DISCRIMINATIVE INFINITE LATENT FEATURE MODELS

Minjie Xu, Jun Zhu

Department of Computer Science and Technology
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory of Information Science and Technology
Tsinghua University, Beijing, 100085, China

ABSTRACT

Latent feature models (LFMs) have been widely used to model ordinal rating data and relational network data in various tasks such as collaborative filtering and link prediction, typically in a generative way. Alternatively, one might incorporate max-margin learning into the model via the principle of Maximum Entropy Discrimination (MED) to learn a more discriminative latent feature space that favors the supervised learning task. Another dimension to extend LFMs is to employ Bayesian nonparametric methods to make LFMs self-adaptive to the number of latent features, which is crucial for model complexity control. In this paper we review several recent progresses that have been made in the above two extensions for the task of collaborative filtering and link prediction.

***Index Terms*—** max-margin, latent feature, Bayesian nonparametrics

1. INTRODUCTION

Latent feature models posit each entity is characterized by a latent feature vector which can be either correlated with some observed features or, especially when no observed features are available at all, totally inferred from data. Such data normally comprise relations between entities, e.g. whether a person follows or is friend of another in a social network as in link prediction where entities are people in the network, or in the case of collaborative filtering where entities are users and items, how many stars a user rates an item (e.g. book, movie, hotel, etc.) on a review and rating website.

By introducing latent features into the model, we gain more flexibility than sticking with the original ones since it becomes easier for us to inject certain prior knowledge (e.g. structure, sparsity, etc.) so that the learned latent representation would be more favorable to the task in hand.

Actually the concept of "latent features" is fairly broad and borrowed in many cases and problems. To narrow down our discussion, we in this paper only focus on a specific kind of latent feature model, where latent features only take binary values 0 or 1, and for each entity i or entity pair (i, j), its corresponding latent feature vector(s) $\mathbf{f}_i$ and $\mathbf{f}_j$ only appear in the model as a linear or bilinear form, namely $\boldsymbol{\alpha}^\top \mathbf{f}_i$ or $\mathbf{f}_i^\top W \mathbf{f}_j$, where $\boldsymbol{\alpha}$ is a real-valued coefficient vector and W is a real weight matrix. We can represent such latent features by a binary matrix Z, where the ith row corresponds to the latent feature vector of entity i, namely $\mathbf{f}_i^\top$, and each column corresponds to one specific latent feature. Then collectively, we have the following matrix notation

$$Z\boldsymbol{\alpha} \text{ or } ZV \text{ or } ZWZ^\top, \tag{1}$$

where V is an entity-specific real coefficient matrix.

Simple as they are, latent features models of such kind turn out to be useful in a large family of relational data problems [1, 2, 3]. Take link prediction as an example. Suppose entity i and j are two people in a social network, and we want to predict whether i is a friend of j based on many other already observed friendship connections in the network. Then $\mathbf{f}$ naturally arises as a vector recording the presence of such binary features as "is male", "graduates from Tsinghua University", "is beatlemania", "likes cycling" and so on. $\boldsymbol{\alpha}$ or V or W then represents the weights that will individually contribute to the prediction when their corresponding binary features are on.

Although typically formalized as generative models, LFMs can also be built as discriminative ones to better deal with supervised learning tasks. Besides, through borrowing Bayesian nonparametrics into the model, we are provided with a far more flexible and efficient way of deciding the number of latent features that is normally set *a priori* via computationally expensive model selection procedures such as cross-validation. Such two extensions bring about the discriminant infinite latent feature models (DILFMs) that we are going to introduce in this paper. As an example, we focus on the infinite probabilistic max-margin matrix factorization (iPM^{3}F) [4] for collaborative filtering and briefly covers another max-margin infinite latent feature relational model (MedLFRM) [5] for link prediction.

The rest of the paper is structured as follows. In Section 2, we briefly review two fundamental techniques that serve the basis of our extensions; In Section 3, we introduce two DILFM examples and discuss their learning algorithms; In Sec-

tion 4, we give empirical results that verify the advantages of the models; And finally, we conclude in Section 5.

2. BACKGROUNDS

We introduce MED [6] and the Indian buffet process (IBP) [7], which are two key elements underlying DIFLMs.

