Factorizing Personalized Markov Chains for Next-Basket Recommendation

Steffen Rendle
Department of Reasoning for intelligence
The Institute of Scientific and Industrial Research
Osaka University, Japan
rendle@ar.sanken.osaka-u.ac.jp

Introduction

- Two of the most popular approaches are based on matrix factorization (MF) and Markov chains (MC).
- MF methods learn the general taste of a user by factorizing the matrix over observed user-item preferences.
- MC methods model sequential behavior by learning a transition graph over items that is used to predict the next action based on the recent actions of a use.
- In this paper, we present a method bringing both approaches together.
- Our method is based on personalized transition graphs over underlying Markov chains. That means for each user an own transition matrix is learned thus in total the method uses a transition cube and farctorize it with a pairwise interaction model (Tucker Decomposition special case)
- Empirically, we show that our FPMC (Factorized Personalized Markov Chains) model outperforms both the common matrix factorization and the unpersonalized MC model both learned with and without factorization.

Basic Method

- Introduce personalized Markov chains relying on personalized transition matrices. This allows to capture both sequential effects and long term user-taste. We show that this is a generalization of both standard MC and MF models.

- To deal with the sparsity for the estimation of transition probabilities, we introduce a factorization model that can be applied both to personalized and normal transition matrices. This factorization approach results in less parameters and due to generalization to a better quality than full parametrized models.

FACTORIZING PERSONALIZED MARKOV CHAINS (FPMC)

- First, we introduce MC for sequential set data and extend this to personalized MCs.

- We discuss the weakness of Maximum Likelihood Estimates for the transition cubes.

- To solve this, we introduce factorized transition cubes where information among transitions is propagated.

Markov Chains for Sequential Sets

- Markov Chain for Sets:

$$p(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-m} = x_{t-m})$$

- In recommender scenarios without sets, usually longer chains (e.g. m = 3) are preferable, but In our case with sets, even a chain with length m = 1 is reasonable because it relies already on many items (all items of the basket)

$$a_{l,i} := p(i \in B_t | l \in B_{t-1})$$

- Transition over the basket can be defined as probability of selecting the item with the items in the basket from past order.

And the full Markov chain over baskets can be expressed by:

$$p(B_t | B_{t-1}) \propto \prod_{i \in B_t} p(i | B_{t-1}) \qquad (5)$$

Personalized Markov Chains for Sets

- We represent each MC by the transitions over items, but now user-specific:

$$a_{u,l,i} := p(i \in B_t^u | l \in B_{t-1}^u)$$

- Thus also the prediction depends only on the user’s transitions

$$p(i \in B_t^u | B_{t-1}^u) := \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} p(i \in B_t^u | l \in B_{t-1}^u) \qquad ($$

