Nonparametric Bayesian Multi-task Learning with Max-margin Posterior Regularization

Jun Zhu
Dept. of Computer Science & Tech., TNList Lab, Tsinghua University, Beijing 100084, China
dcszj@tsinghua.edu.cn

Abstract

Learning a common latent representation can capture the relationships and share statistic strength among multiple tasks. To automatically resolve the unknown dimensionality of the latent representation, nonparametric Bayesian methods have been successfully developed with a generative process describing the observed data. In this paper, we present a discriminative approach to learning nonparametric Bayesian models under a computational framework called regularized Bayesian inference. In particular, we will discuss how to use the successful principle of max-margin learning to improve the prediction performance of nonparametric Bayesian multi-task models. We will discuss both variational approximation and Markov chain Monte Carlo methods to do posterior inference, with real-world experimental results demonstrating their efficacy.

1 Introduction

In real world applications, we usually deal with multiple tasks that are potentially related in various ways, such as sharing the same inputs or arising from a similar physical process, and the training signal in one task is potentially of great use for improving the learning of other tasks. A naive solution that treats the related tasks separately (known as single task learning, or STL) has the risk of missing important information. In order to better capture the dependency between multiple tasks and achieve inductive transfer between tasks, multi-task learning (or MTL) [5] has received a lot of attention and various approaches have been developed. Many different approaches have been developed for multi-task learning (See [7] for a review). In particular, learning a common latent representation shared by all the related tasks has proven to be an effective way to capture task relationships [1, 2, 10]. But existing work either needs to pre-specify the dimensionality of the latent space or has a weak predictive power due to a generative formulation.

In this paper, we present a nonparametric Bayesian method for multi-task learning, by conjoining the ideas of Bayesian nonparametrics to resolve the latent dimension in a data-driven manner and the ideas of large-margin learning to discover a predictive latent representations. Though it is not intuitively obvious how to achieve such a goal, we show that this can be done under regularized Bayesian inference (RegBayes) [16, 17], a new computational framework that incorporates rich domain knowledge or side information by imposing posterior regularization on the desired posterior distribution. We present two approaches to defining the large-margin posterior regularization terms with either an averaging classifier or a Gibbs classifier, with efficient inference algorithms.

2 Regularized Bayesian Inference

In this section, we overview the generic computational framework of regularized Bayesian inference with posterior constraints.

2.1 Variational Formulation of Bayes' Theorem

Let $\mathbb{M}$ denote a space of feasible models, and $\mathcal{M} \in \mathbb{M}$ represents an atom in this space. Given a collection of observed data $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$, which are often assumed to be i.i.d given the model, the Bayes' theorem establishes the following relationship among the prior $\pi(\mathcal{M})$, a likelihood model $p(\mathbf{x}|\mathcal{M})$, and the posterior distribution $p(\mathcal{M}|\mathcal{D})$: $p(\mathcal{M}|\mathcal{D}) = \frac{\pi(\mathcal{M})p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})}$, where $p(\mathcal{D})$ is the marginal likelihood or evidence of observed data. We can show that the posterior distribution due to the Bayes' theorem is the solution of the problem

$$\min_{q(\mathcal{M})\in\mathcal{P}_{\text{prob}}} \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) - \int \log p(\mathcal{D}|\mathcal{M})q(\mathcal{M})d\mathcal{M} \tag{1}$$

where $\mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M}))$ is the Kullback-Leibler (KL) divergence, and $\mathcal{P}_{\text{prob}}$ is the space of probability distributions with an appropriate dimension.

2.2 Regularized Bayesian Inference

In the above variational formulation of Bayes' theorem, the constraint on $q(\mathcal{M})$ is merely due to the law of conservation of belief, which does not capture any domain knowledge or structures of the model or data. But this optimization formulation makes it straightforward to generalize Bayesian inference to a richer type of posterior inference, by replacing the trivial normality constraint on q with a wide spectrum of knowledge-driven and/or data-driven constraints or regularization. Formally, we define *Regularized Bayesian Inference* (or RegBayes) as a generalized posterior inference procedure that solves a constrained optimization problem due to some additional regularizations imposed on q:

$$\begin{aligned} \min_{q(\mathcal{M}),\boldsymbol{\xi}} \quad & \mathcal{L}(q(\mathcal{M})) + U(\boldsymbol{\xi}) \\ \text{s.t.}: \quad & q(\mathcal{M}) \in \mathcal{P}_{\text{post}}(\boldsymbol{\xi}), \end{aligned} \tag{2}$$

where $\mathcal{P}_{\text{post}}(\boldsymbol{\xi})$ is a subspace of distributions that satisfy a set of constraints and $\mathcal{L}(q)$ is the objective of problem (1). The auxiliary parameters $\boldsymbol{\xi}$ are nonnegative slack variables. $U(\boldsymbol{\xi})$ is a convex function, which usually corresponds to a surrogate loss (e.g., hinge loss) of a prediction rule, as we shall see. By absorbing the slack variables, the problem can be equivalently rewritten as:

$$\min_{q(\mathcal{M})\in\mathcal{P}_{\text{prob}}} \quad \mathcal{L}(q(\mathcal{M})) + \Omega(q(\mathcal{M})), \tag{3}$$

where $\Omega(q(\mathcal{M}))$ is a regularization term on the posterior distribution q.

When the posterior regularization is defined with some linear operator (e.g., expectation), the problem is convex and convex analysis theory can be applied to solve it. We refer the readers to [17] for more details. Below, we present a concrete example of RegBayes for multi-task learning.

3 Multi-task Infinite Latent Support Vector Machines

In this section, we concretize the ideas of RegBayes by particularly focusing on developing a multi-task learning model consisting of multiple large-margin classifiers that share a common latent projection matrix. To resolve the dimensionality of the projection matrix, we conjoin the ideas of Bayesian nonparametrics and large-margin learning by using the well-studied Indian Buffet Process (IBP, aka Beta Process) prior on the binary projection matrix.

3.1 Indian Buffet Process

Indian buffet process was proposed in [6] and has been successfully applied in various fields, such as link prediction [9, 14] and multi-task learning [10]. We focus on its stick-breaking construction [11], which is good for developing efficient inference methods. Let $\pi_k \in (0, 1)$ be a parameter associated with column k of the binary matrix $\mathbf{Z}$. Given π_k, each z_{nk} in column k is sampled independently from $\mathrm{Bernoulli}(\pi_k)$. The parameters $\boldsymbol{\pi}$ are generated by a stick-breaking process

$$\pi_1 = \nu_1, \text{ and } \pi_k = \nu_k\pi_{k-1} = \prod_{i=1}^{k} \nu_i, \tag{4}$$

where $\nu_i \sim \mathrm{Beta}(\alpha, 1)$. This process results in a decreasing sequence of probabilities π_k. Specifically, given a finite dataset, the probability of seeing feature k decreases exponentially with k.

3.2 The Model with Averaging Classifiers

Suppose we have M related tasks. Let $\mathcal{D}_m = \{(\mathbf{x}_{mn}, y_{mn})\}_{n \in \mathcal{I}_{\text{tr}}^m}$ be the training data for task m. We consider binary classification tasks, where $\mathcal{Y}_m = \{+1, -1\}$. Extension to multi-way classification or regression tasks can be easily done. If the latent matrix $\mathbf{Z}$ is given, we define the latent discriminant function for task m as

$$f_m(\mathbf{x}, \mathbf{Z}; \boldsymbol{\eta}_m) \equiv (\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x} = \boldsymbol{\eta}_m^\top (\mathbf{Z}^\top \mathbf{x}). \tag{5}$$

This definition provides two views of how the M tasks get related. If we let $\varsigma_m = \mathbf{Z}\boldsymbol{\eta}_m$, then ς_m are the actual parameters of task m and all ς_m in different tasks are coupled by sharing the same latent matrix $\mathbf{Z}$. Another view is that each task m has its own parameters $\boldsymbol{\eta}_m$, but all the tasks share the same latent features $\mathbf{Z}^\top \mathbf{x}$, which is a projection of the input features $\mathbf{x}$ and $\mathbf{Z}$ is the latent projection matrix. As such, our method can be viewed as a nonparametric Bayesian treatment of alternating structure optimization (ASO) [1], which learns a single projection matrix with a pre-specified latent dimension. Moreover, different from [7], which learns a binary vector with known dimensionality to select features or kernels on $\mathbf{x}$, we learn an unbounded projection matrix $\mathbf{Z}$ using nonparametric Bayesian techniques.

