Bayesian Matrix Factorization

By
Subham Kumar, 160707
Sudhanshu Bansal, 160710
Suryateja B.V., 160729

Overview

- Tackling Inductive Bias in recommender systems by modelling user exposure to items as latent variables
  Reference - *Modeling User Exposure in Recommendation, Liang, Dawen and Charlin, Laurent and McInerney, James and Blei, David M.*

- Incorporating item-item co-occurrence matrix, a non-linear transformation of user-item matrix, as a regularizer
  Reference - *Factorization Meets the Item Embedding: Regularizing Matrix Factorization with Item Co-occurrence, Liang, Dawen and Altosaar, Jaan and Charlin, Laurent and Blei, David M.*

- Explore ways to incorporate fairness in recommender systems using constrained bandits.
  Reference - *Controlling polarization in personalization: An algorithmic framework, L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi*

Previous Work

Rudimentary Model

$$\boldsymbol{\theta}_u \sim \mathcal{N}\left(\boldsymbol{\theta}_u | 0, \lambda_\theta \mathbf{I}^{-1}\right)$$
$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}_i | 0, \lambda_\beta \mathbf{I}^{-1}\right)$$
$$y_{ui} \sim \mathcal{N}\left(y_{ui} | \boldsymbol{\theta}_u^T \boldsymbol{\beta}_i, \sigma_y^2 \mathbf{I}^{-1}\right)$$

Inductive Bias while modelling implicit data :(

Weighted Matrix Factorization (WMF)

$$\boldsymbol{\theta}_u \sim \mathcal{N}\left(\boldsymbol{\theta}_u | 0, \lambda_\theta \mathbf{I}^{-1}\right)$$
$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}_i | 0, \lambda_\beta \mathbf{I}^{-1}\right)$$
$$y_{ui} \sim \mathcal{N}\left(y_{ui} | \boldsymbol{\theta}_u^T \boldsymbol{\beta}_i, c_{ui} \mathbf{I}^{-1}\right) \qquad c_{y=1} > c_{y=0}$$

Setup a hyper parameter which gives more weightage to seen data. Issue? Not a fully generative model. Heuristic

Modeling User Exposure

$$
\begin{aligned}
\boldsymbol{\theta}_u &\sim \mathcal{N}\left(\boldsymbol{\theta}_u|0, \lambda_\theta \mathbf{I}_K^{-1}\right) \\
\boldsymbol{\beta}_i &\sim \mathcal{N}\left(\boldsymbol{\beta}_i|0, \lambda_\beta \mathbf{I}_K^{-1}\right) \\
\mu_{ui} &= \sigma\left(\boldsymbol{\psi}_u^T \mathbf{x}_i\right) \\
a_{ui} &\sim \text{Bernoulli}\left(\mu_{ui}\right) \\
y_{ui}|a_{ui} = 1 &\sim \mathcal{N}\left(y_{ui}|\boldsymbol{\theta}_u^T \boldsymbol{\beta}_i, \lambda_y^{-1}\right) \\
y_{ui}|a_{ui} = 0 &\sim \delta_0
\end{aligned}
$$

$$
\begin{aligned}
&\log p(a_{ui}, y_{ui} \mid \mu_{ui}, \boldsymbol{\theta}_u, \boldsymbol{\beta}_i, \lambda_y^{-1}) \\
&\quad = \log \text{Bernoulli}(a_{ui} \mid \mu_{ui}) + a_{ui} \log \mathcal{N}(y_{ui} \mid \boldsymbol{\theta}_u^\top \boldsymbol{\beta}_i, \lambda_y^{-1}) \\
&\quad + (1 - a_{ui}) \log \mathbb{I}[y_{ui} = 0], \qquad (2)
\end{aligned}
$$

- ψ_u and x_i are exposure covariate and task based representation respectively.More intuitively if x_il is recommender system and user u is interested in this topic then we would like ψ_{ul} to be high.
- Here δ_0 is 1 if y_ui is 0 which makes sense as in such a case user won't be able to click
- Any y_ui>0 implies a_ui=1.Only when y_ui is 0 then a_ui is latent(may be not interested or unexposed).
- Also note the log-joint at the bottom.It generalizes all the previous two discussed related works.When a_ui=1,it is normal MF and similar to WMF, it shares same feature to selectively downweighting the evidence from the click matrix(can be noted when a_ui=0).

