Bayesian Multitask Distance Metric Learning

Piyush Rai, Wenzhao Lian, Lawrence Carin
ECE Department, Duke University
Durham, NC 27708
{piyush.rai,wenzhao.lian,lcarin}@duke.edu

Abstract

We present a Bayesian approach for jointly learning distance metrics for a large collection of potentially related learning tasks. We assume there exists a relatively smaller set of *basis distance metrics* and the distance metric for each task is a *sparse*, positively weighted combination of these basis distance metrics. The set of basis distance metrics and the combination weights are learned from data. Moreover, taking a nonparametric Bayesian approach, the number of basis distance metrics need not be set *a priori*. Our proposed construction significantly reduces the number of parameters to be learned, especially when the number of tasks and/or data dimensionality is large. Several existing methods for multitask/transfer distance metric learning arise as special cases of our model. Preliminary results on real-world data show that our model outperforms various baselines. We also discuss some possible extensions of our model and future work.

1 Introduction

Computing distances between data points is a key step in many problems such as classification, clustering, and ranking. In many cases, the standard Euclidean distance is not appropriate and *problem-specific* distance functions are deemed more suitable. Distance metric learning [14, 1] algorithms are appealing in such cases as they allow learning data-driven distance metrics. Specifically, the distance between two data points x_i and x_j is defined as $d = \sqrt{(x_i - x_j)^\top A(x_i - x_j)}$ where A is a $D \times D$ positive semi-definite matrix denoting the distance metric. Distance metric learning algorithm try to learn the "right" distance metric A, given a set of constraints (pairwise similarities/dissimilarities, or relative preferences) provided as a form of supervision.

Often, we are interested in solving not just one but $T > 1$ learning tasks and wish to learn *multiple* distance metrics $A_1, \ldots, A_T$ (one per task). Since the tasks could possibly be related, it is desirable to *jointly* learn these distance metrics in order to share statistical strengths across the multiple learning tasks, especially when the amount of training data and/or the number of distance-based constraints known for each task is small. This has been the motivation behind some recent methods for *transfer/multitask* distance metric learning [10, 17, 16, 18]. However, the task-relatedness is usually unknown *a priori*. It is beneficial to learn the task-relatedness while jointly learning the distance metrics for the multiple tasks.

In this paper, we present a Bayesian approach to the multitask distance metric learning problem. Our proposed approach is appealing due to several reasons. Firstly, our approach discovers the task-relatedness and allows a proper sharing of statistical strength among the multiple learning tasks. Secondly, the Bayesian formulation naturally provides a full posterior distribution over the distance metrics [15], rather than a point estimate, which gives the solution more robustness against overfitting when the amount of training data is small.

Specifically, our proposed formulation expresses the distance metric of each task as a *sparse* weighted combination of a set of *basis* distance metrics (intuitively, the degree of similarity of two tasks would be proportional to the number of basis distance metrics they share). Note that this is akin to a sparse coding [6, 8] of each task-specific distance metric using the elements from a *distance metric dictionary*. The sparse code for each task as well as the distance metric dictionary will be

learned from data using a nonparametric Bayesian approach. Our model allows incorporating both strictly pairwise (similar or dissimilar) as well as relative preference (e.g., triplets) based constraints.

2 Bayesian Multitask Distance Metric Learning

We assume that we are given T tasks, with their corresponding distance metrics denoted as $A_1, \ldots, A_T$. We further assume that each $D \times D$ distance metric A_t can be written as a sparse, positively weighted combination of a small set of K basis metrics $M_1, \ldots, M_K$, plus an offset M_0 shared by all the tasks. Each M_k, $k = 0, \ldots, K$, is assumed to be a symmetric positive definite matrix, which is further assumed to be a low-rank matrix of the form $B_k B_k^\top$ where $B_k \in \mathbb{R}^{D \times L}$ with $L \leq D$. Specifically, the task t distance metric A_t is modeled as follows:

$$
\begin{aligned}
A_t &= \sum_{k=1}^{K} W_{tk} M_k + M_0 \\
M_k &= B_k B_k^\top \quad \text{where} \quad B_k \in \mathbb{R}^{D \times L} \\
B_{kl} &\sim \mathcal{N}or(0, \sigma_b^2 I_D) \\
W_{tk} &= Z_{tk} S_{tk} \\
Z_{tk} &\sim \mathcal{B}er(\pi_k), \quad \pi_k \sim \mathcal{B}et(\alpha/K) \\
S_{tk} &\sim \mathcal{HN}(0, \sigma_s^2)
\end{aligned}
$$

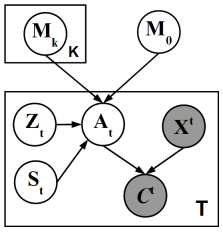

Figure 1: Graphical model in plate notation. Shaded nodes are observed.

