij}^2 \to \underbrace{\sum_{ij}\Omega_{ij}^2}_{=1} - \eta\underbrace{\sum_{ij}\Omega_{ij}(C\Omega)_{ij}}_{\mathsf{Tr}(\Omega C\Omega)\equiv\kappa} - \eta\underbrace{\sum_{ij}\Omega_{ij}(C\Omega)_{ji}}_{\kappa} + \mathcal{O}(\eta^2). \tag{37}$$

Terms of order $\mathcal{O}(\eta^2)$ are neglected and the two term terms linear in η are equal. The update of Ω taking the normalization into account reads

$$\begin{aligned}\Omega \;\to\; \frac{\Omega - \frac{1}{2}\eta(C\Omega + \Omega C)}{\sqrt{1-2\eta\kappa}} \;&\approx\; \left(\Omega - \frac{1}{2}\eta\,(C\Omega + \Omega C)\right)\;(1+\eta\,\kappa) \\ &\approx\; \Omega - \frac{1}{2}\,\eta\,[\,C\Omega + \Omega C - 2\kappa\Omega\,] + \mathcal{O}(\eta^2). \end{aligned} \tag{38}$$

Here, the stationarity condition for symmetric Ω becomes $C\Omega + \Omega C = 2\kappa\Omega$.

Now we exploit the fact that a symmetric Ω can be written as a linear combination of the matrices $\mathbf{v}_l\mathbf{v}_m^T$ constructed from eigenvectors of C:

$$\Omega = \sum_{lm} a_{lm}\mathbf{v}_l\mathbf{v}_m^T \;\text{ with }\; a_{lm} = a_{ml}. \tag{39}$$

Inserting the ansatz in Eq. (38) we can re-write the update scheme for Ω as

$$\begin{aligned}\sum_{jk} a_{jk}(t+1)\mathbf{v}_j\mathbf{v}_k^\top &= \sum_{jk} a_{jk}(t)\mathbf{v}_j\mathbf{v}_k^\top \\ &\quad - \frac{\eta}{2}\left[\sum_{jk} a_{jk}(t)\lambda_j\mathbf{v}_j\mathbf{v}_k^\top + \sum_{jk} a_{jk}(t)\lambda_k\mathbf{v}_j\mathbf{v}_k^\top - 2\kappa(t)\sum_{jk} a_{jk}(t)\mathbf{v}_j\mathbf{v}_k^\top\right].\end{aligned}$$

Since the vectors $\mathbf{v}_j$ are orthonormal, the evolution of the coefficients a_{lm} reads

$$\begin{aligned} a_{lm}(t+1) &= a_{lm}(t) - \frac{\eta}{2}\,a_{lm}\,(\lambda_l + \lambda_m - 2\kappa(t)) \\ &= a_{lm}(t)\;\left[1 - \frac{\eta}{2}(\lambda_l + \lambda_m - 2\kappa(t))\right]. \end{aligned} \tag{40}$$

which preserves, of course, the symmetry $a_{lm} = a_{ml}$. Consequently, the stationary condition for non-zero a_{lm} becomes $\tilde{\kappa} = \frac{1}{2}(\lambda_l + \lambda_m)$. Since $\tilde{\kappa}$ has to be unique for all pairs (l, m), all coefficients except one have to become zero in the stationary state. Here we neglect potential degeneracies in the eigenvalue spectrum, as before.

We conclude by the same argument as in the non-symmetrized case that the stationary value is

$$\tilde{\kappa} = \lambda_1, \qquad A_{11} > 0, \qquad A_{jk} = 0 \;\text{ for }\; (j,k) \neq (1,1).$$

with λ_1 the smallest eigenvalue of matrix C. Consequently, Ω converges to the normalized matrix $\mathbf{v}_1\mathbf{v}_1^T$.

B.2 Regularized matrix learning

In analogy to Sec. 4.1, we derive here the stationary behavior for the regularized training of a symmetric matrix $\Omega = \Omega^\top$. The corresponding versions of Eqs. (16) and (17) read

$$\Omega_{ij} \to \Omega_{ij} - \frac{\eta}{2}\left((C\Omega)_{ij} + (\Omega C)_{ij}\right) + \mu\,\eta\,\Omega_{ij}^{-1}$$

and

$$\sum_{ij}\Omega_{ij}^2 \to \underbrace{\sum_{ij}\Omega_{ij}^2}_{=1} - \eta\underbrace{\sum_{ij}\Omega_{ij}(C\Omega)_{ij}}_{\sum_j(\Omega C\Omega)_{jj}} - \eta\underbrace{\sum_{ij}\Omega_{ij}(\Omega C)_{ij}}_{\sum_i(\Omega C\Omega)_{ii}} + 2\mu\eta\underbrace{\sum_{ij}\Omega_{ij}(\Omega^{-1})_{ij}}_{\sum_j(\Omega\Omega^{-1})_{jj}=N} + \mathcal{O}(\eta^2).$$