Inference - Generalized EM Algorithm

- In the E-step we calculate the E[a_ui | y_ui=0](p_ui) else p_ui=1 if y_ui=1.Note that E-step can be done lazily i.e. only required parts of matrix A is constructed for updates.This saves a lot of storage.

$$\mathbb{E}[a_{ui} \,|\, \boldsymbol{\theta}_u, \boldsymbol{\beta}_i, \mu_{ui}, y_{ui} = 0] = \frac{\mu_{ui} \cdot \mathcal{N}(0|\boldsymbol{\theta}_u^\top \boldsymbol{\beta}_i, \lambda_y^{-1})}{\mu_{ui} \cdot \mathcal{N}(0|\boldsymbol{\theta}_u^\top \boldsymbol{\beta}_i, \lambda_y^{-1}) + (1 - \mu_{ui})}.$$

- In M-step updates are as usual except that we now need to update exposure covariates as well. Its update were not in closed form so used batch gradient descent.

$$\boldsymbol{\theta}_u \leftarrow (\lambda_y \textstyle\sum_i p_{ui} \boldsymbol{\beta}_i \boldsymbol{\beta}_i^\top + \lambda_\theta I_K)^{-1} (\textstyle\sum_i \lambda_y p_{ui} y_{ui} \boldsymbol{\beta}_i)$$
$$\boldsymbol{\beta}_i \leftarrow (\lambda_y \textstyle\sum_u p_{ui} \boldsymbol{\theta}_u \boldsymbol{\theta}_u^\top + \lambda_\beta I_K)^{-1} (\textstyle\sum_u \lambda_y p_{ui} y_{ui} \boldsymbol{\theta}_u)$$

Inference for exposure covariate

$$\boldsymbol{\psi}_u^{\text{new}} \leftarrow \boldsymbol{\psi}_u + \eta \nabla_{\boldsymbol{\psi}_u} \mathcal{L},$$

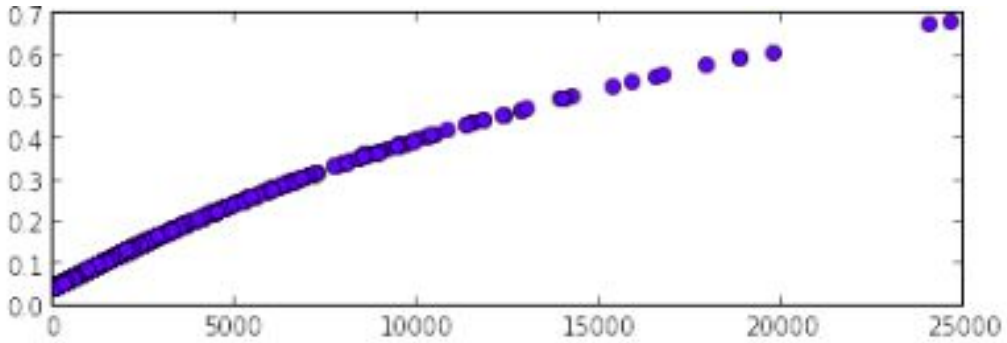
for some learning rate η, where

$$\nabla_{\boldsymbol{\psi}_u} \mathcal{L} = \tfrac{1}{I} \textstyle\sum_i (p_{ui} - \sigma(\boldsymbol{\psi}_u^\top \mathbf{x}_i)) \mathbf{x}_i.$$

Experiments

	recall@20	recall@50	NDCG@100	MAP@100
W/o exposure covariate	0.0906	0.1450	0.0917	0.0324
With exposure covariate	0.1288	0.1988	0.1252	0.0477

Probability exposure prior



Item popularity

Item-item co-occurrence Matrix

- Inspired from word2vec word-word co-occurrence matrix. We can think of users as documents and sequential items collected by a user as sequential words of a document.
- Jointly learn item embeddings.
- w_i and c_j are item and context biases resp.
- When we look at the loss function, we just see an additional regularization term!
- Matrix M is not *new* data. It is just a non-linear transformation of matrix Y. We create M as a pointwise mutual information (PMI) matrix.
- (ij)th entry of M is directly proportional to number of users picking both item i and item j.

$$\boldsymbol{\theta}_u \sim \mathcal{N}\left(\boldsymbol{\theta}_u | 0, \lambda_\theta \mathbf{I}_K^{-1}\right)$$
$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}_i | 0, \lambda_\beta \mathbf{I}_K^{-1}\right)$$
$$\boldsymbol{\gamma}_i \sim \mathcal{N}\left(\boldsymbol{\gamma}_i | 0, \lambda_\gamma \mathbf{I}_K^{-1}\right)$$
$$y_{ui} \sim \mathcal{N}\left(y_{ui} | \boldsymbol{\theta}_u^T \boldsymbol{\beta}_i, c_{ui}^{-1}\right) \qquad c_{y=1} > c_{y=0}$$
$$m_{ij} = \boldsymbol{\beta}_i^T \boldsymbol{\gamma}_j - \mathbf{w}_i - \mathbf{c}_j$$

Note : M is not a part of generative story!

