Nonparametric Max-Margin Matrix Factorization for Collaborative Prediction

Minjie Xu, Jun Zhu and Bo Zhang

Department of Computer Science and Technology, LITS, TNList, Tsinghua University

Abstract

We present a ***probabilistic*** formulation of ***max-margin matrix factorization*** and build accordingly a ***nonparametric*** Bayesian model which automatically resolves the unknown number of latent factors.

Our work demonstrates a successful example that integrates ***Bayesian nonparametrics*** and ***max-margin learning***, which are conventionally two separate paradigms and enjoy complementary advantages.

We develop an efficient ***variational*** algorithm for posterior inference, and our extensive empirical studies on large-scale MovieLens 1M and EachMovie data sets appear to justify the aforementioned dual advantages.

Method Outline

1. **PM^{3}F**: The first key step is a general probabilistic formulation of the standard max-margin matrix factorization (**M^{3}F**), which is based on the maximum entropy discrimination (**MED**) principle.

2. **iPM^{3}F**: We can then extend it to a nonparametric model, which in theory has an unbounded number of latent factors. To avoid over-fitting, we impose a sparsity-inducing Indian buffet process (**IBP**) prior on the latent coefficient matrix, selecting only an *appropriate* number of active factors.

3. **iBPM^{3}F**: To make one step further towards a hierarchical Bayesian-style model (**HBM**), we introduce priors for hyper-parameters and perform fully-Bayesian inference, where model parameters and hyper-parameters are integrated out when making prediction.

M^{3}F + **MED** → **PM^{3}F** **P**robabilistic

+ **IBP** → **iPM^{3}F** **I**nfinite

+ **HBM** → **iBPM^{3}F** **B**ayesian

From M^{3}F to iBPM^{3}F

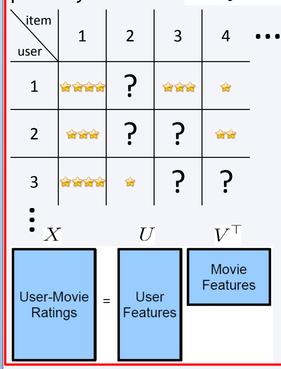
Given a preference matrix for N users and M items, which is only partially observed: $Y \in \mathcal{Y}^{N\times M}$, $\mathcal{I} = \{(i,j)|Y_{ij} \text{ is available}\}$, M^{3}F approximates it with X under a **trace norm** regularizer and **hinge loss** error measure (binary case):

$$\min_X \ \|X\|_* + C\sum_{ij\in\mathcal{I}} h\left(Y_{ij}X_{ij}\right)$$

Or equivalently (due to the variational form of trace norm):

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

M^{3}F



Problem:
Original formulation scales poorly;
Variational formulation requires an explicit selection of the dimensionality of the latent feature space.

To solve the problem, we first endow M^{3}F with a probabilistic background according to the principle of MED.

For a training data set $\{(\mathbf{x}_d, y_d)\}_{d=1}^D$ $(y_d \in \{\pm 1\})$ and a discriminant function $F(\mathbf{x};\boldsymbol{\eta})$, MED seeks to learn a **distribution** $p(\boldsymbol{\eta})$ by solving

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

and predicts by $\hat{y} = \mathrm{sign}\left(\mathbb{E}_p[F(\mathbf{x};\boldsymbol{\eta})]\right)$.

MED

MED provides an elegant way to integrate discriminative max-margin learning and Bayesian generative modeling. In fact, MED subsumes SVM as a special case.

In the case of M^{3}F, we define $F((i,j);U,V) = U_iV_j^\top$ and thus have:

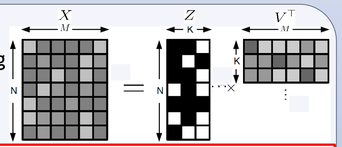
$$\min_{p(U,V)} \ \mathrm{KL}(p(U,V)\|p_0(U,V)) + C\sum_{ij\in\mathcal{I}} h_\ell\left(Y_{ij}\mathbb{E}_p[U_iV_j^\top]\right)$$

- observed feature vector: entry indices (i,j)
- prediction rule: $\widehat{Y}_{ij} = \mathrm{sign}\left(\mathbb{E}_p[U_iV_j^\top]\right)$
- subsumes M^{3}F as a special case
- still suffers from explicit dimensionality selection problem without proper treatment

PM^{3}F

To automatically infer the latent feature dimension, we place an IBP prior over U, resulting in a binary coefficient matrix with an **unbounded** number of columns (features).

