Factorizing Markov Models for Categorical Time Series Prediction

Christoph Freudenthaler*, Steffen Rendle† and Lars Schmidt-Thieme*

*Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim, Germany
†Social Network Analysis, University of Konstanz, Germany

Abstract. During the last decade, recommender systems became a popular class of models for many commercial websites. One of the best state-of-the-art methods for recommender systems are Matrix and Tensor Factorization models. Besides, Markov Chain models are common for representing sequential data problems (e.g. categorical time series data). The item recommendation problem of recommender systems in fact is a categorical time series problem where each user represents an individual categorical time series. In this paper we combine factorization models with Markov Chain models. To increase efficiency of parameter estimation we introduce our generalized Factorized Markov Chain model.

Keywords: Categorical Time Series, Markov Chain, Factorization Models, Recommender Systems
PACS: 02.50.Fz Stochastic analysis

INTRODUCTION

Recommender systems are an important feature of modern websites. Especially, commercial websites benefit from a boost in customers loyalty, click-through rates and revenue when implementing recommender systems.

Typically data is collected by explicitly asking users for ratings of previously seen items (e.g. movies). Another way is to collect the data implicitly by examining logfiles. In this case only binary information (e.g. did the user select the movie or not) is available (Hu et al. [1]). We focus on the item recommendation problem with implicit data which is in fact a categorical time series problem, i.e. nominal predictors and a nominal target variable.

Markov Chains (MC) of order m are devised for modeling all sequential dynamics between categorical predictors and a categorical target variable (e.g. item prediction task with users and past items as categorical predictors) which makes them different to the time-aware models proposed by Koren [2] and Xiong et al. [3] who predicted ratings, i.e. real-valued target variables. To reduce overfitting while maintaining the expressiveness of an m-order MC this work adapts the Markov Chain model by replacing the standard maximum likelihood (ML) principle on independent transition probabilities by a factorization approach of the MC transition tensor (FMC). Extending FMC with personal information, i.e. allowing each user an individual transition tensor, leads to FPMC a personalized Markov Chain model learnt with factorized transition probabilities.

Although we focus on item recommendation our approach is more general: any categorical time series prediction problem can be modeled with our factorization approach. Replacing the item and user variables of the item prediction task by their appropriate substitutes is sufficient. With respect to our intended application area of recommender systems, scalability is another issue which we address with an efficient yet simple stochastic gradient descent algorithm.

Notation. Before describing our approach for efficiently learning, i.e. factorizing, personalized Markov Chains as a solution for categorical time series prediction problem, we introduce the notation of this paper which is in line with the sequential item recommendation scenario.

Let $U = \{u_1, \ldots, u_{|U|}\}$ be the set of all users and $I = \{i_1, \ldots, i_{|I|}\}$ the set of all items in the training sample S. For each user $u \in U$, a sequence of selected items $S_{u,t} = (i_t, i_{t-1}, \ldots, i_{t-m})$ is known for any point in time t the respective user actively selected an item[1]. The multiset of all sequences $S_{u,t}$ together is denoted by S.

The sequential item recommendation task can be formalized in creating a personal, dynamic relation $>_{u,t} \subseteq I^2$ which defines a ranking relation over all pairs of items for user u at time t. With this ranking, we can recommend known users their top n items.

[1] For the first m item selections where the user history is shorter than m we assume randomly missing values.

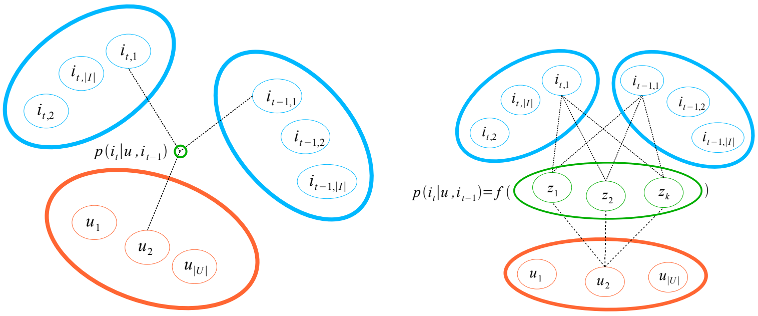

FIGURE 1. Left panel: a personalized Markov Chain of any order m (here: order 1) has $O(|U||I|^{m+1})$ many independent parameters, i.e. transition probabilities (here: only one parameter $p(i_t|u,i_{t-1}$ is depicted). Right panel: Factorized Markov Chain models decompose each relation using a function of latent variable representations $z_1,\ldots,z_k$ where each variable in the corresponding relation has its own set of $k \in \mathbb{N}^+$ latent feature representations.