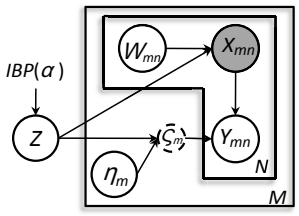


Figure 1: Graphical structure of multi-task infinite latent SVM.

We take the fully Bayeisan treatment (i.e., $\boldsymbol{\eta}_m$ are also random variables) and define the *effective discriminant function* for task m as the expectation

$$f_m(\mathbf{x}; p(\mathbf{Z}, \boldsymbol{\eta})) \equiv \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[f_m(\mathbf{x}, \mathbf{Z}; \boldsymbol{\eta}_m)] = \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}. \tag{6}$$

Then, the prediction rule for task m is naturally $y_m^* \equiv \text{sign} f_m(\mathbf{x})$. With these definitions, we do regularized Bayesian inference by defining $U(\boldsymbol{\xi}) \equiv C \sum_{m,n \in \mathcal{I}_{\text{tr}}^m} \xi_{mn}$ and

$$\mathcal{P}_{\text{post}}(\boldsymbol{\xi}) \equiv \left\{ p(\mathbf{Z}, \boldsymbol{\eta}) \;\middle|\; \forall m,\ \forall n \in \mathcal{I}_{\text{tr}}^m : \quad \begin{array}{l} y_{mn} \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{mn} \geq 1 - \xi_{mn} \\ \xi_{mn} \geq 0 \end{array} \right\}. \tag{7}$$

Minimizing $U(\boldsymbol{\xi})$ with the above constraints is equivalent to minimizing the hinge-loss $\mathcal{R}_h$ of the multiple binary prediction rules, where

$$\mathcal{R}_h = C \sum_{m,n \in \mathcal{I}_{\text{tr}}^m} \max(0, 1 - y_{mn} \mathbb{E}_{p(\mathbf{Z},\boldsymbol{\eta})}[\mathbf{Z}\boldsymbol{\eta}_m]^\top \mathbf{x}_{mn}). \tag{8}$$

Finally, we may also be interested in modeling the input data. This can be done by defining a likelihood model

$$p(\mathbf{x}_{mn} | \mathbf{w}_{mn}, \mathbf{Z}, \lambda_{mn}^2) = \mathcal{N}(\mathbf{x}_{mn} | \mathbf{Z}\mathbf{w}_{mn}, \lambda_{mn}^2 I), \tag{9}$$

where $\mathbf{w}_{mn}$ is a vector. We assume $\mathbf{W}$ has an independent prior $\pi(\mathbf{W}) = \prod_{mn} \mathcal{N}(\mathbf{w}_{mn} | 0, \sigma_{m0}^2 I)$. Fig. 1 (b) illustrates the graphical structure of MT-iLSVM. For testing, we do Bayesian inference on both training and test data. The difference is that training data are subject to large-margin constraints, while test data are not. The hyper-parameters σ_{m0}^2 and λ_{mn}^2 can be estimated from data.

3.3 The Model with Gibbs Classifiers

An alternative way to define a multi-task iLSVM model is to follow the ideas of Gibbs classifiers [8]. Specifically, if the latent matrix $\mathbf{Z}$ is given, we define the same latent discriminant function $f_m(\mathbf{x}, \mathbf{Z}; \boldsymbol{\eta}_m)$ for task m as above. Then, instead of making prediction with an averaging classifier, a Gibbs classifier uses the latent discriminant function to make prediction directly, that is, $y_m^*(\mathbf{Z}, \boldsymbol{\eta}_m) = \text{sign} f_m(\mathbf{x}, \mathbf{Z}; \boldsymbol{\eta}_m)$; and minimizes the expected hinge loss of the multiple latent predictive rules:

$$\mathcal{R}'(q(\mathbf{Z}, \boldsymbol{\eta}) = C \sum_{m,n \in \mathcal{I}_{\text{tr}}^m} \mathbb{E}_{q(\mathbf{Z},\boldsymbol{\eta})}[\max(0, 1 - y_{mn}(\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x}_{mn})]. \tag{10}$$

With these definitions, we do regularized Bayesian inference in the form of (3) with the posterior regularization $\mathcal{R}'(q(\mathbf{Z}, \boldsymbol{\eta}))$. Comparing to the averaging model, we can show that $\mathcal{R}'$ is an upper bound of $\mathcal{R}$.

