Presented at the NIPS 2011 Workshop, "Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity at Granada, Spain.

Mirror Descent for Metric Learning

Gautam Kunapuli
University of Wisconsin-Madison
1300 University Avenue
Madison, WI 53705
kunapuli@wisc.edu

Jude W. Shavlik
University of Wisconsin-Madison
1300 University Avenue
Madison, WI 53705
shavlik@cs.wisc.edu

Introduction. The concepts of similarity, distance or metric are central to a many well-known and popular algorithms such as k-means clustering [13], the nearest neighbor algorithm [6], locally-linear embedding [14], multi-dimensional scaling [7] and semi-supervised clustering [18]. While there are many approaches to metric learning, a large body of work is focussed on learning the Mahalanobis distance, which amounts to learning a transformation and computing the distance in the transformed space. Among these approaches are the work of Xing et al., [20], relevant components analysis [17], the large-margin nearest neighbor (LMNN) algorithm [19], Globerson and Roweis' method of collapsing classes [10], information-theoretic metric learning (ITML) [8] and BoostMetric [16]. Aside from the batch approaches above, online algorithms such as the online ITML algorithm [8] and the pseudo-metric online learning algorithm (POLA) [15] have proven successful.

All these approaches are characterized by diverse loss functions and projection methods, which naturally begs the question: is there a wider framework that can generalize many of these existing methods? In addition, ever persistent issues are those of scalability to large data sets and the question of kernelizability. Thus, we propose a unified approach to Mahalanobis metric learning: an online *regularized metric learning* algorithm based on the ideas of *composite objective mirror descent* (COMID) [9]. We propose to formulate the metric learning problem as a regularized *positive semi-definite matrix learning* problem, whose update rules can be derived using the COMID framework. This approach aims to be scalable, kernelizable, and admissible to many different types of Bregman and loss functions which allows for the tailoring of several different classes of algorithms. The most novel contribution is the use of the *trace norm*, which yields a sparse metric in its eigenspectrum, thus *simultaneously performing feature selection along with metric learning*.

Unifying Framework. The goal is to incrementally learn a squared Mahalanobis metric $d(\mathbf{x},\,\mathbf{z})^2 = (\mathbf{x}_t - \mathbf{z}_t)' M (\mathbf{x}_t - \mathbf{z}_t)$, given training data of the form $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^T$, where labels $y_t = \pm 1$ indicate similarity (+1) or dissimilarity (-1). We formulate the problem by defining a *metric function* (as in [15]) that measures the fidelity of the training data with respect to metric and bias pair ($M \succeq 0$, $\mu \geq 1$): $m(M, \mu;\, \mathbf{x}_t, \mathbf{z}_t, y_t) = y_t \left(\mu - (\mathbf{x}_t - \mathbf{z}_t)' M (\mathbf{x}_t - \mathbf{z}_t)\right)$. This now allows us to define several symmetric loss functions on the metric; for instance, the hinge-loss $\ell_t(M, \mu) = \max\{\, 0,\, 1 - m(\mathbf{x}_t, \mathbf{z}_t, y_t)\,\}$, or the logistic loss $\ell_t(M, \mu) = \log\left(1 + e^{-m(\mathbf{x}_t, \mathbf{z}_t, y_t)}\right)$. Several classes of algorithms arise from an appropriate choice of Bregman function [4], which is used to compute the updates using the composite objective mirror descent rule:

$$M_{t+1} = \arg\min_{M \succeq 0} \quad B_\psi(M,\, M_t) + \eta \left\langle \nabla_M \ell_t(M_t, \mu_t),\, M - M_t \right\rangle \quad + \eta\, \rho \;|||\, M |||, \tag{1}$$

$$\mu_{t+1} = \arg\min_{\mu \geq 1} \quad B_\psi(\mu,\, \mu_t) + \eta\, \nabla_\mu \ell_t(M_t, \mu_t)'\, (\mu - \mu_t), \tag{2}$$

Bregman functions (ψ) can be generalized to symmetric psd matrices, which enables us to define different types of Bregman divergences $B_\psi(\cdot,\,\cdot)$ over matrices. The Euclidean norm, $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$, induces the squared-Frobenius divergence, $B_\psi(X, Y) = \frac{1}{2}\|X - Y\|_F^2$; this leads to additive updates. The convex function $\psi(\mathbf{x}) = \sum_i x_i \log x_i - x_i$ induces the matrix generalization of KL divergence called the von Neumann divergence, $B_\psi(A, Y) = \mathsf{tr}\,(X \log X - X \log Y - X + Y)$; this leads to multiplicative updates. The function $\psi(\mathbf{x}) = -\sum_i \log x_i$ induces the matrix generalization of Itakura-Saito divergence called the Burg divergence $B_\psi(X, Y) = \mathsf{tr}\, XY^{-1} - \log\det XY^{-1} - n$;

this leads to inversive updates.