Concerning factorization models the column vector $v_c = (v_{c,1},\ldots,v_{c,k})^T \in \mathbb{R}^k$ will denote the representation of the corresponding categorical variable $c \in C$ (e.g. $u \in U$ or $i_t, i_{t-1},\ldots,i_{t-m} \in I$) in the latent feature space $\mathbb{R}^k$. Moreover, we will use θ to denote all parameters of a discussed model, i.e. all transition probabilities of order-m Markov Chain models or all latent feature representations of factorized Markov Chain models.

FACTORIZING MARKOV CHAINS

Any categorical time series prediction problem especially sequential item prediction can be modeled with Markov Chain models. A Markov Chain model of order m is defined by its $m+1$-dimensional transition tensor

$$A_{i_t,\ldots,i_{t-m}} = p(i_t|i_{t-1},\ldots,i_{t-m}) \qquad \mathrm{i_t,\ldots,i_{t-m} \in I.} \tag{1}$$

Since personalization is important for recommender systems requesting personalized Markov Chains, i.e. one Markov Chain per user, leads to an additional dimension in the transition tensor:

$$A_{u,i_t,\ldots,i_{t-m}} = p(i_t|u,i_{t-1},\ldots,i_{t-m}),\ \mathrm{i_t,\ldots,i_{t-m} \in I, u \in U} \tag{2}$$

In general, each Markov Chain model is parametrized with one parameter per transition probability. This results in two major issues: first, the probabilities are estimated independently from each other, e.g. the two sets of observations used to estimate $p(i_t|u,i_{t-1},\ldots,i_{t-m})$ and $p(i_t|u^*,i_{t-1},\ldots,i_{t-m})$ are disjoint just because one state is different. Second, the number of model parameters is exponential in the number of items. In the case of recommender systems where U and I may have several thousands or even millions of different states this makes the parameter estimation intractable.

Figure 1 shows a graphical representation of an order 1 personalized Markov Chain: due to the independence assumption there exists an individual ternary relation for each triple of a user u, an item i_t and its predecessor i_{t-1} representing the independent transition probability. The extension to higher order Markov Chain models is obvious.

Our factorization approach relaxes the independence assumption. This is done by modeling each transition probability by latent features representations $v_c \in \mathbb{R}^k$ of its categorical variables c. Each observed predictor is mapped individually in the latent space $\mathbb{R}^k$ $k \in \mathbb{N}^k$.

$$p(i_t|u,i_{t-1},\ldots,i_{t-m}) := f(v_u, v_{i_t},\ldots,v_{i_{t-m}}). \tag{3}$$

By factorizing the parameters, the independence of the model parameters is removed because each transition probability shares the same latent features for every shared individual state. In contrast to the independence assumption which rules out any dependency between transition probabilities factorization models (eq. 3) include dependency as soon as at least one state is shared. This relaxation enables information sharing between previously independent transition probabilities via the jointly used k latent features. There exist several ways of learning the feature representation depending on the definition of the objective criterion and the latent feature model. For general categorical sequential prediction problems plugging any instance of (eq. 3) into the multinomial objective function and adding a 0-mean normal prior for the model parameters $p(\theta) \sim N(0, \frac{1}{\lambda_\theta})$ yields the general objective function for low-norm factorized

Markov Chain models. For the special case of item recommendation we apply a similar objective function introduced by Rendle et al. [4] which directly optimizes the latent feature representation θ for ranking. Generalizing BPR-opt to our categorical sequential prediction problem yields

$$p(\theta|S) := \prod_{u\in U}\prod_{i_t\in I}\prod_{i_t^*\in I}\prod_{i_{t-1}\in I}\cdots\prod_{i_{t-m}\in I} p(i_t > i_t^*|\theta)^n\ p(\theta) \quad n = |\{(r_1,r_2) : r_1 = (u,i_t,i_{t-1},\ldots)\in S,\ r_2 = (u,i_t^*,i_{t-1},\ldots)\notin S\}| \tag{4}$$

In our evaluation we use this ranking criterion for low-norm factorized Markov Chain models but bear in mind that the multinomial objective criterion is more appropriate in the general categorical time series prediction case.