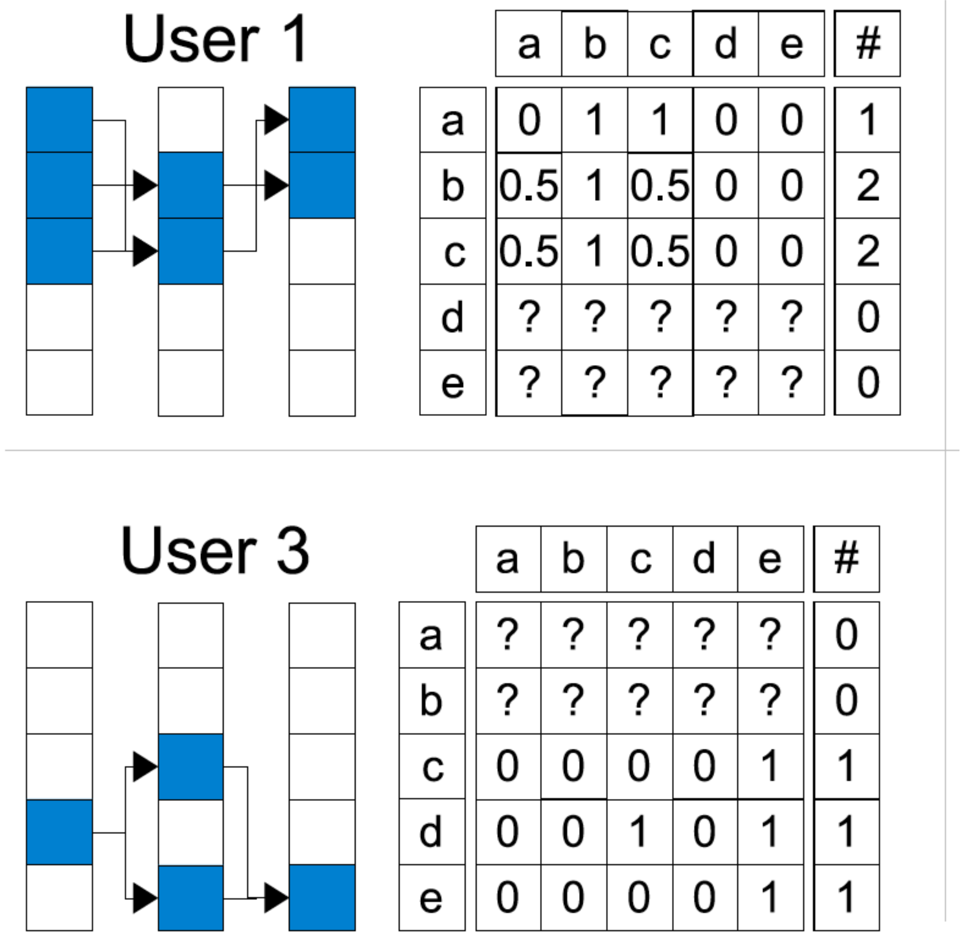
Limitations

- MLE estimates each transition parameter $a_{l,i}$ independently from the others, i.e. none of the cooccurrences (l, i) will contribute to another transition probability estimator

- This is even worse for personalized where another parameter of user is added and one value of user does not effects the value of different user hence the computation efforts are increased more significantly.

Factorizing Transition Graphs

- That means we model the unobserved transition tensor A by a low rank approximation Aˆ. The advantage of this approach over a full parametrization is that it can handle sparsity and generalizes to unobserved data.



User 1

	a	b	c	d	e	#
a	0	1	1	0	0	1
b	0.5	1	0.5	0	0	2
c	0.5	1	0.5	0	0	2
d	?	?	?	?	?	0
e	?	?	?	?	?	0

User 3

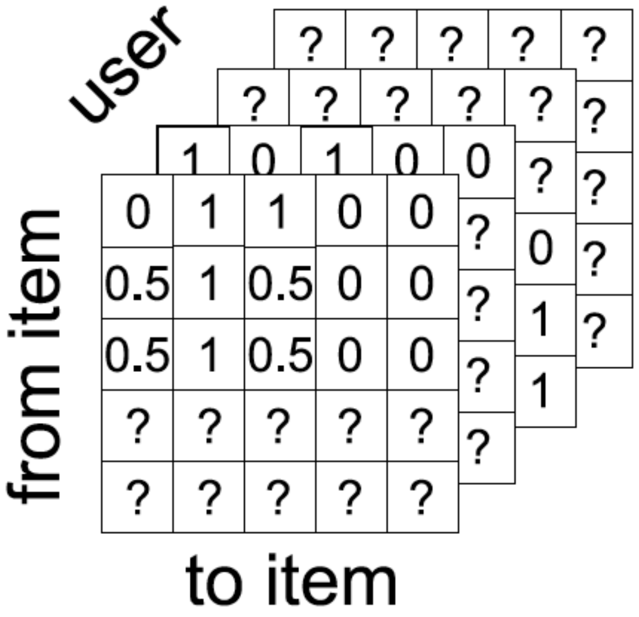
	a	b	c	d	e	#
a	?	?	?	?	?	0
b	?	?	?	?	?	0
c	0	0	0	0	1	1
d	0	0	1	0	1	1
e	0	0	0	0	1	1

Factorizing Transition Cubes

- A general linear factorization model for estimating the tensor A is the Tucker Decomposition (TD):