2.1. Maximum entropy discrimination

We consider binary classification since it suffices for our discussion. Given a set of training data $\mathcal{D} = \{(\mathbf{x}_d, y_d)\}_{d=1}^{D}$ ($y_d \in \{\pm 1\}$) and a discriminant function $F(\mathbf{x}; \boldsymbol{\eta})$ parameterized by $\boldsymbol{\eta}$, maximum entropy discrimination (MED) [6] seeks to learn a distribution $q(\boldsymbol{\eta})$ rather than a point estimate as is the case with standard SVMs that typically lack a probabilistic interpretation. Accordingly, MED takes expectation over $F(\mathbf{x}; \boldsymbol{\eta})$ with respect to $q(\boldsymbol{\eta})$ and has the following prediction rule

$$\hat{y} = \operatorname{sign}\left(\mathbb{E}_q[F(\mathbf{x}; \boldsymbol{\eta})]\right). \tag{2}$$

To find the target $q(\boldsymbol{\eta})$, MED solves a regularized risk minimization problem

$$\min_{q(\boldsymbol{\eta})} \ \mathrm{KL}\left(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})\right) + C \sum_d h_\ell\left(y_d \mathbb{E}_q[F(\mathbf{x}_d; \boldsymbol{\eta})]\right), \tag{3}$$

where $p_0(\boldsymbol{\eta})$ is the pre-specified prior distribution of $\boldsymbol{\eta}$; $\mathrm{KL}(q\|p_0)$ is the Kullback-Leibler divergence, or relative entropy, between two distributions; C is the regularization constant and $h_\ell(x) = \max(0, \ell - x)$ ($\ell > 0$) is the generalized hinge loss with margin parameter ℓ.

By defining the discriminant function F as the log-likelihood ratio of a Bayesian generative model, MED provides an elegant way to integrate discriminative max-margin learning and Bayesian generative modeling. Alternatively, F can be directly specified as any normal discriminant function without reference to probabilistic models [6], which makes MED far more applicable and flexible and in fact, MED naturally bestows support vector machines (SVMs) a probabilistic interpretation when $F(\mathbf{x}; \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \mathbf{x}$ and $p_0(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\mathbf{0}, I)$.

There are several interesting observations to the above MED formulation (3). In a broader sense, it is closely related in spirit to the Bayes' theorem

$$p(\boldsymbol{\eta}|\mathcal{D}) \propto p_0(\boldsymbol{\eta}) p(\mathcal{D}|\boldsymbol{\eta}) \tag{4}$$

that says the posterior distribution p is guided by the prior p_0 and then updated by the likelihood $p(\mathcal{D}|\boldsymbol{\eta})$ after data are observed; While in MED, the KL-divergence term ensures that the target distribution $q(\boldsymbol{\eta})$ stays not too far away from the prior and the empirical risk term, on the other hand, drives $q(\boldsymbol{\eta})$ towards one that more accurately captures the intrinsic rule explaining, not necessarily generating, the data. If we consider the Bayes' theorem as the golden rule underlying generative models, MED might be the counterpart for discriminative models since it directly solves for the posterior[1]. Meanwhile, the explicit representation of the empirical loss makes MED even more suitable for supervised discriminative learning, especially for binary or discrete ordinal data for which hinge loss is an appropriate choice.

Owing to these nice properties, MED has been widely used to build discriminative probabilistic models. It has also been extended to incorporate latent variables [8, 9] and perform structured output prediction [10].

2.2. Indian buffet process

One key element influencing the performance of latent feature models is the number of latent features to use, or equivalently in our case, the number of columns in Z. Typically a larger number indicates more parameters and hence more time to explore the solution space during learning while a smaller number puts model complexity at risk and normally gives unsatisfactory results. A typical solution relies on some general model selection procedure, e.g., cross-validation, which enumerates and compares many candidate models with different number of features and thus can be computationally expensive. The Indian buffet process (IBP) [7] is proposed in consequence to allow for probabilistic inference in LFMs with an unbounded number of latent features, the exact number of which is to be determined only *a posteriori*.

Specifically, IBP defines a stochastic process that generates sparse binary matrices of an unbounded number of columns. Think of binary matrix Z as recoding row-wisely customers' behavior of sampling dishes from an infinite long buffet, each dish corresponding to one column. Then IBP works as follows,

1. The first customer samples the first $\mathrm{Poisson}(\alpha)$ number of dishes;
2. The ith customer first samples dishes that have already been taken by previous customers, according to the dishes' popularity m_k/i where m_k is the number of previous customers who have sampled that dish; Then he tries a $\mathrm{Poisson}(\alpha/i)$ number of new dishes.