3.4 Inference Algorithms

For the MT-iLSVM model with averaging classifiers, we can develop an approximate algorithm by imposing additional truncated mean-field constraints on q with the stikck-breaking representation of IBP, which includes the auxiliary variables $\boldsymbol{\nu}$, and infer the posterior $p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta})$. Formally, we impose the truncated mean-field constraint that

$$p(\boldsymbol{\nu}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\eta}) = p(\boldsymbol{\eta}) \prod_{k=1}^{K} \Big(p(\nu_k|\boldsymbol{\gamma}_k) \prod_{d=1}^{D} p(z_{dk}|\psi_{dk}) \Big) \prod_{mn} p(\mathbf{w}_{mn}|\Phi_{mn}, \sigma_{mn}^2 I), \tag{11}$$

where K is the truncation level; $p(\mathbf{w}_{mn}|\Phi_{mn}, \sigma_{mn}^2 I) = \mathcal{N}(\mathbf{w}_{mn}|\Phi_{mn}, \sigma_{mn}^2 I)$; $p(z_{dk}|\psi_{dk}) = \text{Bernoulli}(\psi_{dk})$; and $p(\nu_k|\boldsymbol{\gamma}_k) = \text{Beta}(\gamma_{k1}, \gamma_{k2})$. Then, we can derive an iterative procedure to solve the MT-iLSVM problem in a constrained form (2). We refer the readers to [17] for details.

For the MT-iLSVM model with Gibbs classifiers, although it is not obvious how to derive a variational approximation algorithm, we can develop a Markov chain Monte Carlo (MCMC) algorithm by introducing some augmented variables. Namely, let $\omega_{mn} = 1 - y_{mn}(\mathbf{Z}\boldsymbol{\eta}_m)^\top \mathbf{x}_{mn}$ and $\psi(y_{mn}|\mathbf{x}_{mn}, \mathbf{Z}, \boldsymbol{\eta}_m) = \exp(-2C \max(0, 1 - \omega_{mn}))$ be an un-normalized pseudo-likelihood. Then, the optimization problem is to solve

$$\min_{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) \in \mathcal{P}_{\text{prob}}} \text{KL}(q(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) \| p_0(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W})) - \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) + \log \psi(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta})], \tag{12}$$

where $\psi(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}) = \prod_{m,n} \psi(y_{mn}|\mathbf{x}_{mn}, \mathbf{Z}, \boldsymbol{\eta}_m)$. It can be shown that the optimal solution is $q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) \propto p_0(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) \psi(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\eta})$, with the normalization constant safely ignored. The last step to derive a simple MCMC sampler is to introduce some augmented variables. For the pseudo-likelihood, we can show that

$$\psi(y_{mn}|\mathbf{x}_{mn}, \mathbf{Z}, \boldsymbol{\eta}_m) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{mn}}} \exp\Big(- \frac{(\lambda_{mn} + C\omega_{mn})^2}{2\lambda_{mn}} \Big) d\lambda_{mn}. \tag{13}$$

Therefore, we can infer $q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta})$ through drawing examples from the augmented posterior $q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}, \boldsymbol{\lambda}) \propto p_0(\mathbf{Z}, \boldsymbol{\eta}, \mathbf{W}) p(\mathbf{X}|\mathbf{Z}, \mathbf{W}) \prod_{m,n} \frac{1}{\sqrt{2\pi\lambda_{mn}}} \exp(-\frac{(\lambda_{mn} + C\omega_{mn})^2}{2\lambda_{mn}})$ with a Gibbs sampler of which each conditional distribution has an analytical form. We omit the details to save space. A similar example can be found in [15].

4 Experiments

Now, we present some empirical results to demonstrate the efficacy of MT-iLSVM.

4.1 Description of the Data

Scene and Yeast Data: These datasets are from the UCI repository, and each data example has multiple labels. As in [10], we treat the multi-label classification as a multi-task learning problem, where each label assignment is treated as a binary classification task. The Yeast dataset consists of 1500 training and 917 test examples, each having 103 features, and the number of labels (or tasks) per example is 14. The Scene dataset consists 1211 training and 1196 test examples, each having 294 features, and the number of labels (or tasks) per example for this dataset is 6.

