Distance Metric Learning by Optimization on the Stiefel Manifold

Ankita Shukla ANKITAS@IIITD.AC.IN
Saket Anand ANANDS@IIITD.AC.IN
IIIT Delhi, New Delhi, 110020.

Abstract

Distance metric learning has proven to be very successful in various problem domains. Most techniques learn a global metric in the form of a $n \times n$ symmetric positive semidefinite (PSD) Mahalanobis distance matrix, which has $\mathcal{O}(n^2)$ unknowns. The PSD constraint makes solving the metric learning problem even harder making it computationally intractable for high dimensions. In this work, we propose a flexible formulation that can employ different regularization functions, while implicitly maintaining the positive semidefiniteness constraint. We achieve this by eigendecomposition of the rank p Mahalanobis distance matrix followed by a joint optimization on the Stiefel manifold $\mathcal{S}_{n,p}$ and the positive orthant $\mathbb{R}^p_+$. The resulting nonconvex optimization problem is solved by employing an alternating strategy. We use a recently proposed projection free approach for efficient optimization over the Stiefel manifold. Even though the problem is nonconvex, we empirically show competitive classification accuracy on UCI and USPS digits datasets.

1. Introduction

Distance metric learning has received a lot of attention in the last decade owing to its success in many application domains like computer vision, classification and clustering. The default Euclidean distance equally weights each dimension in the input space and is often inadequate to capture the semantics of the data. Metric learning techniques use training examples to learn a distance function that is semantically consistent with the data. The commonly used Mahalanobis distance metric is characterized by a positive semidefinite (PSD) matrix that applies a linear transformation in the input space.

Many popular techniques (Davis et al., 2007; Kulis et al., 2009; Jain et al., 2012; Law et al., 2014) set up the metric learning problem in a constrained optimization framework. The imposed constraints capture the intuition that same class point pairs have small distances, while sample points from different classes have a large distance. The challenge in solving such problems is efficient projection on to the constraint space while maintaining the positive semidefiniteness of the Mahalanobis distance matrix.

Techniques that rely on projection on to $\mathcal{S}^n_+$ like (Weinberger & Saul, 2009; Xing et al., 2002; Law et al., 2014), usually require an eigen-decomposition or SVD in each iteration resulting in an additional cost of $\mathcal{O}(n^3)$. Projection free approaches like (Davis et al., 2007; Kulis et al., 2009; Jain et al., 2012) use special regularization functions leading to updates that guarantee positive semidefiniteness. In this paper, we explore a projection free approach that permits the flexibility to use different regularization functions.

The remainder of the paper is organized as follows. We provide a brief review of Stiefel manifold in Section 2 followed by the details of the proposed framework in Section 3. The experimental results are presented in Section 4. We discuss the extension of our parametrization in Section 5 and conclude in Section 6.

2. Optimization on the Stiefel Manifold

The set of $n \times p$ orthornormal matrices has a Riemannian structure and is called the Stiefel manifold, $\mathcal{S}_{n,p} = \{\mathbf{U} \in \mathbb{R}^{n\times p} : \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p, n \geq p\}$ (Edelman et al., 1998). An alternate interpretation is that of a quotient space of the orthogonal group $\mathbf{O}_n = \{\mathbf{Q} \in \mathbb{R}^{n\times n} : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n\}$,*i.e.* , $\mathcal{S}_{n,p} = \mathbf{O}_n/\mathbf{O}_{n-p}$.The tangent space at a point $\mathbf{U} \in \mathcal{S}_{n,p}$ is given by $T_\mathbf{U} = \{\Delta \in \mathbb{R}^{n\times p} : \Delta^\top\mathbf{U} = -\mathbf{U}^\top\Delta\}$.

Preliminary work. Under review by the Indian Workshop on Machine Learning (*i*WML). Do not distribute.