The process above induces a distribution for *lof*-equivalent class of binary matrices [7]. Matrices are considered equivalent if they are identical under some permutation of columns, which is desirable since we are not interested in distinguishing between different latent features. Together with Eq. (1), we find that all the all-zero columns in Z are hence ignorable and we may concentrate on a more compact Z with only a finite number of *active* features K_+, which follows a $\mathrm{Poisson}(\alpha H_N)$ where N is the number of entities (customers) and H_N is the Nth harmonic number.

IBP has another equivalent augmented stick-breaking construction [11]. Specifically, let $\pi_k \in (0, 1)$ be a parameter

[1] Hence we use $q(\boldsymbol{\eta})$ rather than $p(\boldsymbol{\eta}|\mathcal{D})$ in the formulation to distinguish it from the posterior as induced by the Bayes' rule.

associated with each column of Z. Then IBP can be described as given by the following generative process

$$\begin{aligned} &Z_{ik} \sim \text{Bernoulli}(\pi_k), \text{ i.i.d. for } i = 1, \ldots, N \ (\forall k), \\ &\pi_1 = \nu_1, \pi_k = \nu_k \pi_{k-1} = \prod_{l=1}^{k} \nu_l, \text{where} \\ &\nu_l \sim \text{Beta}(\alpha, 1), \text{ i.i.d. for } l = 1, \ldots, +\infty. \end{aligned} \tag{5}$$

Note that when $\boldsymbol{\nu}$ is integrated out, the marginal distribution of Z, with respect to the equivalent class, is identical to that induced from the above stochastic process.

Since its introduction, IBP has been widely used as the prior in lots of infinite LFMs [7, 1, 2, 3], most of which are generative and solved via MCMC sampling. Below, we review two representative discriminative infinite LFMs for relational data and their variational inference solution.

3. DISCRIMINATIVE INFINITE LFMS

3.1. iPM^{3}F for collaborative filtering

Collaborative filtering is a task of predicting users' potential preferences on currently unrated items (e.g., movies) based on their observed preferences and their relations with others'. One typical setting formalizes it as a matrix completion problem, i.e., to fill in missing entries into a partially observed user-by-item preference matrix $Y \in \mathbb{R}^{N \times M}$, where N and M are respectively the number of users and items. We denote the observed entry indices by $\mathcal{I}$.

Among other popular approaches, matrix factorization models a user's rating of an item as the linear combination of their latent factors (or features), hence the factorization $Y \simeq UV^\top$ with latent feature matrices $U \in \mathbb{R}^{N \times K}$ for users and $V \in \mathbb{R}^{M \times K}$ for items with K latent features. Various methods have been successfully developed to implement such an idea, including probabilistic matrix factorization (PMF) [12, 13] and deterministic reconstruction error minimization, e.g., max-margin matrix factorization (M^{3}F) [14, 15]. Note that in this paper, we don't consider side information (observed features) and hence data are only comprised of entity index pairs (i, j). For models where side information are incorporated, please refer to [16, 17].

PMF builds a probabilistic generative model for U and V with priors being zero-mean spherical Gaussian and likelihood induced from Gaussian observation noise. The resulting MAP estimation of PMF is thus equivalent to a regularized risk minimization problem with Frobenious norm regularizer and squared loss. M^{3}F solves an alternative minimization problem with the same regularizer but hinge loss[2]

$$\min_{U,V} \ \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + C \sum_{ij \in \mathcal{I}} h\left(Y_{ij} U_i V_j^\top\right), \tag{6}$$

[2]Due to space limit, we only discuss the binary case where $Y_{ij} \in \{\pm 1\}$. Please find more details in [15, 4].

where $h(x) = \max(0, 1 - x)$ is the hinge loss function and we use U_i to denote the ith row of U, and V_j likewise.

A major problem in (6), which is also common to many other matrix factorization methods and LFMs is how to determine an appropriate value of K. For PMF models, [1] proposed to use the IBP prior on U and developed an nonparametric Bayesian matrix factorization model where K is inferred from data. For the deterministic M^{3}F however, we first need to extend it to a probabilistic model, since only after that can IBP be introduced likewise.

MED happens to fit in for such an extension. More specifically, we take $\boldsymbol{\eta} = (U, V)$ and accordingly for the discriminant function,

$$F((i, j); U, V) = U_i V_j^\top \tag{7}$$

Substituting Eq. (7) into (3), we have the following discriminative probabilistic M^{3}F problem

$$\min_{q(U,V)} \ \text{KL}(q(U, V) \| p_0(U, V)) + C \sum_{ij \in \mathcal{I}} h_\ell\left(Y_{ij} \mathbb{E}_q[U_i V_j^\top]\right). \tag{8}$$

Under a proper choice of the prior p_0 and a rather mild mean-field assumption $p(U, V) = p(U)p(V)$, one can prove that problem (8) naturally reduces to M^{3}F (6).