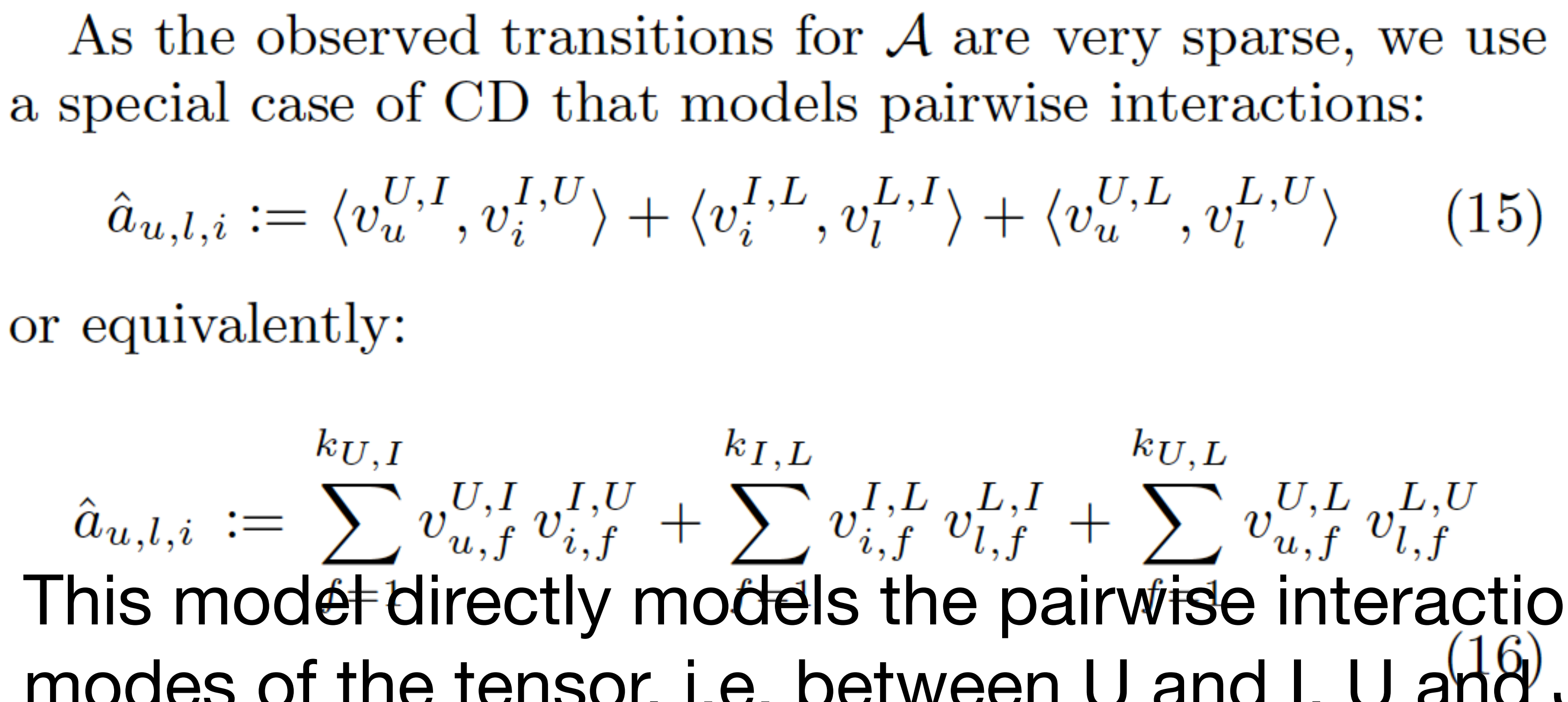
$$\hat{\mathcal{A}} := \mathcal{C} \times_U V^U \times_L V^L \times_I V^I$$

- Where C is a core tensor and V_U is the feature matrix for the users, V_L is the feature matrix for the items in the last transition (outgoing nodes) and $V_I$is the feature matrix for the items to predict (ingoing nodes). We improve it with Canonical Decomposition (CD) aka parallel factor analysis (PARAFAC).



0	1	1	0	0
0.5	1	0.5	0	0
0.5	1	0.5	0	0
?	?	?	?	?
?	?	?	?	?

Factorizing Transition Cubes



As the observed transitions for $\mathcal{A}$ are very sparse, we use a special case of CD that models pairwise interactions:

$$\hat{a}_{u,l,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle \qquad (15)$$

or equivalently:

$$\hat{a}_{u,l,i} := \sum_{f=1}^{k_{U,I}} v_{u,f}^{U,I}\, v_{i,f}^{I,U} + \sum_{f=1}^{k_{I,L}} v_{i,f}^{I,L}\, v_{l,f}^{L,I} + \sum_{f=1}^{k_{U,L}} v_{u,f}^{U,L}\, v_{l,f}^{L,U} \qquad (16)$$

- This model directly models the pairwise interaction between all three modes of the tensor, i.e. between U and I, U and J as well as J and I (i.e. user U, item I, item J),

- This decomposition is especially important for applications with a high number of items where a full parametrization with $|I|^2$ parameters might not be feasible.

Optimization Criterion S-BPR

- To model the ranking, we assume there is an estimator ˆx : U X T X I → R – e.g. the buying probability of the personalized Markov Chain – which is used to define the ranking

- Next we derive the sequential BPR (S-BPR) optimization criterion analogously to the general Bayesian Personalized Ranking approach.