Feature Selection. If the SVD of $A = U\text{diag}(\boldsymbol{\sigma})V$, the trace norm of A is the matrix analogue of the ℓ_1-norm for vectors and is defined as the sum of singular values of the matrix i.e., $|||A||| = \mathbf{e}'\boldsymbol{\sigma}$. For symmetric M, minimizing $|||M|||$ attempts to produce a sparse eigenspectrum of M. This essentially amounts to *shrinking* the eigenvalues based on a threshold [2, 3], which in this case is the regularization constant on the nuclear norm term, $\eta\rho$. If the eigenvalue decomposition (EVD) of $M_t = V_t\Lambda_t V_t'$ (with $\Lambda_t = \text{diag}(\boldsymbol{\lambda}_t)$), the optimal solution to (1) can be computed as follows:

$$\begin{aligned} \text{(update eigendecomposition)} \quad & V_{t+1}\,\Lambda_{t+1}\,V_{t+1}' = V_t\nabla\psi(\Lambda_t)V_t' - \eta\nabla_M\ell_t(M_t,\,\mu_t) \\ \text{(shrink eigenvalues and project)} \quad & M_{t+1} = V_{t+1}\,\nabla\psi^{-1}\left(\,S_{\eta\rho}(\Lambda_{t+1})\,\right)\,V_{t+1}' \end{aligned}$$

where the *shrinkage operator* $S_{\eta\rho}(x) = \text{sign}(x)\,\max\{(|x| - \eta\rho),\,0\}$. The final $M(= L'L)$ is sparse in its spectrum, and we have that $L = V\sqrt{\Lambda}$. If M has $r < n$ non-zero eigenvalues, we can consider just the reduced eigendecomposition in calculating the distances: $\tilde{L} = V_r\Lambda_r$. Finally, the optimal solution to (2) gives: $\mu_{t+1} = \max\left\{\,\nabla\psi^{-1}\left(\nabla\psi(\mu_t) - \eta\,\nabla_\mu\ell_t(M_t,\mu_t)\right),\;\;1\right\}$.

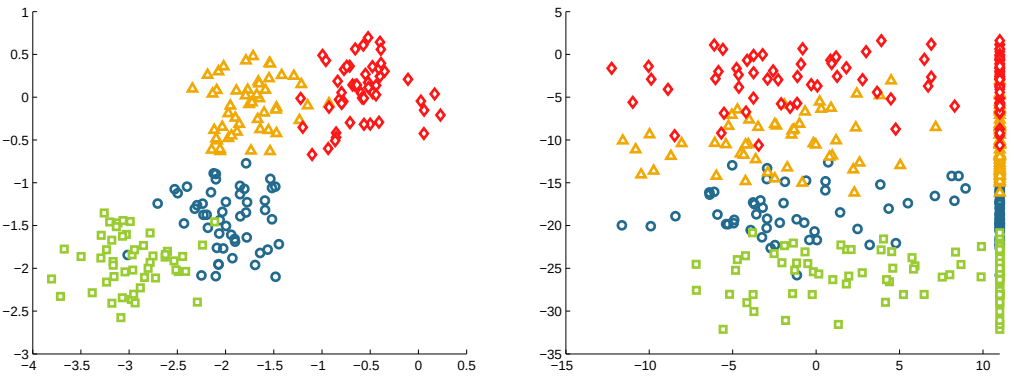
Efficient Computation and Scalability. While it may appear, at first glance, that a full eigen-decomposition is constructed at every step, this is not the case. For many loss functions, the gradient $\nabla_M\ell_t = \alpha(\mathbf{x}_t - \mathbf{z}_t)(\mathbf{x}_t - \mathbf{z}_t)'$, i.e., we only have to compute a *rank-one update to an existing eigendecomposition*. This is a well-studied problem, whose solution can be computed very efficiently [1, 12] by exploiting the Eigenvalue Interleaving Theorem [11]. In fact, if there are a large number of repeated eigenvalues then, owing to interleaving, all but one of those eigenvalues and eigenvectors in the update have to be computed. Thus, as the trace-norm shrinkage introduces more zeros into the problem, it becomes progressively more efficient. Finally, the non-repeating eigenvalues and their corresponding eigenvectors can be computed independently of each other making this approach is highly parallelizable. This latter aspect is currently under investigation.