Wen and Jin (Wen & Yin, 2013) proposed an efficient constraint preserving update on the Stiefel manifold based on the Cayley transformation. The key idea is to relax the constraint of moving along geodesics and use retraction (Absil et al., 2008) to smoothly map a tangent vector to manifold. For a given point $\mathbf{U} \in \mathcal{S}_{n,p}$, let $\mathbf{G}$ be the gradient of $\mathcal{F}(\mathbf{U})$. A skew symmetric matrix $\mathbf{A} = \mathbf{G}\mathbf{U}^\top - \mathbf{U}\mathbf{G}^\top$ is then defined to get the following update in closed form (Wen & Yin, 2013).

$$\mathbf{V}(\tau) = \mathbf{Q}\mathbf{U} \text{ , where } \mathbf{Q} = \left(\mathbf{I} + \frac{\tau}{2}\mathbf{A}\right)^{-1}\left(\mathbf{I} - \frac{\tau}{2}\mathbf{A}\right) \quad (1)$$

Since we seek fast updates on the Stiefel manifold, we resort to this update scheme in designing our metric learning algorithm.

3. Proposed Framework

3.1. Notations

We denote the set of data points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$, with $\mathbf{x}_i \in \mathbb{R}^n$, $\;i = 1, \ldots, m$ and their corresponding class labels by ℓ_i.The n-dimensional real space is denoted by $\mathbb{R}^n$ and its positive orthant as $\mathbb{R}^n_+$. As we use pairwise constraints for metric learning, the constraint point pairs are grouped into two sets: $\mathcal{C}_s$, the set of similar pairs and $\mathcal{C}_d$, the set of dissimilar pairs. The complete set of constraints is denoted by $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_d$.

3.2. Problem Formulation

Our formulation for metric learning is based on the premise that PSD matrices have nonnegative eigenvalues and orthogonal eigenvectors. Thus we work with the representation of the rank p Mahalanobis matrix obtained by its eigendecomposition, $\mathbf{M}_{n\times n} = \mathbf{U}\mathbf{W}\mathbf{U}^\top$, $\;\mathbf{W} = \text{Diag}(\mathbf{w})$, where $\mathbf{U} \in \mathcal{S}_{n,p}$ is the orthonormal matrix of eigenvectors, and $\mathbf{w} \in \mathbb{R}^p_+$ is the vector of eigenvalues. We rewrite the metric learning problem as a joint optimization over $\mathcal{S}_{n,p} \times \mathbb{R}^p_+$ and use $||\mathbf{w}||_2^2$ as the regularization function, which is equivalent to $||\mathbf{M}||_F^2$, the squared Frobenius norm of $\mathbf{M}$.

The convex metric learning problem with the Frobenius norm regularizer is

$$\begin{aligned} \min_{\mathbf{M}\in\mathcal{S}^n_+} \quad & ||\mathbf{M} - \mathbf{M}_0||_F^2 & \quad (2)\\ \text{subject to} \quad & \mathbf{z}_{ij}^\top \mathbf{M}\mathbf{z}_{ij} \le s, & \forall\, i,j \in \mathcal{C}_s\\ & \mathbf{z}_{ij}^\top \mathbf{M}\mathbf{z}_{ij} \ge d, & \forall\, i,j \in \mathcal{C}_d \end{aligned}$$

where the vectors $\mathbf{z}_{ij}$ are the difference vectors $\mathbf{x}_i - \mathbf{x}_j$ obtained from the constraint pairs in $\mathcal{C}$, s and d are the desired distances for constraints in $\mathcal{C}_s$ and $\mathcal{C}_d$ respectively. $\mathbf{M}_0$ is the initial Mahalanobis distance matrix, often initialized to identity or the data covariance matrix. Since the problem in (2) could be infeasible, we introduce slack variables $\boldsymbol{\xi}$ and rewrite the relaxed problem as

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p_+, \mathbf{U}\in\mathcal{S}_{n,p}, \boldsymbol{\xi}\in\mathbb{R}^{|\mathcal{C}|}} & ||\mathbf{w} - \mathbf{w}_0||_2^2 + \gamma||\boldsymbol{\xi} - \boldsymbol{\xi}_0||_2^2 \quad (3)\\ \text{s.t. } & \mathbf{z}_{ij}^\top \mathbf{U}\text{Diag}(\mathbf{w})\mathbf{U}^\top \mathbf{z}_{ij} \le \xi_{ij}, \forall\, i,j \in \mathcal{C}_s\\ & \mathbf{z}_{ij}^\top \mathbf{U}\text{Diag}(\mathbf{w})\mathbf{U}^\top \mathbf{z}_{ij} \ge \xi_{ij}, \forall\, i,j \in \mathcal{C}_d \end{aligned}$$