In the above construction, Z_{tk} denotes whether basis k (given by M_k) is chosen by task t and $S_{tk} \in \mathbb{R}_+$ specifies the *weight*. Note that in the above construction, the Beta-Bernoulli prior distribution over the basis selection matrix Z assumes a finite K; for large enough K it approximates the Indian Buffet Process (IBP) [3]. Alternatively, Z can be drawn from the IBP, in which case K need not be set beforehand. Likewise, the number of columns L of each low-rank matrix B_k can be inferred nonparametrically using the multiplicative gamma process prior [2].

Our model requires estimating $Z \in \{0,1\}^{T \times K}$, $S \in \mathbb{R}_+^{T \times K}$, and $\{B_k\}_{k=1}^K$ with each $B_k \in \mathbb{R}^{D \times L}$. Therefore, the total number of parameters to be estimated is $\mathcal{O}(TK + KDL)$. In contrast, the multitask metric learning model in [10] which learns a separate distance metric for each task (plus a shared distance matrix) requires estimating $\mathcal{O}(TD^2)$ parameters which can be expensive when the number of tasks (T) and/or the number of features (D) per task is large.

3 Full Model and Inference

Figure 1 shows the full model. The training data for task t is given in form of N_t examples $X^t = [x_1^t, \ldots, x_{N_t}^t]$ and a set $\mathcal{C}^t = \{\mathcal{S}_t, \mathcal{D}_t, \mathcal{R}_t\}$ of constraints defined between examples where $\mathcal{S}_t$, $\mathcal{D}_t$, and $\mathcal{R}_t$ denote pairwise similarity, pairwise dissimilarity, and *relative* comparison based constraints (given as triplets of the form "example i is more similar to j than to k), respectively. The goal is to infer the distribution $P(\Theta \mid \{X^t, C^t\}_{t=1}^T)$ over the latent variables, given data from all the T tasks, where Θ collectively refers to the set of all the latent variables $\{Z, S, \{B_k\}_{k=0}^K\}$ that we need to infer (note that we do not need to explicitly maintain M_k but only B_k since $M_k = B_k B_k^\top$).

Pairwise Constraints: Each (similarity/dissimilarity based) pairwise constraint is modeled using a logistic function with a margin μ [15]:

$$
P(y_{ij}^t | x_i^t, x_j^t, \Theta, \mu) = \frac{1}{1 + \exp(y_{ij}^t (d_{A_t}^2(x_i^t, x_j^t) - \mu))}
$$

where $y_{ij}^t = +1$ if x_i^t, x_j^t are similar, and $y_{ij}^t = -1$ otherwise. Two points are likely to be assigned to the same class only when their distance is less than μ. Here $d_{A_t}^2(x_i^t, x_j^t) = (x_i^t - x_j^t)^\top A_t (x_i^t - x_j^t)$ is the squared distance between two examples x_i^t, x_j^t from task t, under the distance metric A_t.

Preference Based Constraints: For task t, the relative preference based constraints $\mathcal{R}^t$ are of the form (i, j, k) which means that data point x_i^t is more similar to x_j^t than to x_k^t. We impose the large

margin assumption [10] on the triplet constraints and require the following condition to be satisfied for each triplet (i, j, k) of task t: $d^2_{A_t}(x^t_i, x^t_k) \geq d^2_{A_t}(x^t_i, x^t_j) + 1$.