Kernel Metric Learning. The framework of Chatpatanasiri et al., [5] can be used to kernelize this approach. Consider the (possibly nonlinear) mapping ϕ that maps all data $\mathbf{x}$ in the input space to a high-dimensional feature space, with an associated kernel function $\kappa(\cdot,\,\cdot)$. In feature space, the squared-Mahalanobis distance is computed as $d(\phi(\mathbf{x}),\;\phi(\mathbf{z}))^2 = (\phi(\mathbf{x}) - \phi(\mathbf{z}))'L'L(\phi(\mathbf{x}) - \phi(\mathbf{z}))$. Let A and Φ be the training data in input space and feature space respectively. Now, we parameterize $L' = \Phi G'$, and we have $d(\phi(\mathbf{x}),\;\phi(\mathbf{z}))^2 = (\phi(\mathbf{x}) - \phi(\mathbf{z}))'\Phi G'G\Phi'(\phi(\mathbf{x}) - \phi(\mathbf{z}))$. The vector $\Phi'\mathbf{x}$ is just the column of the kernel matrix corresponding to $\mathbf{x}$ and thus we have:

$$d_\kappa(\mathbf{x},\,\mathbf{z})^2 = (\boldsymbol{\kappa}(A,\mathbf{x}) - \boldsymbol{\kappa}(A,\mathbf{z}))'\,M\,(\boldsymbol{\kappa}(A,\mathbf{x}) - \boldsymbol{\kappa}(A,\mathbf{z})), \tag{3}$$

where $M = G'G$. Finally, once the matrix M is learnt, the Mahalanobis distance of some test point $\tilde{\mathbf{x}}$ with respect to a data point $\mathbf{x}$ can easily be computed as $d_\kappa^2(\tilde{\mathbf{x}},\,\mathbf{x}) = \boldsymbol{\kappa}(A,\tilde{\mathbf{x}}) - \boldsymbol{\kappa}(A,\mathbf{x}))'\,M\,(\boldsymbol{\kappa}(A,\tilde{\mathbf{x}}) - \boldsymbol{\kappa}(A,\mathbf{x}))$.

Figure 1: **A simple proof-of-concept**; (**left**) A 4-class 2d data set; 8 spurious dimensions are added to create a 10d training set for learning. Frobenius divergence and hinge loss are chosen, with $\eta = 5$ and $\rho = 15$; (**right**) Data projected on to the top two eigenvectors of the learned M, and on the far right, data projected onto the largest eigenvector of the learned M, showing that the approach is able to perform feature selection while learning an effective metric. Also, $\|\boldsymbol{\lambda}\|_0 = 2$, meaning that there are only two non-zero eigenvalues in the final M, and the algorithm was able to drop the spurious features.



References

[1] James R. Bunch, Christopher P. Nielsen, and Danny C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, 1978.

[2] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, March 2010.

[3] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, December 2009.

[4] Yair Al Censor and Stavros A. Zenios. *Parallel Optimization: Theory, Algorithms and Applications*. Oxford University Press, 1997.

[5] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. On kernelization of supervised mahalanobis distance learners. *Computing Research Repoisitory (CoRR)*, abs/0804.1441, 2008.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13(1):21–27, January 1967.

[7] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.

[8] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 209–216, Corvalis, Oregon, USA, 2007.

[9] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

[10] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems 18*, 2005.

[11] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences)(3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October 1996.

[12] Ming Gu and Stanley C. Eisenstat. A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. *SIAM Journal on Matrix Analysis and Applications*, 15(4):1266–1276, 1994.

[13] J. MacQueen. On convergence of k-means and partitions with minimum average variance. *Annals of Mathematical Statistics*, 36:1084ff, 1965.

[14] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[15] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the 21st International Conference on Machine Learning*, pages 94–102, New York, NY, USA, 2004. ACM.

[16] Chunhua Shen, Junae Kim, Lei Wang, and Anton van den Hengel. Positive Semidefinite Metric Learning with Boosting. In *Advances in Neural Information Processing Systems 22*, pages 629–633. MIT Press, 2009.

[17] Noam Shental, Tomer Hertz, Daphna Weinshall, and Misha Pavel. Adjustment learning and relevant component analysis. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 776–792, London, UK, UK, 2002. Springer-Verlag.

[18] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[19] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2006.

[20] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.