We focus on the stick-breaking construction of IBP since it enables efficient variational inference.



IBP prior

- distribution over equivalent classes of binary matrices with a finite number of rows and an unbounded number of columns

stick-breaking construction

- **augment** the representation with stick-breaking variables $\nu_{1,\ldots,\infty}$

$$\pi_1 = \nu_1,\ \pi_k = \nu_k\pi_{k-1} = \prod_{i=1}^k \nu_i,\ \text{where } \nu_i \sim \mathrm{Beta}(\alpha,1) \qquad \text{i.i.d. for } i = 1,\ldots,+\infty$$
$$Z_{ik} \sim \mathrm{Bernoulli}(\pi_k) \qquad \text{i.i.d. for } i = 1,\ldots,N \quad (\forall k)$$

- results in a descending sequence of π_k
- number of active factors $K_+ \sim \mathrm{Poisson}(\alpha H_N)$

IBP

We place an isotropic Gaussian prior over V and we have iPM^{3}F:

$$\min_{p(\nu,Z,V)} \ \mathrm{KL}(p(\nu,Z,V)\|p_0(\nu,Z,V)) + C\sum_{ij\in\mathcal{I}} h_\ell\left(Y_{ij}\mathbb{E}_p[Z_iV_j^\top]\right)$$

where $p_0(\nu,Z,V) = p_0(\nu)p_0(Z|\nu)p_0(V)$ with

$$\nu_k \sim \mathrm{Beta}(\alpha,1) \qquad \text{i.i.d. for } k = 1,\ldots,+\infty,$$
$$Z_{ik}|\nu \sim \mathrm{Bernoulli}(\pi_k) \qquad \text{i.i.d. for } i = 1,\ldots,N \quad (\forall k),$$
$$V_{jk} \sim \mathcal{N}(0,\sigma^2) \qquad \text{i.i.d. for } j = 1,\ldots,M,\ k = 1,\ldots,+\infty$$

iPM^{3}F

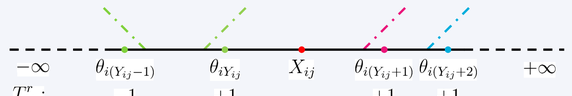
We may further place a Gaussian-Wishart prior over V and its hyper-parameters μ and Ω to build a **HBM** and perform fully Bayesian inference, hence the **iBPM^{3}F** (omitted to save space).

For ordinal ratings where $\mathcal{Y} = \{1,\ldots,L\}$, we introduce thresholds $-\infty = \theta_0 \le \theta_1 \le \cdots \le \theta_{L-1} \le \theta_L = +\infty$ and

- predict by $\widehat{Y}_{ij} = \max\left\{r|\mathbb{E}_p[U_iV_j^\top] \ge \theta_r\right\} + 1$
- rewrite loss as $\sum_{ij\in\mathcal{I}}\sum_{r=1}^{L-1} h_\ell\left(T_{ij}^r(\theta_r - \mathbb{E}_p[U_iV_j^\top])\right)$, where $T_{ij}^r = \begin{cases}+1 & \text{for } r \ge Y_{ij}\\ -1 & \text{for } r < Y_{ij}\end{cases}$
- the loss is an upper bound of sum absolute error
- introduce user-specific thresholds θ_{ir}
- treat thresholds as random variables, hereby the full model

$$\min_{p(U,V,\theta)} \ \mathrm{KL}(p(U,V,\theta)\|p_0(U,V,\theta)) + C\sum_{ij\in\mathcal{I}}\sum_{r=1}^{L-1} h_\ell\left(T_{ij}^r(\mathbb{E}_p[\theta_{ir}] - \mathbb{E}_p[U_iV_j^\top])\right)$$

