Factorizing Personalized Markov Chains for Next-Basket Recommendation

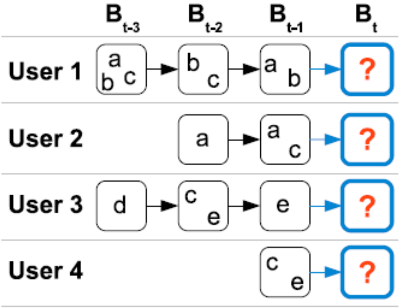
Prashant Gaurav

School of Computational Science & Engineering
College of Computing
Georgia Institute of Technology

March 2, 2011

Next-Basket Recommendation

- Scenario: Sequential basket data is given per user (e.g. online shopping)
- Goal: To reccommend items to the user that he might want to buy in his next visit



Factorized Personalized Markov Chains (FPMC)

- A generalization of Matrix Factorization (MF) and Markov Chain (MC) models
  — Allows to capture both sequential and long term user-taste.
- A factorization model that results in less parameter and overcomes the limitations of MLE
- Outperforms MF and MC models both on sparse and dense datasets
- Addresses the problem setting with *set data*
  — e.g. in online-shopping usually a basket of products is bought at the same time

Notation

$U = \{u_1, \ldots, u_{|U|}\}$ denotes a set of users.

$I = \{i_1, \ldots, i_{|I|}\}$ denotes a set of items.

For each user u, a purchase history $\mathbf{B}^u$ of his baskets is known:
$\mathbf{B}^u := (B_1^u, \ldots, B_{t_u-1}^u)$ with $B_t^u \subset I$.

The purchase history of all users is $\mathbf{B} := \{\mathbf{B}^{u_1}, \ldots, \mathbf{B}^{u_{|U|}}\}$

Given this history, item recommendation task can be formalized in creating personal ranking

$$<_{u,t} \subset I^2$$

over all pairs of items for user u for his t-th basket.

Unpersonalized Markov Chains (MC) for Sets

For basket problem, a MC of order $m = 1$ with $p(B_t|B_{t-1})$.
— $m = 1$ is resonable in this case
— The dimension of the transition matrix A would be $2^{|I|} * 2^{|I|}$

Instead model over $|I|$ binary variables $a_{l,i} := p(i \in B_t | l \in B_{t-1})$
— The dimension of the transition matrix A is $|I|^2$.
— The transition matrix A is not stochastic i.e. $\sum_{i \in I} a_{l,i} \neq 1$

The probability of purchasing an item given last basket of a user is

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

The MLE estimate for $a_{l,i}$ given the data **B** is:

$$\hat{a}_{l,i} = \frac{|(B_t, B_{t-1} : i \in B_t \wedge l \in B_{t-1})|}{|(B_t, B_{t-1}) : l \in B_{t-1}|}$$

Personalized Markov Chains for Sets

Extending to personalized MC per user:

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

The prediction becomes:

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

The MLE estimate for $a_{u,l,i}$ given the data $\mathbf{B}^u$ is:

$$\hat{a}_{u,l,i} = \frac{|(B_t^u, B_{t-1}^u : i \in B_t^u \wedge l \in B_{t-1}^u)|}{|(B_t^u, B_{t-1}^u) : l \in B_{t-1}^u|}$$

Instead of transition matrix, we have a transition tensor $\mathbf{A} \in [0,1]^{|U|*|I|*|I|}$

Limitations of MLE method

- MLE estimates each transition parameter indendent of other parameters
- In current scenario, data is extremely sparse, MLE model may suffer undefitting.

Factorizing Transition Tensor

The factorization model for the tensor models the pairwise interaction between all modes of tensor (user U, item I, item L):

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

or equivalently:

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

For each mode, the pair of factorization matrices are :

- $U - I : V^{U,I} \in \mathbb{R}^{|U| * k_{U,I}}, V^{I,U} \in \mathbb{R}^{|I| * k_{U,I}}$
- $I - L : V^{I,L} \in \mathbb{R}^{|I| * k_{I,L}}, V^{L,I} \in \mathbb{R}^{|I| * k_{I,L}}$
- $U - L : V^{U,L} \in \mathbb{R}^{|U| * k_{U,L}}, V^{L,U} \in \mathbb{R}^{|I| * k_{U,L}}$

Summary: FPMC

We model $p(i \in B_t^u | l \in B_{t-1}^u)$ with factorization cube as :

$$\hat{p}(i \in B_t^u | B_{t-1}^u) = \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \hat{a}_{u,l,i}$$
$$= \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} (\langle v_u^{U,I}, v_i^{I,U} \rangle + \langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle)$$

Since factorization (U,I) is independent of L :

$$\hat{p}(i \in B_t^u | B_{t-1}^u) = \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \left(\langle v_i^{I,L}, v_l^{L,I} \rangle + \langle v_u^{U,L}, v_l^{L,U} \rangle \right)$$

Optimization formulation (S-BPR)

To model the ranking, an estimator $\hat{x} : U * T * I \to \mathbb{R}$

$$i >_{u,t} j \;\; :\Leftrightarrow \;\; \hat{x}_{u,t,i} >_{\mathbb{R}} \hat{x}_{u,t,j}$$

The best ranking $>_{u,t} \subset I^2$ 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

Assuming the independence of users and backets, the maximum MAP estimator of the model parameters is :

$$\underset{\Theta}{\operatorname{argmax}} \prod_{u \in U} \prod_{B_t \in \mathcal{B}^u} p(>_{u,t} |\Theta) \, p(\Theta)$$

Optimization formulation ...