where $\mathbf{w}_0$ is the vector of eigenvalues of $\mathbf{M}_0$. The initial vector of slack variables $\boldsymbol{\xi}_0$ of length $|\mathcal{C}|$ takes values $(\boldsymbol{\xi}_0)_{ij} = \{s, d\}$ based on whether $i,j \in \mathcal{C}_s$ or $i,j \in \mathcal{C}_d$. Note that the problem becomes nonconvex because of the domain of $\mathbf{U}$, which is the Stiefel manifold. The solution to the problem (3) yields $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{U}}$, which are used to reconstruct the Mahalanobis matrix $\widehat{\mathbf{M}} = \widehat{\mathbf{U}}\text{Diag}(\widehat{\mathbf{w}})\widehat{\mathbf{U}}^\top$.

Intuitively, the solution to (3) gives an orthogonal basis $\widehat{\mathbf{U}}$ of the p-dimensional subspace of $\mathbb{R}^n$, along with minimal scaling required to satisfy the distance constraints. While we cannot theoretically guarantee good generalization, our experiments in Section 4 show that results are competitive with metric learned by solving (2).

3.3. Algorithm

We solve the problem developed in (3) jointly over $\mathcal{S}_{n,p} \times \mathbb{R}^p_+$ by taking an alternating minimization approach. We initialize the algorithm with the Euclidean metric in a p-dimensional space with $\mathbf{w}_0$ as a vector of ones and $\mathbf{U}_0$ as a randomly picked point on $\mathcal{S}_{n,p}$.

The algorithm alternates between two steps: it solves for $\mathbf{U}$ in (3) while keeping $\mathbf{w}$ fixed and vice versa till the convergence criteria is satisfied.

The optimization problem in (3), with fixed $\mathbf{U}$ is a constrained least square problem. We write the corresponding unconstrained Lagrangian in (4) and obtain updates for $\mathbf{w}$ and $\boldsymbol{\xi}$ using KKT conditions for single constraint $(i,j) \in \mathcal{C}$.

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p_+, \boldsymbol{\xi}\in\mathbb{R}^{|\mathcal{C}|}} & ||\mathbf{w} - \mathbf{w}_0||_2^2 + \gamma||\boldsymbol{\xi} - \boldsymbol{\xi}_0||_2^2 \quad (4)\\ & + \lambda_{ij}^t\; y_{ij}(\mathbf{z}_{ij}^\top \mathbf{U}^t \text{Diag}(\mathbf{w}^t)\mathbf{U}^{t\top}\mathbf{z}_{ij} - \boldsymbol{\xi}_{ij}^t) \end{aligned}$$

Here $y_{ij} = -1$, if $i,j \in \mathcal{C}_s$ and $y_{ij} = 1$ if $i,j \in \mathcal{C}_d$ and $\lambda_{ij} \ge 0$ are the Lagrange multipliers. With an updated $\mathbf{w}$, we then solve for $\mathbf{U}$ for the same constraint pair (i,j). This is achieved by solving the following problem over the Stiefel manifold using updates in (1)

$$\min_{\mathbf{U}^t\in\mathcal{S}_{n,p}} \quad \lambda_{ij}^t\; y_{ij}(\mathbf{z}_{ij}^\top \mathbf{U}^t \text{Diag}(\mathbf{w}^t)\mathbf{U}^{t\top}\mathbf{z}_{ij} - \boldsymbol{\xi}_{ij}^t). \quad (5)$$

We pick another constraint and repeat the updates (4) and (5) till convergence. Since p could be as large as

n, and the updates (1) require inversion of a $2p \times 2p$ matrix (Wen & Yin, 2013), we use a block coordinate descent like strategy proposed in (Collins et al., 2014) to speed up this step. The key idea in (Collins et al., 2014) is to parametrize $\mathbf{U}$ by a point on a smaller Stiefel manifold. To obtain this parametrization, a set of $k \leq n$ rows $\mathcal{K}$, is selected from $\mathbf{U}$ to construct a smaller matrix $\mathbf{H}_{k\times p}$. If $\mathcal{I}$ is the set of linearly independent columns of $\mathbf{H}$, the parametrization is given as

$$\mathbf{U}(\mathbf{V}) = \begin{bmatrix} \mathbf{V}\mathbf{P}^{1/2} & \mathbf{V}\mathbf{P}^{1/2}\mathbf{R} \\ \mathbf{U}_{\bar{\mathcal{K}},\mathcal{I}} & \mathbf{U}_{\bar{\mathcal{K}},\bar{\mathcal{I}}} \end{bmatrix} \tag{6}$$