- place a Gaussian prior, guiding a set of large-margin thresholds



Learning and Inference

We perform variational inference under the truncated mean-field assumption (K the truncation level):

$$p(\nu,Z,V,\theta) = \textstyle\prod_{k=1}^K p(\nu_k)p(Z)p(V)p(\theta),\ \nu_k \sim \mathrm{Beta}(\gamma_{k1},\gamma_{k2})$$

For iPM^{3}F (iBPM^{3}F likewise), we have (details in Appendix C & D)

- $p(V) = \prod_{j=1}^M \mathcal{N}(V_j|\Lambda_j,\sigma^2 I)$ and each Λ_j is solution to a linear SVM
- γ_{k1},γ_{k2} follows the same approximate update rule as developed by (Doshi-Velez, et al., AISTATS'09)
- $p(Z) = \prod_{i=1}^N\prod_{k=1}^K \mathrm{Bernoulli}(Z_{ik}|\psi_{ik})$ and each ψ_i can be easily updated via coordinate descent
- $p(\theta) = \prod_{i=1}^N\prod_{r=1}^{L-1}\mathcal{N}(\theta_{ir}|\varrho_{ir},\varsigma^2)$ and each ϱ_{ir} is solution to a linear SVM

We iterate through each component until convergence.

Note that the number of ***active*** factors ***a posteriori*** follows

$$\mathbb{E}_p[K_+] = \textstyle\sum_{k=1}^K\left(1 - \prod_{i=1}^N(1-\psi_{ik})\right)$$

Experimental Results

We test on MovieLens 1M and EachMovie data sets and compare with M^{3}F, PMF and BPMF under NMAE error measure.

Table 1: NMAE performance of different models on MovieLens and EachMovie.

Algorithm	MovieLens		EachMovie	
	weak	strong	weak	strong
M^{3}F [11]	.4156 ± .0037	.4203 ± .0138	.4397 ± .0006	.4341 ± .0025
PMF [13]	.4332 ± .0033	.4413 ± .0074	.4466 ± .0016	.4579 ± .0016
BPMF [12]	.4235 ± .0023	.4450 ± .0085	.4352 ± .0014	.4445 ± .0005
M^{3}F*	.4176 ± .0016	.4227 ± .0072	.4348 ± .0023	.4301 ± .0034
iPM^{3}F	**.4031** ± .0030	.4135 ± .0109	**.4211** ± .0019	**.4224** ± .0051
iBPM^{3}F	.4050 ± .0029	**.4089** ± .0146	.4268 ± .0029	.4403 ± .0040

Performance of iPM^{3}F on MovieLens 1M

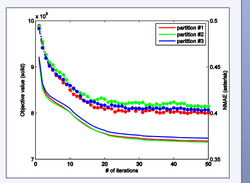
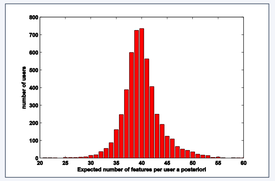


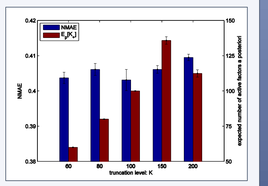
Table 3: Performance of iPM^{3}F with and without probabilistic treatment of θ

Algorithm	MovieLens	EachMovie	pEachMovie
w/ prob.	.4031 ± .0030	.4211 ± .0019	.3954 ± .0026
w/o prob.	.4056 ± .0043	.4256 ± .0011	.4026 ± .0023
margin	.0024 ± .0013	.0045 ± .0016	.0072 ± .0045

Expected number of features per user *a posteriori* (3 *a priori*)



Influence of the truncation level



Conclusion

We endow M^{3}F with a probabilistic background, which further enables the integration of nonparametric Bayesian techniques to *automatically* infer the dimensionality of the latent feature space and thus bypass explicit model complexity control. Our extensive experimental studies verify this benefit and demonstrate that our methods have competitive performance on real word collaborative filtering data sets as well.