Learning Markov Chain Models

Maximum Likelihood Approach

The solution for learning transition probabilities using Maximum Likelihood estimation is counting:

$$\hat{p}(i_t|u,i_{t-1},\ldots,i_{t-m}) = \frac{|\{(u,i_t,\ldots,i_{t-m}) \in S\}|}{|\{(u,i_{t-1},\ldots,i_{t-m}) \in S\}|}. \tag{5}$$

Problems of MLE. Equation 5 reveals that for estimating transition probabilities only the corresponding *full* relations of order $m+2$, and $m+1$ respectively, are counted whereas similar relations differing in at least one state are totally disregarded. The reason for the wasteful use of data is the aforementioned independence assumption, i.e. if at least one of the predictors in any pair of relations r_i and r_j is different the corresponding transition probabilities are independent from each other. Thus, observations of one transition are not used for inferring the probability of any other transition even if some of the predictors in both relations are the same. Neglecting this information requires the number of observations to grow linearly with the number of model parameters. This is infeasible since the number of model parameters is $O(|U||I|^{m+1})$ and typical application scenarios for recommender systems include hundreds of thousands of users and items.

Latent Factor Models

So far we have discussed different objective criteria and chose (eq. 4) for our ranking problem. Thus, the remaining task is to select a factorization model $f(v_u, v_{i_t}, \ldots, v_{i_{t-m}})$. Standard factorization models from the literature are the Tucker decomposition [5] or or the Canonical decomposition (CD) [6, 7]: The Tucker decomposition (TD) allows each categorical variable its own latent feature space with different numbers of latent features. CD unifies the latent feature spaces to one k-dimensional latent feature space $\mathbb{R}^k$ with mutually independent latent dimensions. The expressiveness of both decomposition models is the same since k can be increased as much as necessary. Recently, Rendle et al. [8] have proposed a pairwise interaction tensor factorization model (PITF), modeling entries of tensors as a sum of pairwise factorization models for each possible pair of dimensions, e.g user and selected item dimension. This gives[2]

$$f^{PITF}(v_u, v_{i_t}, \ldots, v_{i_{t-m}}) := \sum_{j>i}^{n}\sum_{f=1}^{k} v_{i,f} v_{j,f}. \tag{6}$$

This model is less expressive since it neglects all higher-order interactions except two-way interactions. However, we will use it following the strategy of Occam's razor, i.e. modeling lower order interactions first. MOreover, by merging the factorization model (eq. 6) with the ranking criterion BPR-opt we need way less parameters as we will see. Before our factorization model meets the BPR-opt criterion we need to define the ranking probability in terms of our factorization model:

$$p(i_t > i_t^*|\theta) = \sigma(f(r) - f(r^*)) \quad \text{with } S_{u,t} = (v_u, v_{i_t}, \ldots, v_{i_{t-m}}),\ S^*_{u,t} = (v_u, v_{i_t^*}, \ldots, v_{i_{t-m}}) \tag{7}$$

[2] To simplify notation we will substitute indices in the order $(i_t, u, i_{t-1}, \ldots, i_{t-m})$ by integers $(1, 2, \ldots, n)$ when necessary.

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function. For PITF, this has the advantage that all pairwise interactions vanish which do not represent an interaction with the target variable i_t.

For learning any of the factorization models scalable, fast, and easy to implement algorithm for learning θ are stochastic gradient descent algorithms. For the BPR-opt learning we adapted the sampling and gradient step such that for each iteration a sequence $S_{u,t} = (u, i_t, i_{t-1}, \ldots, i_{t-m}) \in S)$ together with a negative target state $i_t^* := (u, i_t^*, i_{t-1}, \ldots, i_{t-m}) \notin S$ is sampled. Using both the gradient is computed.