School Data: This dataset comes from the Inner London Education Authority and has been used to study the effectiveness of schools. It consists of examination records from 139 secondary schools in years 1985, 1986 and 1987. The dataset is publicly available and has been extensively evaluated in various multi-task learning methods [3, 4, 13], where each task is defined as predicting the exam scores of students belonging to a specific school based on four student-dependent features (year of the exam, gender, VR band and ethnic group) and four school-dependent features (percentage of students eligible for free school meals, percentage of students in VR band 1, school gender and school denomination). In order to compare with the above methods, we follow the same setup and similarly we create dummy variables for those features that are categorical forming a total of 19 student-dependent features and 8 school-dependent features. We use the same 10 random splits[1] of the data, so that 75% of the examples from each school (task) belong to the training set and 25% to the test set. On average, the training set includes about 80 students per school and the test set about 30 students per school.

[1] Available at: http://ttic.uchicago.edu/∼argyriou/code/index.html

Table 1: Multi-label classification performance on Scene and Yeast datasets.

Model	Yeast			Scene		
	Acc	F1-Micro	F1-Macro	Acc	F1-Micro	F1-Macro
yaxue [10]	0.5106	0.3897	0.4022	0.7765	0.2669	0.2816
piyushrai-1 [10]	0.5212	0.3631	0.3901	0.7756	0.3153	0.3242
piyushrai-2 [10]	0.5424	0.3946	0.4112	0.7911	0.3214	0.3226
MT-IBP+SVM	0.5475 ± 0.005	0.3910 ± 0.006	0.4345 ± 0.007	0.8590 ± 0.002	0.4880 ± 0.012	0.5147 ± 0.018
MT-iLSVM	0.5792 ± 0.003	0.4258 ± 0.005	0.4742 ± 0.008	0.8752 ± 0.004	0.5834 ± 0.026	0.6148 ± 0.020
Gibbs MT-iLSVM	0.5851 ± 0.005	0.4294 ± 0.005	0.4763 ± 0.006	0.8855 ± 0.004	0.6494 ± 0.011	0.6458 ± 0.011

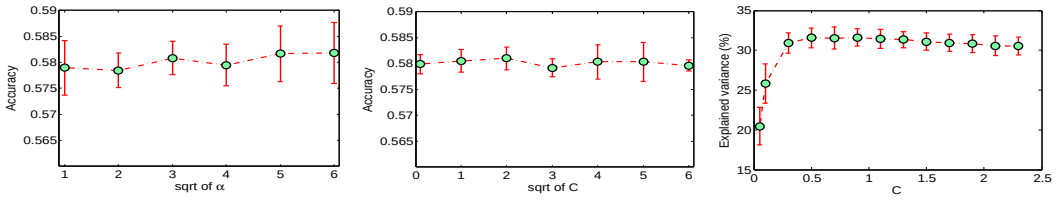
Table 2: Percentage of explained variance by various models on the School dataset.

STL	BMTL	MTGP	MTRL	MT-IBP+SVM	MT-iLSVM	MT-IBP+SVM^f	MT-iLSVM^f
23.5 ± 1.9	29.5 ± 0.4	29.2 ± 1.6	29.9 ± 1.8	20.0 ± 2.9	30.9 ± 1.2	28.5 ± 1.6	**31.7** ± 1.1

4.2 Results

Scene and Yeast Data: We compare with the closely related nonparametric Bayesian methods [10, 12], which were shown to outperform the independent Bayesian logistic regression and a single-task pooling approach [10], and a decoupled method *MT-IBP+SVM*[2] that uses IBP factor analysis model to find shared latent features among multiple tasks and then builds separate SVM classifiers for different tasks. For MT-iLSVM and MT-IBP+SVM, we use the mean-field inference method in Sec 3.4 and report the average performance with 5 randomly initialized runs. We also report some results of the MT-iLSVM model using Gibbs classifiers. For comparison with [10, 12], we use the overall classification accuracy, F1-Macro and F1-Micro as performance measures. Table 1 shows the results. We can see that the large-margin MT-iLSVM performs much better than other nonparametric Bayesian methods and MT-IBP+SVM, which separates the inference of latent features from learning the classifiers; and Gibbs MT-iLSVM performs better than MT-iLSVM mainly due to its MCMC algorithms that do not make strict mean-field assumptions.