Following other metric learning methods [13, 10], we express the "loss" associated with triplet (i, j, k) of task t to be: $\max(1 - \{d^2_{A_t}(x^t_i, x^t_k) - d^2_{A_t}(x^t_i, x^t_j)\}, 0)$. and use the following pseudo-likelihood for each triplet based likelihood term:

$$\exp(-\max(1 - \{d^2_{A_t}(x^t_i, x^t_k) - d^2_{A_t}(x^t_i, x^t_j)\}, 0))$$

Exact inference in our model is intractable. We use MCMC to perform approximate inference in our model. We use Gibbs sampling [3] to sample each binary-valued entry of the basis selection matrix Z, and use elliptical slice-sampling [9] to sample for S and $\{B_k\}_{k=0}^K$. For brevity, we skip the details of the inference.

4 Special Cases

For specific choices of the basis selection matrix Z (i.e., when it is set to a *fixed* value) and the global shared distance metric M_0, our model leads to some spacial cases such as:

- The case when Z is an identity matrix of size $T \times T$ and $M_0 = 0$ is equivalent to learning independent distance metric for each task, i.e., $A_t = M_t$.
- The case when Z is an identity matrix of size $T \times T$ and $M_0 \neq 0$ is equivalent to the method proposed in [10] which assumes that the distance metric of each task is a sum of a global distance metric and a task-specific distance metric, i.e., $A_t = M_0 + M_t$.
- The case when Z is a matrix of all zeros, each distance metric $A_t = M_0$, which is equivalent to all the tasks sharing a single global distance metric M_0.

Our model, in addition to subsuming the above-mentioned cases, can flexibly model different relatedness between tasks by inferring metric basis $\{M_k\}$ and basis selection matrix Z.

Also note that our model can be used to learn *class-specific* distance metrics [13] in single-task learning. In this case, each task corresponds to a class and the data for each task only consists of examples from the corresponding class.

5 Possible Extensions and Future Work

Our model can also be extended to *zero-data* transfer learning [5] settings. For instance, in many problems, features-descriptors/covariates for tasks may be available [5] which, in our framework, can be leveraged to predict the basis combination weights $Z_t \odot S_t$, especially for a *new* task that may not have any labels/distance-based-constraints. To model the basis combination weights of such tasks, one possibility could to replace the Beta-Bernoulli/IBP prior on Z_t by a *feature-dependent* IBP prior such as the linear probit model [11]: Suppose, a new task t has a task-descriptor feature vector $f_t \in \mathbb{R}^P$ then we could model Z_t as $P(Z_{tk} = 1) = \Phi_{0,1}(\beta_k^\top f_t + \Phi^{-1}_{0,1}(a_k))$ where a_k is the *a priori* probability of basis k to be chosen, $\beta_k \in \mathbb{R}^P$ are regression weights on the task descriptors, and Φ represent the normal CDF. Note that the set of regression weights $\{\beta_k\}_{k=1}^K$ and basis usage probabilities $\{a_k\}_{k=1}^K$ would be learned from the previous tasks. This would allow predicting the basis combination weights for a new task, for which no supervision (in form of constraints) is available, solely based on its task-descriptor feature vector.

Another possible extension could be doing *active* distance metric learning [15] in our multitask distance metric learning setting, which is expected to further reduce the number of constraints needed for each task. Our Bayesian framework would naturally allow doing this.

Finally, scaling up the model to large-data problems is another avenue of future work, especially to handle the enormously large number of pairwise/triplet constraints in the training data which make the likelihood computations a bottleneck in efficient inference. In this direction, two approaches seem worth pursuing: (1) finding the most useful "support" pairs/triplets such that computing the likelihood using only those can provide an approximation to the likelihood on the entire set of constraints, e.g., using the idea of Firefly Monte Carlo [7]; and (2) instead of MCMC, using online variational inference methods such as stochastic variational inference [4], which would be a promising way to scale up our model for larger data sets.

6 Experiments

We report preliminary results of our model on a real-world data - **Isolet** [10]. The Isolet data set consists of 5 tasks, constructed from speech data from 150 speakers with 5 groups. Each task is a multiclass classification problem (classifying an utterance into one of 26 English alphabets). We experiment with the more challenging *label-incompatible* setting [10] where the number of labels in could be different across the different tasks (we construct the data such that some tasks had less than 26 alphabets in the utterances). The data originally had 617 features and PCA was applied as a preprocessing step (as done in [10]) to reduce the dimensionality to 169 using principal components that capture 95% variance. We use a subset of the data which consists of 360 training, 120 test, and 120 validation examples for each task. In our experiments, we only use triplet constraints; note however that, if provided, the proposed model is capable of using pairwise constraints. To generate the triplets, we follow the strategy used in [12]: for each training example, we choose 3 nearest neighbors from the same class and 10 nearest neighbors from different classes.