where $\mathbf{P} = \mathbf{H}_{\cdot,\mathcal{I}}^{\top}\mathbf{H}_{\cdot,\mathcal{I}}$ is positive definite, the $\bar{\mathcal{K}}$ and $\bar{\mathcal{I}}$ denote the complementary sets of $\mathcal{K}$ and $\mathcal{I}$ respectively. The matrix $\mathbf{R} \in \mathbb{R}^{|\mathcal{I}|\times|\bar{\mathcal{I}}|}$ is the linear transformation that maps $\mathbf{H}_{\cdot,\mathcal{I}}$ to $\mathbf{H}_{\cdot,\bar{\mathcal{I}}}$ and the orthonormal matrix $\mathbf{V}$ is a point on the smaller Stiefel manifold $\mathcal{S}_{k,|\mathcal{I}|}$. Collins (2014) show that a descent curve on $\mathcal{S}_{k,|\mathcal{I}|}$ gets mapped to the original manifold $\mathcal{S}_{n,p}$ by (6) in a direction of descent. As each block of k rows is updated on a smaller Stiefel manifold, we get efficient updates for $\mathbf{U}$. Moreover, this block coordinate descent type strategy can be parallelized by using disjoint sets of rows $\mathcal{K}_i$ such that $|\cup_i \mathcal{K}_i| \leq n$.

4. Experiments

We refer our formulation in (3) as SMML (Stiefel Manifold based Metric learning) and evaluate the learned metric against the Euclidean distance metric in terms of classification accuracy of a 3-nearest neighbor classifier on the UCI benchmark data sets and USPS digits. We compared the run time of SMML to solve (3) with that of SeDuMi (Sturm, 1999) to solve the relaxed version of (2). The experiments ran on a laptop with a core i7 quad core processor and 8 GB RAM with only two cores enabled. The threshold values s and d in (3) for similarity and dissimilarity constraints are set to the 1^{st} and 99^{th} percentile of all pairwise distances.

For high dimensional data, we optimize simultaneously over multiple $\mathcal{S}_{k,p}$ by selecting disjoint sets of $|\mathcal{K}|$ rows, whereas a sequential approach is used in case of low dimension data to avoid communication overheads between parallel threads. The optimal choice of $\mathcal{K}$ is found heuristically.

USPS digits [1] dataset consists of 16×16 grayscale images with 1100 images for each digit. The images are represented as 256 dimensional vector formed by concatenating the columns of image. The results for UCI datasets and USPS digits are summarized in Table 1.

We also compare our approach with convex formulation for USPS digits. We used PCA to reduce the dimensionality of the data with 99%, 95% and 90% energy. While the results for convex formulation and our proposed method are same for lower dimension representation with improvement in computation time. However, in case of higher dimensions, the learning with SeDuMi solver becomes computationally expensive in terms of memory usage with impractical run times. The accuracy and run time comparisons are summarized in Table 2.

Table 1. Classification Accuracy and Run Time Results

	USPS	Wine	Inosphere
# samples	11000	178	351
# constraints ($|\mathcal{C}|$)	900	630	900
# dimension	256	13	34
# dimension after PCA	114		
# Training points	150	45	30
# Testing points	2000	133	148
# classes	10	3	2
$|\mathcal{K}|$	24	5	8
Classification Accuracy%			
Euclidean	76.10	72.3	69.4
CVX	93.7	94.6	98
SMML	93.1	95	97.3
Run Time(in secs)			
CVX	846.3	7.2	22.6
SMML	346	13.6	39.2

Table 2. Results on USPS digits for different dimensions from PCA

	Run time(in mins)/Accuracy(%)	
PCA dimension, $|\mathcal{K}|$	CVX	SMML
38,10	10/**83**	8.7/82
66,22	54/80	17/**89.4**
152,30	-	39/**93.7**