Consequently, the normalized update of Ω becomes

$$\begin{aligned}\Omega \to & \frac{\Omega - \frac{\eta}{2}(C\Omega + \Omega C) + \mu\eta\Omega^{-1}}{\sqrt{1 - 2\eta\sum_j (\Omega C\Omega)_{jj} + 2\mu\eta N}} \\ & \approx (\Omega - \frac{\eta}{2}(C\Omega + \Omega C) + \mu\eta\Omega^{-1})(1 + \eta\sum_j (\Omega C\Omega)_{jj} - \mu\eta N) \\ & \approx \Omega - \frac{\eta}{2}\Big[(C\Omega + \Omega C) - 2\underbrace{\Big(\sum_j (\Omega C\Omega)_{jj} - \mu N\Big)}_{\kappa}\Omega - 2\,\mu\Omega^{-1}\Big] + \mathcal{O}(\eta^2). \end{aligned} \tag{41}$$

Eq. (41) yields the stationarity condition

$$-\kappa\Omega + \frac{1}{2}(C\Omega + \Omega C) = \mu\Omega^{-1} \tag{42}$$

which leads to Eq. (24) after multiplication with Ω.

C Global Metrics Adaptation

In global matrix relevance LVQ1, one unique matrix Ω defines the distance measure employed for all data. Formally, the update is identical with (3) or $(3')$, respectively, replacing Ω^* by the one and the same Ω independent of the winning prototype. Again, we can write the matrix C as an empirical average, but now every example contributes to the update of Ω:

$$C^{global} = \frac{1}{P}\sum_{\mu=1}^{P}\sum_{k=1}^{M} \phi(\xi^\mu, \mathbf{w}^k)\ \psi(s^k, \sigma^\mu)\ (\xi - \mathbf{w}^k)\ (\xi - \mathbf{w}^k)^\top. \tag{43}$$

The respective winner is singled out by the indicator function ϕ with $\phi(\xi, \mathbf{w}) = 1$ if $\mathbf{w}$ is the prototype closest to ξ and $\phi(\xi, \mathbf{w}) = 0$ else. As above, $\psi(s, \sigma) = +1\,(-1)$ if the class label $s = \sigma$ ($s \neq \sigma$), respectively.

Note that C depends on Ω via the assignment of data to the winning prototype. Assuming stationarity of the prototypes and the *winner configuration* we can follow the lines of the analysis for local matrix updates and obtain analogous results. Here, the eigendirections of C reflect the geometry of clusters, their position relative to each other, and the cluster composition. While the mathematical structure of the stationary state is the same as for local distance measures, its interpretation is less obvious. In particular, C is defined with respect to several centers and does not resemble a simple covariance matrix, in general.

References

[1] M. Biehl, B. Hammer, M. Verleysen, T. Villmann (eds.), *Similiarity Based Clustering, Recent Developments and Biomedical Application*, Springer Lecture Notes in Artificial Intelligence Vol. 5400, Springer (2009)

[2] T. Kohonen, *Self-Organizing Maps*, Springer (1997)

[3] P. Schneider, M. Biehl, B. Hammer, *Relevance Matrices in LVQ*, in: M. Verleysen (ed.), *European Symposium on Artificial Neural Networks (ESANN 2007)*, d-side publishing, 37-42 (2007)

[4] P. Schneider, M. Biehl, B. Hammer, *Adaptive Relevance Matrices in Learning Vector Quantization*, Neural Computation (in press)

[5] P. Schneider, M. Biehl, B. Hammer, *Distance learning in discriminative vector quantization*, Neural Computation (in press)

[6] A.S. Sato, K. Yamada, *Generalized Learning Vector Quantization*, in: M.C. Mozer, D.S. Touretzky, M.E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, MIT Press (Cambridge, MA), 423-429 (1996)

[7] B. Hammer, T. Villmann, *Generalized Relevance Learning Vector Quantization*, Neural Networks **15**, 1059-1068 (2002)

[8] S. Seo, K. Obermayer, *Soft Learning Vector Quantization*, Neural Computation **15**, 1589-1604 (2003)

[9] T. Bojer, B. Hammer, D. Schunk, K. Tluk von Toschanowitz, *Relevance determination in Learning Vector Quantization*, in: M. Verleysen (ed.), *European Symposium on Artificial Neural Networks (ESANN 2001)*, d-facto publications, 271-276 (2001)

[10] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, M. Biehl, *Regularization in Matrix Relevance LVQ*, Machine Learning Reports (www.uni-leipzig.de/compint/mlr/mlr_03_2008.pdf), Rep. 02/2008, Univ. Leipzig (2008)

[11] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, M. Biehl, *Regularization in Matrix Relevance Learning*, submitted.