For these experiments, we use nearest neighbors classification to predict the labels for the test data. For this step, the number of nearest neighbors for each baseline as well as our method is chosen using cross-validation on the validation set. We compare our model BMDML (for Bayesian Multitask Distance Metric Learning) with the following baselines:

- Independent task learning with Euclidean distance (Ind-Euc).
- Global multitask distance metric learning (gMDML) which learns a single distance metric shared by all the tasks, i.e., $A_t = M_0$ for $t = 1, \ldots, T$.
- Multitask large-margin distance metric learning [10] which assumes each distance metric to be of the form $A_t = M_0 + M_t$.

In our experiments, for our model we set $K = 20$ and $L = 50$, which worked well for our experiments Alternatively, these values can be inferred from data using the Indian Buffet Process [3] and the multiplicative gamma process [2] prior on Z_t's and B_k's, respectively.

Table 1: Classification accuracies on Isolet data

	Ind-Euc	gMDML	mt-LMNN	BMDML
Task 1	90.83%	91.12%	92.56%	**93.74%**
Task 2	95.00%	94.48%	96.12%	**97.53%**
Task 3	90.83%	91.04%	93.02%	**93.95%**
Task 4	87.50%	87.11%	89.35%	**92.24%**
Task 5	93.33%	94.12%	94.92%	**96.72%**
Average	91.50%	91.37%	93.19%	**94.83%**

Table 1 shows the results of our model and the various baselines. Our results for each experiment are obtained by averaging the distance metrics over the posterior samples after burn-in. We report results in terms of the classification accuracies on each task as well as the average classification accuracy over all the tasks. As shown in Table 1, BMDML outperforms all the baselines which demonstrates the model's effectiveness in appropriately sharing the right amount of information across the multiple tasks. We also notice that the gMDML baseline sometimes gets outperformed by the simpler method Ind-Euc, which is probably a result of the small size of the training data and/or adverse effects due to pooling all the tasks' data to learn a single shared distance metric.

7 Conclusion

We have presented a model for learning distance metrics for multiple tasks by appropriately sharing information across the different tasks. Our models leads to a flexible way of jointly learning multiple distance metrics and, as discussed in Section 5, our model can be extended in several useful ways such as *zero-data* transfer learning to new tasks with no supervised information, active distance metric learning, and can be given a fully nonparametric Bayesian treatment. We are currently exploring these possibilities, as well as ways to scale up the model to larger data sets by employing alternative, more efficient inference methods.

Acknowledgements: This work is supported in part by AOR, DARPA, DOE, NGA, and ONR.

References

[1] A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[2] A. Bhattacharya and D. B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

[3] T. L. Griffiths and Z. Ghahramani. The indian buffet process: An introduction and review. *The Journal of Machine Learning Research*, 12:1185–1224, 2011.

[4] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[5] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[6] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[7] D. Maclaurin and R. P. Adams. Firefly monte carlo: Exact mcmc with subsets of data. *arXiv preprint arXiv:1403.5693*, 2014.

[8] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[9] I. Murray, R. P. Adams, and D. J. MacKay. Elliptical slice sampling. In *AISTATS*, 2010.

[10] S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. In *NIPS*, 2010.

[11] N. Quadrianto, V. Sharmanska, D. A. Knowles, and Z. Ghahramani. The supervised ibp: Neighbourhood preserving infinite latent feature models. In *UAI*, 2013.

[12] Y. Shi, A. Bellet, and F. Sha. Sparse compositional metric learning. In *AAAI*, 2014.

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

[14] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002.

[15] L. Yang, R. Jin, and R. Sukthankar. Bayesian active distance metric learning. *arXiv preprint arXiv:1206.5283*, 2012.

[16] P. Yang, K. Huang, and C.-L. Liu. Geometry preserving multi-task metric learning. *Machine learning*, 92(1):133–175, 2013.

[17] Y. Zhang and D.-Y. Yeung. Transfer metric learning by learning task relationships. In *KDD*, 2010.

[18] Y. Zhang and D.-Y. Yeung. Transfer metric learning with semi-supervised extension. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):54, 2012.