Then the nonparametric extension is only steps away since one can again adopt the IBP prior for U and any appropriate prior for V, e.g. $p_0(V) = \prod_{j=1}^{M} \mathcal{N}(V_j | \mathbf{0}, \sigma^2 I)$. Note that although U and V are assumed to have an infinite number of columns, they only have a finite K^+ number of columns (features) that are actually active.

3.2. MedLFRM for link prediction

Link prediction is a fundamental problem in analyzing social network or relational data, and its goal is to predict unseen links between entities given the observed links. It is similarly formulated as predicting missing entries in an $N \times N$ partially observed relational link matrix Y, where N is the number of entities and $Y_{ij} = \pm 1$ indicates the presence or absence of a link between entity i and j. Sometimes we have observed attributes $X_{ij} \in \mathbb{R}^D$ that affect the link between i and j.

Various approaches based on probabilistic models have been developed, one class of which utilizes a latent feature matrix and a link function (e.g., the commonly used sigmoid function) [3] to define the link formation probability distribution. In contrary to these generative models that require a normalized link likelihood, MedLFRM [5] proposed to directly minimize the hinge loss that measures the quality of link prediction. The model, which also stems from MED, takes $\boldsymbol{\eta} = (Z, W, \boldsymbol{\alpha})$ and defines the discriminant function as

$$F((i, j), X_{ij}; Z, W, \boldsymbol{\alpha}) = Z_i W Z_j^\top + \boldsymbol{\alpha}^\top X_{ij} \tag{9}$$

where Z is the binary latent feature matrix, and W and $\boldsymbol{\alpha}$ are the latent weights. As for the priors $p_0(Z, W, \boldsymbol{\alpha})$, a natural

Table 1. NMAE comparison of matrix factorization methods.

Algorithm	MovieLens	EachMovie
M^{3}F [15]	$.4156 \pm .0037$	$.4397 \pm .0006$
PMF [12]	$.4332 \pm .0033$	$.4466 \pm .0016$
BPMF [13]	$.4235 \pm .0023$	$.4352 \pm .0014$
iPM^{3}F	$\mathbf{.4031} \pm .0030$	$\mathbf{.4211} \pm .0019$

Table 2. AUC comparison of LFRM models.

Algorithm	NIPS coauthorship
LFRM [3]	$.9509 \pm ./$
MedLFRM	$\mathbf{.9642} \pm .0026$

way to incorporate Bayesian nonparametrics is to assume independent priors and use the IBP prior for Z and standard normal priors for W and $\boldsymbol{\alpha}$.

3.3. Learning and inference

Due to space limit, we only briefly address several common problems when performing learning and inference in the above two DILFMs. For details please refer to the corresponding papers [5, 4].

Firstly, exact solution is intractable because the latent variables, e.g. U and V in iPM^{3}F and Z and W in MedLFRM, appear coupled in the discriminant function. Therefore we have to resort to some approximate learning algorithms. Specifically, we find the truncated mean-field variational algorithm to be both applicable and efficient once we employed the augmented stick-breaking construction of IBP as shown in Eq. (5). As a result, we alternatively solve for $q(\boldsymbol{\eta}, \boldsymbol{\nu})$.

As for the truncated mean-field assumptions, we assume that the posterior factorizes into component-wise independent ones and we set a finite truncation level K so that Z with more than K non-zero columns are directly rejected. Hence the truncation level limits the computational costs to a finite amount with acceptable approximation.

The variational algorithm then goes as usual for parametric models as we alternate between the components, solving a conditional subproblem each time. Thanks to the linear or bilinear forms in the discriminant function (7) & (9) and the linear expectation operator in the loss term (3), the subproblems are all relatively easy to solve.

4. EXPERIMENTS AND DISCUSSIONS

We now demonstrate the benefits of DILFMs over normal LFMs through empirical studies. iPM^{3}F is tested against PMF models and the deterministic parametric M^{3}F on two popular movie rating data sets, MovieLens 1M and EachMovie. While MedLFRM is tested against the generative infinite LFRM on the NIPS coauthorship data.