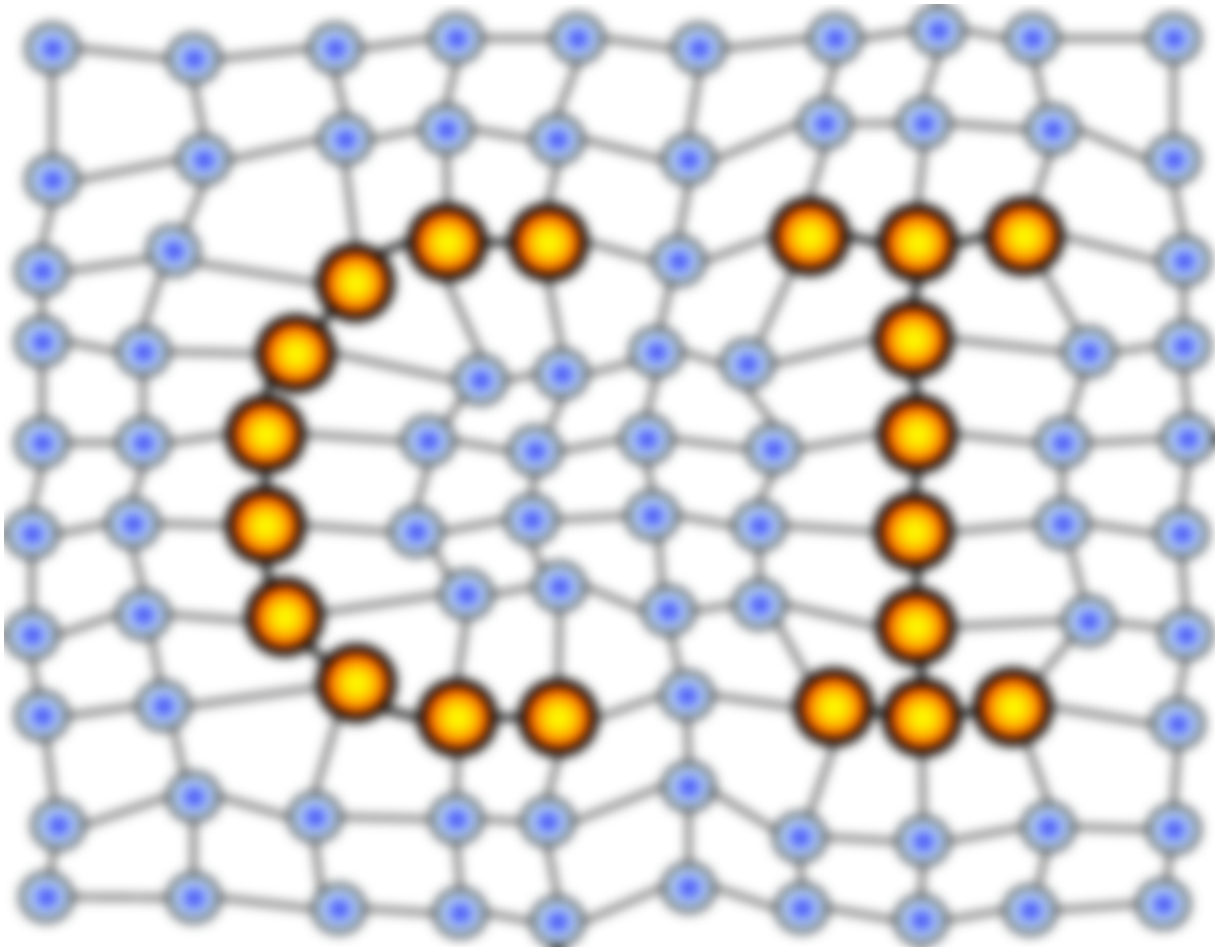
[12] K. Brandt Petersen, M. Syskind Pedersen, *The Matrix Cookbook*, www.matrixcookbook.com

[13] R. Plato, *Concise Numerical Mathematics,* American Mathematical Society (2003)

[14] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, M. Biehl, *Discriminative Visualization by Limited Rank Matrix Learning*, Machine Learning Reports (www.uni-leipzig.de/compint/mlr/mlr_03_2008.pdf), Rep. 03/2008, Univ. Leipzig (2008)

MACHINE LEARNING REPORTS

Report 01/2009



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann & Dr. rer. nat. Frank-Michael Schleif
Medical Department, University of Leipzig
Semmelweisstrasse 10, D-04103 Leipzig, Germany •
http://www.uni-leipzig.de/compint

▽ Acknowledgments

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