Expanding $>_{u,t}$ for all item pairs $(i,j) \in I^2$, and assuming the independence, the probability of $p(>_{u,t} |\Theta)$

$$\prod_{u \in U} \prod_{B_t \in \mathcal{B}^u} \prod_{i \in B_t} \prod_{j \notin B_t} p(i >_{u,t} j|\Theta)$$

An equivalent way to express $p(i >_{u,t} j|\Theta)$ is :

$$\begin{aligned} p(i >_{u,t} j|\Theta) &= p(\hat{x}_{u,t,i} >_{\mathbb{R}} \hat{x}_{u,t,j} \,|\, \Theta) \\ &= p(\hat{x}_{u,t,i} - \hat{x}_{u,t,j} >_{\mathbb{R}} 0 \,|\, \Theta) \end{aligned}$$

We define $p(z > 0) := \sigma(z) = \frac{1}{1+e^{-z}}$:

$$p(i >_{u,t} j|\Theta) = \sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j})$$

Optimization Problem

Overall, the MAP-estimator becomes :

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

where λ_Θ is the regularization constant.

FPMC : Simpler Model

First let us assume,

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

LEMMA 1 (INVARIANCE OF (U,L) DECOMPOSITION). *For ranking of items and optimization with S-BPR, the FPMC model is invariant to the (U,L) decomposition, i.e.* $\hat{x}'$ *is invariant to* $\hat{x}$ *with:*

$$\hat{x}_{u,t,i} := \langle v_u^{U,I}, v_i^{I,U} \rangle + \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} \langle v_i^{I,L}, v_l^{L,I} \rangle$$

We have the result that :

$$\forall u, t, i, j : \ \hat{x}'_{u,t,i} - \hat{x}'_{u,t,j} = \hat{x}_{u,t,i} - \hat{x}_{u,t,j}$$

FPMC - A generalization

Setting $k_{U,I} = 0 \implies$ pure FMC model

$$\hat{x}^{\mathrm{FMC}}_{u,t,i} := \frac{1}{|B_{t-1}|} \sum_{l \in B_{t-1}} \langle v_i^{I,L}, v_l^{L,I} \rangle$$

Setting $k_{I,L} = 0 \implies$ pure MF model

$$\hat{x}^{\mathrm{MF}}_{u,t,i} = \langle v_u^{U,I}, v_i^{I,U} \rangle$$

CLearly, FPMC is linear combination of both the model:

$$\hat{x}^{\mathrm{FPMC}}_{u,t,i} = \hat{x}^{\mathrm{MF}}_{u,t,i} + \hat{x}^{\mathrm{FMC}}_{u,t,i}$$

Learning Algorithm: Stochastic gradient descent

$$\frac{\partial}{\partial\theta}\left(\ln\sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) - \lambda_\theta\,\theta^2\right)$$
$$= (1 - \sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}))\frac{\partial}{\partial\theta}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) - 2\,\lambda_\theta\,\theta$$

$$\frac{\partial}{\partial v^{U,I}_{u,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = v^{I,U}_{i,f} - v^{I,U}_{j,f}$$
$$\frac{\partial}{\partial v^{I,U}_{i,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = v^{U,I}_{u,f}$$
$$\frac{\partial}{\partial v^{I,U}_{j,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = -v^{U,I}_{u,f}$$
$$\frac{\partial}{\partial v^{L,I}_{l,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = \frac{1}{|B^u_{t-1}|}(v^{I,L}_{i,f} - v^{I,L}_{j,f})$$
$$\frac{\partial}{\partial v^{I,L}_{i,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = \frac{1}{|B^u_{t-1}|}\sum_{l \in B^u_{t-1}} v^{L,I}_{l,f}$$
$$\frac{\partial}{\partial v^{I,L}_{j,f}}(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}) = -\frac{1}{|B^u_{t-1}|}\sum_{l \in B^u_{t-1}} v^{L,I}_{l,f}$$