School Data: We use the percentage of explained variance [3] as the measure of the regression performance, which is defined as the total variance of the data minus the sum-squared error on the test set as a percentage of the total variance. Since we use the same settings, we can compare with the state-of-the-art results of Bayesian multi-task learning (BMTL) [3], multi-task Gaussian processes (MTGP) [4], convex multi-task relationship learning (MTRL) [13], and single-task learning (STL) as reported in [4, 13]. For MT-iLSVM and MT-IBP+SVM, we also report the results achieved by using both the latent features (i.e., $\mathbf{Z}^\top \mathbf{x}$) and the original input features $\mathbf{x}$ through vector concatenation, and we denote the corresponding methods by *MT-iLSVM*^{f} and *MT-IBP+SVM*^{f}, respectively. From the results in Table 2, we can see that the multi-task latent SVM (i.e., MT-iLSVM) achieves better results than the existing methods that have been tested in previous studies. Again, the joint MT-iLSVM performs much better than the decoupled method MT-IBP+SVM, which separates the latent feature inference from the training of large-margin classifiers. Finally, using both latent features and the original input features can boost the performance slightly for MT-iLSVM, while much more significantly for the decoupled MT-IBP+SVM.



(a) Yeast (b) Yeast (c) School

Figure 2: Sensitivity study of MT-iLSVM: (a) classification accuracy with different α; (b) classification accuracy with different C; and (c) percentage of explained variance with different C.

[2]This decoupled approach is in fact an one-iteration MT-iLSVM, where we first infer the shared latent matrix $\mathbf{Z}$ and then learn an SVM classifier for each task.

4.3 Sensitivity Analysis

Figure 2 shows how the performance of MT-iLSVM changes against the hyper-parameter α and regularization constant C on Yeast and School datasets. We can see that on the Yeast dataset, MT-iLSVM is insensitive to α and C. For the School dataset, MT-iLSVM is stable when C is set between 0.3 and 1. MT-iLSVM is insensitive to α on the School data too, which is omitted to save space.

5 Conclusions

Learning a shared representation is an effective method to obtain inductive transfer between multiple related tasks. This paper presents a regularized Bayesian model by conjoining the ideas of Bayesian nonparametrics to resolve the latent dimension and large-margin learning to discover predictive latent representations. The resulting problems can be solved with truncated mean-field or Monte Carlo methods.

Acknowledgments

This work is supported by National Key Foundation R&D Projects (No.s 2013CB329403, 2012CB316301), Tsinghua Initiative Scientific Research Program No.20121088071, and the 221 Basic Research Plan for Young Faculties at Tsinghua University.

References

[1] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, (6):1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. In *NIPS*, 2007.

[3] B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *JMLR*, (4):83–99, 2003.

[4] E. Bonilla, K.M.A. Chai, and C. Williams. Multi-task Gaussian process prediction. In *NIPS*, 2008.

[5] R. Caruana. Multi-task learning. Technical report, TR: CMU-CS-97-203, Carnegie Mellon University, 1997.

[6] T.L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2006.

[7] T. Jebara. Multitask sparsity via maximum entropy discrimination. *JMLR*, (12):75–110, 2011.

[8] D. McAllester. Pac-bayesian model averaging. In *COLT*, 1999.

[9] K. Miller, T. Griffiths, and M. Jordan. Nonparametric latent feature models for link prediction. In *NIPS*, 2009.

[10] P. Rai and H. Daume III. Infinite predictor subspace models for multitask learning. In *AISTATS*, 2010.

[11] Y.W. Teh, D. Gorur, and Z. Ghahramani. Stick-breaking construction of the Indian buffet process. In *AISTATS*, 2007.

[12] Y. Xue, D. Dunson, and L. Carin. The matrix stick-breaking process for flexible multi-task learning. In *ICML*, 2007.

[13] Y. Zhang and D.Y. Yeung. A convex formulation for learning task relationships in multi-task learning. In *UAI*, 2010.

[14] J. Zhu. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

[15] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *ICML*, 2013.

[16] J. Zhu, N. Chen, and E.P. Xing. Infinite latent SVM for classification and multi-task learning. In *NIPS*, 2011.

[17] J. Zhu, N. Chen, and E.P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *arXiv:1210.1766v2*, 2013.