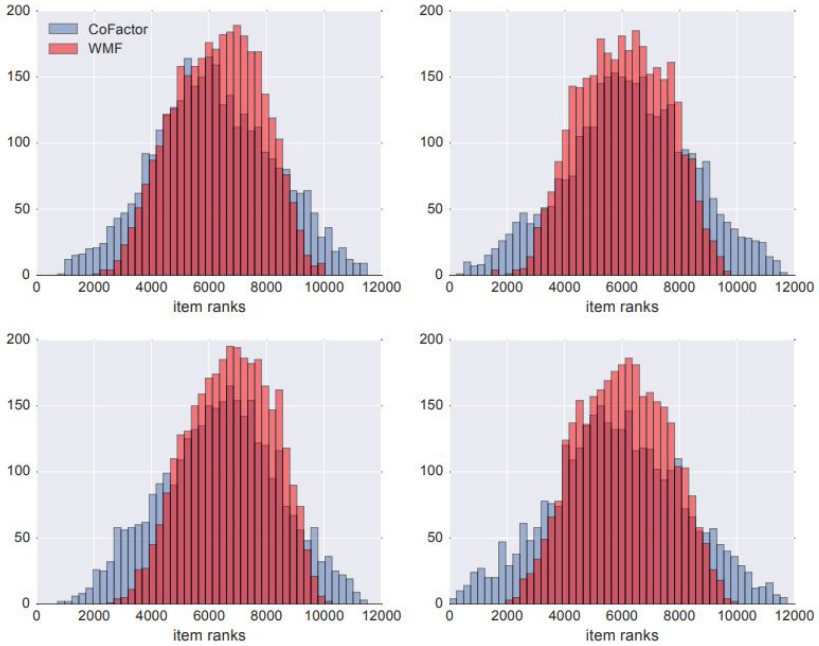
$$\mathcal{L}_{\text{co}} = \overbrace{\sum_{u,i} c_{ui}(y_{ui} - \theta_u^\top \beta_i)^2}^{\text{MF}} + \overbrace{\sum_{m_{ij} \neq 0} \left(m_{ij} - \beta_i^\top \gamma_j - w_i - c_j\right)^2}^{\text{item embedding}}$$
$$+ \lambda_\theta \sum_u \|\theta_u\|_2^2 + \lambda_\beta \sum_i \|\beta_i\|_2^2 + \lambda_\gamma \sum_j \|\gamma_j\|_2^2 \qquad (3)$$

Item Embeddings - Inference and Evaluation

- Closed form coordinate updates. Make sure c_{ui} is not too large, else the usual matrix factorization term dominates.
- What should be taken as the context of an item? We don't usually have timestamps. So we can't adopt a skipgram-like approach. If a user who picks *i* also picks *j,* then j is part of the context of *i*.
- We experimented on **Movielens** (20M) dataset, with timestamps. To binarize data, all ratings greater than 3 were considered as a click. Training, validation and test sets prepared from the same data.
- Evaluation metrics - Recall@M, DCG@M.
- DCG@M - Discount Cumulated Gain: Compute the predictions. Sort them in descending order and consider the top M. If the user actually had picked items from these top M items, our model has worked well. DCG gives greater weight to items at the top of M items.

Item Embeddings - Key Insights

- For users who have consumed less items, WMF suffers from cold-start. However, co-occurrence matrix delivers good signals of pairwise interaction of items, and learns better latent representations
- Another aspect where WMF suffers is diversity. Rare items are ranked middle to low due to their high variance. In case of co-occurrence matrix, they are typically ranked high or low.
  High - for example, we expect rare movies to co-occur.
  Low - Some movies never co-occur!



Exposure + Co-occurrence

- We've added the item embedding regularization term to user exposure loss.

$$\mathcal{L} = \sum_{u,i} \lambda_y a_{ui} \left(y_{ui} - \boldsymbol{\theta}_u^T \boldsymbol{\beta}_i\right)^2 - 2a_{ui} \log \mu_{ui} - 2\left(1 - a_{ui}\right)\left(\log \mathbb{I}\left\{y_{ui} = 0\right\} + \log\left(1 - \mu_{ui}\right)\right)$$

$$\sum_{m_{ij} \neq 0} \left(m_{ij} - \boldsymbol{\beta}_i \boldsymbol{\gamma}_j - \mathbf{w}_i - \mathbf{c}_j\right)^2 + \lambda_\theta \sum_u \|\boldsymbol{\theta}_u\|^2 + \lambda_\beta \sum_i \|\boldsymbol{\beta}_i\|^2 + \lambda_\gamma \sum_j \|\boldsymbol{\gamma}_j\|^2$$

- Derived a similar EM flavored algorithm incorporating the additional regularization terms.
- Very slow training. Not good results
- Possible improvements - try user-user co-occurrence matrix.

Constrained Bandits for Fairness

- Bandits induce polarization - the entire probability mass ends up on a single optimal arm.
- We might have to hide some data, but that can hurt predictive power.
- We can think of each arm belonging to one of G groups (topic modelling, clustering). Impose constraints on the values the probability mass can take - limits for each group set by the users.

$$\ell_i \leq \sum_{a \in G_i} w_a(G_i) \cdot p_a^t \leq u_i \quad \forall i \in [g], \forall t \in [T].$$

- How is this *fair*? Constraints ensure one arm doesn't too high a probability (or too low).
- Algorithm proposed by paper - constrained epsilon-greedy ➜ stay afloat, maintain some probability in each group.

Questions?

Thank You!