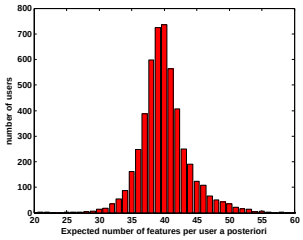

Fig. 1. $\sum_k \mathbb{E}_q[Z_{ik}]$ *a posteriori* for iPM^{3}F on MovieLens (set to 3 *a priori*)

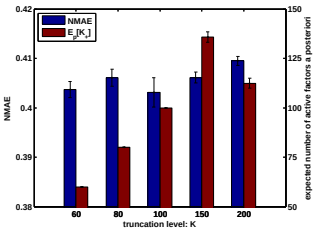

Fig. 2. Influence of K on the performance of iPM^{3}F on MovieLens

Table 1 and 2 show that both models achieve higher prediction accuracy as compared to their generative counterparts thanks to discriminative max-margin learning.

Fig. 1 demonstrates the models' feature adaptation capability inherited from Bayesian nonparametrics. Note that the expected number of features per user *a priori* is totally controlled by the α parameter in the IBP prior and was set to 3 in this case. While after learning, the expected number *a posteriori* automatically shifted to around 40 that better qualifies the model for explaining the observed preference data.

Fig. 2 shows that the models adapt very well to a fairly board range of truncation level Ks, obtaining similar accuracy even though the number of features inferred from data changes with K.

Apart from MED, the more general regularized Bayesian inference framework [18] also provides basis for building such or even more complex DILFM models, e.g. the iLSVM model [19] for general classification and multi-task learning tasks. It remains an active research area to bridge discriminative learning and Bayesian nonparametrics and develop sophisticated models to learn predictive latent feature representations for applications where input features are noisy, hard to obtain, or at a low level (e.g., image pixels) far away from ideal for concepts to be learned upon.

Furthermore, developing highly efficient and accurate inference algorithms (e.g., Markov chain Monte Carlo methods) is yet another key step to make these models successful and deserves attentions from the research community. Note that recently, an alternative iPM^{3}F model has been proposed [20] which, through adopting a loss term induced from Gibbs classifiers, naturally admits Bayesian inference and enjoys an efficient truncation-free solution via Gibbs sampling.

5. CONCLUSIONS

We introduced two meaningful extensions to latent feature models, one for discriminative max-margin learning via MED and the other for automatical model complexity control via IBP from Bayesian nonparametrics. The resulting discriminative infinite LFMs can be efficiently learned through variational algorithms and can automatically adapt to number of features, showing advantages in modeling big relational data.

Acknowledgments

This work is supported by the National Basic Research Program (973 Program) of China (Nos. 2013CB329403, 2012CB316301), National Natural Science Foundation of China (Nos. 91120011, 61273023), and Tsinghua University Initiative Scientific Research Program (No. 20121088071).

6. REFERENCES

[1] F. Wood and T. Griffiths, "Particle filtering for nonparametric bayesian matrix factorization," in *NIPS*, 2006.

[2] E. Meeds, Z. Ghahramani, R. Neal, and S. Roweis, "Modeling dyadic data with binary latent factors," in *NIPS*, 2007.

[3] K.T. Miller, T.L. Griffiths, and M.I. Jordan, "Nonparametric latent feature models for link prediction," in *NIPS*, 2009.

[4] M. Xu, J. Zhu, and B. Zhang, "Nonparametric max-margin matrix factorization for collaborative prediction," in *NIPS*, 2012.

[5] J. Zhu, "Max-margin nonparametric latent feature models for link prediction," in *ICML*, 2012.

[6] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *NIPS*, 1999.

[7] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the indian buffet process," Tech. Rep. GCNU TR 2005-001, Gatsby Computational Neuroscience Unit, 2005.

[8] T. Jebara, "Discriminative, generative and imitative learning," *PhD Thesis*, 2002.

[9] J. Zhu, A. Ahmed, and E.P. Xing, "MedLDA: Maximum margin supervised topic models for regression and classification," in *ICML*, 2009.

[10] J. Zhu and E.P. Xing, "Maximum entropy discrimination markov networks," *JMLR*, vol. 10, pp. 2531–2569, 2009.

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

[12] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *NIPS*, 2008.

[13] R. Salakhutdinow and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo," in *ICML*, 2008.

[14] N. Srebro, J.D.M. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *NIPS*, 2005.

[15] J.D.M. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *ICML*, 2005.

[16] D. Agarwal and B.C. Chen, "Regression-based latent factor models," in *SIGKDD*. ACM, 2009, pp. 19–28.

[17] I. Porteous, A. Asuncion, and M. Welling, "Bayesian matrix factorization with side information and dirichlet process mixtures," in *AAAI*, 2010.

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

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

[20] M. Xu, J. Zhu, and B. Zhang, "Fast max-margin matrix factorization with data augmentation," in *ICML*, 2013.