Stochastic gradient descent

procedure LearnSBPR-FPMC(S)
  draw $V^{U,I}, V^{I,U}, V^{I,L}, V^{L,I}$ from $N(0, \sigma^2)$
  repeat
    draw (u,t,i) uniformly from S
    draw j uniformly from $(I \setminus B_t^u)$
    $\delta \leftarrow (1 - \sigma(\hat{x}_{u,t,i} - \hat{x}_{u,t,j}))$
    for $f \in \{1, \dots, k_{U,I}\}$ **do**
      $v_{u,f}^{U,I} \leftarrow v_{u,f}^{U,I} + \alpha \left(\delta \, (v_{i,f}^{I,U} - v_{j,f}^{I,U}) - \lambda_{U,I} \, v_{u,f}^{U,I} \right)$
      $v_{i,f}^{I,U} \leftarrow v_{i,f}^{I,U} + \alpha \left(\delta \, v_{u,f}^{U,I} - \lambda_{I,U} \, v_{i,f}^{I,U} \right)$
      $v_{j,f}^{I,U} \leftarrow v_{j,f}^{I,U} + \alpha \left(-\delta \, v_{u,f}^{U,I} - \lambda_{I,U} \, v_{j,f}^{I,U} \right)$
    end for
    $\eta \leftarrow \frac{1}{|B_{t-1}^u|} \sum_{l \in B_{t-1}^u} v_{l,f}^{L,I}$
    for $f \in \{1, \dots, k_{I,L}\}$ **do**
      $v_{i,f}^{I,L} \leftarrow v_{i,f}^{I,L} + \alpha \left(\delta \, \eta - \lambda_{I,L} \, v_{i,f}^{I,L} \right)$
      $v_{j,f}^{I,L} \leftarrow v_{j,f}^{I,L} + \alpha \left(-\delta \, \eta - \lambda_{I,L} \, v_{j,f}^{I,L} \right)$
      for $l \in B_{t-1}^u$ **do**
        $v_{l,f}^{L,I} \leftarrow v_{l,f}^{L,I} + \alpha \left(\delta \, \frac{v_{i,f}^{I,L} - v_{j,f}^{I,L}}{|B_{t-1}^u|} - \lambda_{L,I} \, v_{l,f}^{L,I} \right)$
      end for
    end for
  until convergence
  return $V^{U,I}, V^{I,U}, V^{I,L}, V^{L,I}$
end procedure

Data set

The evaluation is performed on anonymized purchase data of online drug store. `http://www.rossmannversand.de`
The dataset is 10-core subset, i.e. every user bought atleast 10 items abd vice versa each item was bought by 10 users.

Table 2: Properties of the MC transition matrix estimated by the counting scheme. For the sparse dataset, only 12% of the entries of the transition matrix are non-zero and non-missing. For the dense subset, 88% are filled.

dataset	total	missing values	non-zero	zero
Drug store 10-core (sparse)	51,552,400 (100%)	1,041,100 (2.0%)	6,234,371 (12.1 %)	44,276,929 (85.9%)
Drug store (dense)	1,004.004 (100%)	0 (0.0%)	889,419 (88.6 %)	114,585 (11.4%)

Results

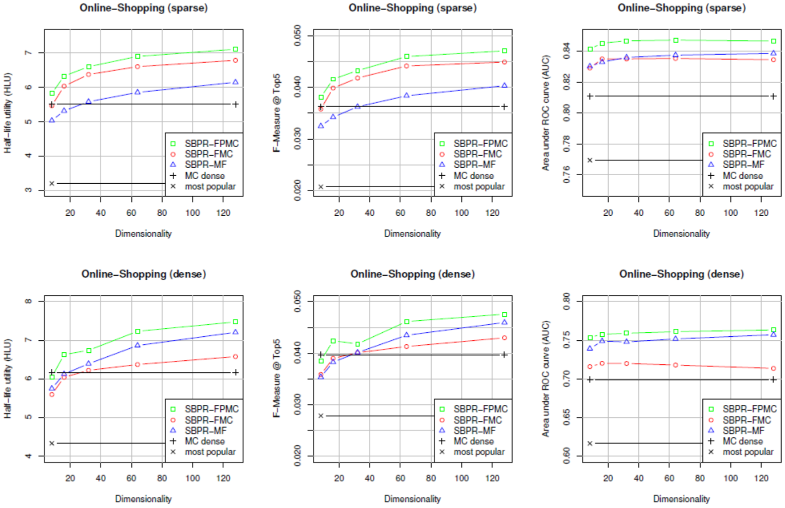


Figure 6: Comparison of factorized personalized Markov chains (FPMC) to a factorized Markov chain (FMC), matrix factorization (MF) [7], a standard dense Markov chain (MC dense) learned with Maximum Likelihood and the baseline 'most-popular'. The factorization dimensionality is increased from 8 to 128.

References

1. Steffen Rendle, Christoph Freudenthaler, Lars Schmidt-Thieme. Factorizing Personalized Markov Chains for Next-Basket Recommendation, in Proceedings of the 19th International World Wide Web Conference (WWW 2010), ACM.
2. S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), 2009.