5. Extension

We later, modify the metric learning formulation in (3) and formulate an unconstrained optimization problem. The two-term objective contains a hinge loss function to control distance constraint violations and a ℓ_2-norm regularizer to ensure smoothness. This formulation optimizes for all the constraints simultaneously instead of solving for only one constraint at every iteration. The formulation is given by

$$\min_{\mathbf{U}\in\mathcal{S}_{n,p},\mathbf{w}\in\mathbb{R}_+^p} \sum_{i,j=1}^{m} \left[y_{ij}(\mathbf{z}_{ij}^{\top}\mathbf{U}\mathrm{Diag}(\mathbf{w})\mathbf{U}^{\top}\mathbf{z}_{ij} - b_{ij}) \right]_+ + \alpha \left\|\mathbf{w} - \mathbf{w}_0\right\|_2^2 \tag{7}$$

Here, the term $[x]_+ = \max(0, x)$ is the hinge loss term that captures the degree of violation of constraints, b_{ij} are the corresponding target distances. and $\alpha > 0$ is a

[1] http://cs.nyu.edu/roweis/data.html

Table 3. Comparison of Classification Accuracies of different metric learning approaches

Dataset	Our	ITML	LMNN
USPS digits	**79.93**±1.54	75.27± 1.26	76.88±6.98
MIT scene	**60.7**±1.54	50.93± 1.72	45.15± 6.53

regularization parameter.

Due to space limitations, we are unable to describe the detailed algorithm for optimizing (7). We use an alternating strategy similar to (Liu et al., 2015).

We present our preliminary results on USPS digits and MIT scene [2] datasets in Table 3. We also compare our approach with state of the art metric learning approaches: ITML (Davis et al., 2007) and LMNN (Weinberger & Saul, 2009). For both the datasets we perform dimensionality reduction using PCA while preserving 95% energy.

6. Conclusion

We proposed a metric learning formulation that poses a joint optimization problem over $\mathcal{S}_{n,p} \times \mathbb{R}^p_+$ to find the eigenvectors and eigenvalues of the learned Mahalanobis distance matrix $\mathbf{M}$. We took an alternate minimization approach by iteratively updating the eigenvalues and the eigenvectors of $\mathbf{M}$ to solve the ensuing nonconvex problem. The proposed method showed competitive performance in classification tasks against convex formulation as well as state of the art methods. Since, our formulation allows the flexibility to replace the regularizer with any convex spectral function, we plan to explore the impact of other functions like $\log\det$ or Burg entropy in future.

Dual Submission

This paper is primarily a version of our work published in Proceedings of the 1st International Workshop on DIFFerential Geometry in Computer Vision for Analysis of Shapes, Images and Trajectories (DIFF-CV 2015) [3], pages 7.1-7.10. BMVA Press, September 2015. This work also includes our formulation accepted at ICIP 2016 [4] and will be available for Open Preview on IEEE Xplore a month before the conference.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

Collins, Maxwell D, Liu, Ji, Xu, Jia, Mukherjee, Lopamudra, and Singh, Vikas. Spectral clustering with a convex regularizer on millions of images. In *Computer Vision–ECCV 2014*, pp. 282–298. Springer, 2014.

Davis, Jason V, Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216. ACM, 2007.

Edelman, Alan, Arias, Tomás A, and Smith, Steven T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Jain, Prateek, Kulis, Brian, Davis, Jason V, and Dhillon, Inderjit S. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research*, 13(1):519–547, 2012.

Kulis, Brian, Sustik, Mátyás A, and Dhillon, Inderjit S. Low-rank kernel learning with bregman matrix divergences. *The Journal of Machine Learning Research*, 10:341–376, 2009.

Law, M.T., Thome, N., and Cord, M. Fantope regularization in metric learning. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1051–1058, June 2014.

Liu, Wei, Mu, Cun, Ji, Rongrong, Ma, Shiqian, Smith, John R., and Chang, Shih-Fu. Low-rank similarity metric learning in high dimensions. In *AAAI Conference on Artificial Intelligence (AAAI)*, Austin, Texas, USA, 2015.

Sturm, Jos F. Using SeDuMi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11:625–653, 1999.

Weinberger, K.Q. and Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

Wen, Zaiwen and Yin, Wotao. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

Xing, Eric P., Jordan, Michael I., Russell, Stuart J, and Ng, Andrew Y. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems NIPS*, pp. 521–528. 2002.

[2] http://people.csail.mit.edu/torralba/code/spatialenvelope/

[3] http://www-rech.telecom-lille.fr/diff-cv2015/

[4] http://2016.ieeeicip.org/default.asp