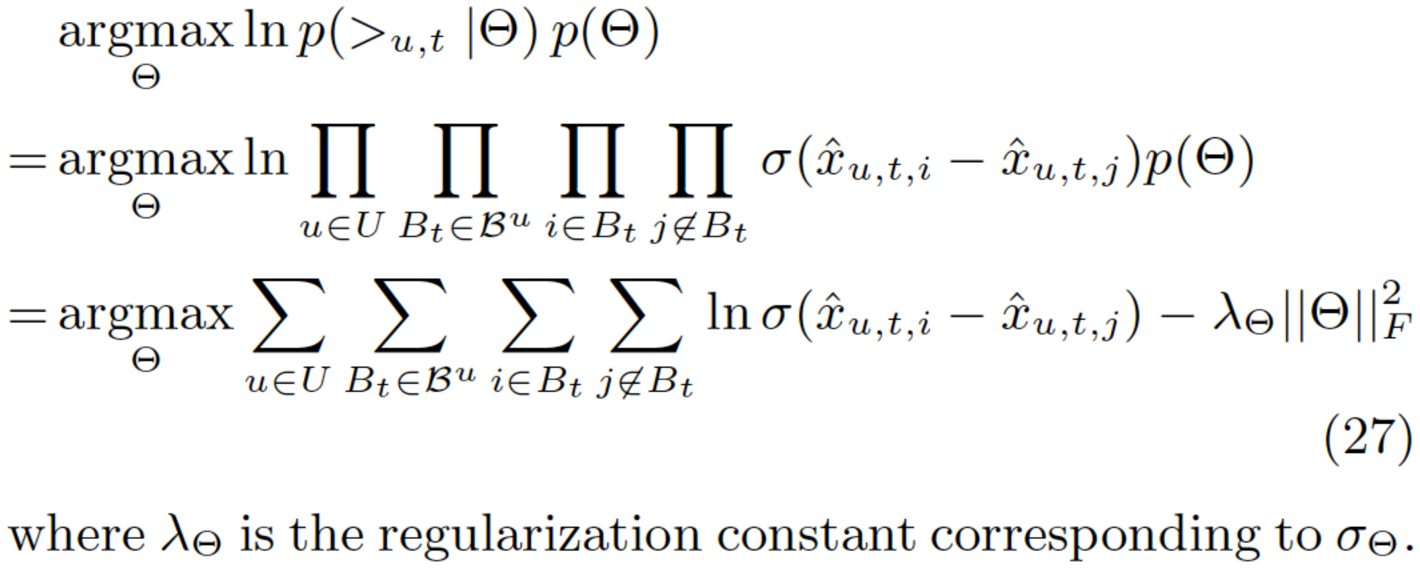
- The best ranking for user u at time t can be formalized as:

$$p(\Theta| >_{u,t}) \propto p(>_{u,t} |\Theta)\, p(\Theta)$$

- Where Θ are the model parameters – in our case the parameters are = {$V_{U,I}$, $V_{I,U}$, $V_{L,I}$, $V_{I,L}$, $V_{U,L}$, $V_{L,U}$}.

Optimization Criterion S-BPR (2)

- Maximum a Posterior (MAP) :



$$\begin{aligned}
&\operatorname*{argmax}_{\Theta} \ln p(>_{u,t} |\Theta)\, p(\Theta) \\
&= \operatorname*{argmax}_{\Theta} \ln \prod_{u\in U} \prod_{B_t\in\mathcal{B}^u} \prod_{i\in B_t} \prod_{j\notin B_t} \sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) p(\Theta) \\
&= \operatorname*{argmax}_{\Theta} \sum_{u\in U} \sum_{B_t\in\mathcal{B}^u} \sum_{i\in B_t} \sum_{j\notin B_t} \ln \sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) - \lambda_{\Theta} ||\Theta||_F^2
\end{aligned} \tag{27}$$

where λ_Θ is the regularization constant corresponding to σ_Θ.

- Where sigmoid is used for each item pair and their loss.

Item Recommendation with FPMC

- For item recommendation with FPMC, we express ˆx by the FPMC model and apply S-BPR.

$$
\begin{aligned}
&\hat{x}'_{u,t,i} := \hat{p}(i \in B_t^u | B_{t-1}^u) \\
&= \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \left(\langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle \right)
\end{aligned}
$$

CONCLUSION

- In this paper, we have introduced a recommender method based on personalized Markov chains over sequential set data.

- Instead of using the same transition matrix for all users, this method uses an individual transition matrix for each user which in total results in a transition cube.

- As direct estimation (e.g. by Maximum Likelihood) over a full parametrized transition cube leads to very poor estimates, we introduce a factorization model that gives a low-rank approximation to the transition cube.

Thank You

–Shivam Mehta