EVALUATION

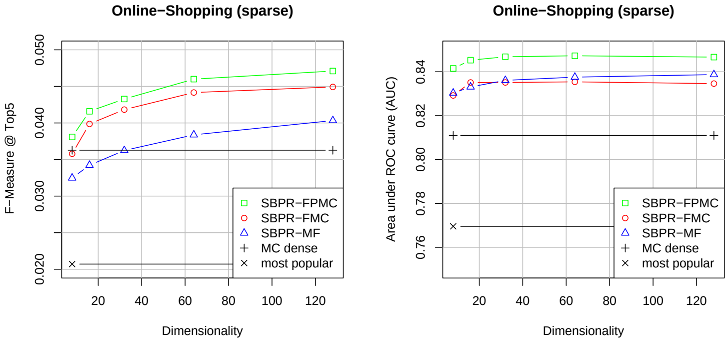

FIGURE 2. Comparison of factorized personalized Markov Chains (FPMC) to factorized Markov Chains (FMC), matrix factorization (MF), a standard dense Markov chain (MC dense) learned with Maximum Likelihood and the baseline 'most-popular'. The factorization dimensionality is increased from 8 to 128.

We empirically compare the recommender quality of our proposed factorized MC methods (factorized personalized Markov chain FPMC and factorized Markov chain FMC) to non-factorized Markov chain ('MC dense'), standard matrix factorization (MF) and the most-popular baseline (MP) – i.e. ranking all items by how often they have been bought in the past. For evaluation we use two different measures: AUC and f-measure. The runtime of model training linearly depends on the number of features. With our implementation, the training of each model took from a few hours up to 34 hours for the FPMC-128 model on the larger (sparse) dataset.

Dataset. We evaluate our recommender on the purchase data of an online shop of a drug store. The dataset we used is a 10-core subset, i.e. every user bought in total at least 10 items and vice versa each item was bought by at least 10 users. The dataset consists of $2,635,125$ purchases of $71,602$ users on $7,180$ items. Typically each user buys a set of items, i.e. a basket. The average basket size is 11.3 items and the average number of baskets per user is 3.2. For these data the fitted Markov Chain models of order 1. Since we have shopping baskets instead of single items, several items are bought simultaneously. This set-behavior can be easily modeled by our order-m MC models by dropping the sequential order assumption. We evaluated by splitting the dataset S into two non overlapping sets: a training set S_{train} and a testing set S_{test}. This split is done by putting the last basket for each user into S_{test} and the remaining ones into S_{train}. We removed those users from the evaluation that have bought less then 10 different items in the past (i.e. S_{train}). Secondly, for each user we removed all items from the test baskets (and the corresponding predictions) that this user has already bought in the past – this is because we want to recommend to the user items that are new/ unknown to him. Note that this makes the prediction task much harder and explains the low f-measure of all methods in figure 2. Otherwise just rerecommending already bought items would be a simple but very successful strategy for non-durable products in drug stores like toothbrushes or cleaner. The quality is measured for each user u on the first basket B_u in the test dataset.

Results. In figure 2 the quality of the different models is shown. For the factorization methods we run each method with $k_{U,I} = k_{I,L} \in \{8, 16, 32, 64, 128\}$ factorization dimension. With reasonable factorization dimensions ($k \geq 32$) all the factorization methods outperform the standard MC method. And in total, the factorized personalized MC (FPMC) outperforms all other methods.

REFERENCES

1. Y. Hu, Y. Koren, and C. Volinsky, “Collaborative Filtering for Implicit Feedback Datasets,” in *IEEE International Conference on Data Mining (ICDM 2008)*, 2008, pp. 263–272.
2. Y. Koren, “Collaborative filtering with temporal dynamics,” in *KDD ’09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, 2009, pp. 447–456, ISBN 978-1-60558-495-9.
3. L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell, “Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization,” in *Proceedings of SIAM Data Mining*, 2010.
4. 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.
5. L. Tucker, *Psychometrika* **31**, 279–311 (1966).
6. J. Carroll, and J. Chang, *Psychometrika* **35**, 283–319 (1970).
7. R. A. Harshman, *UCLA Working Papers in Phonetics* pp. 1–84 (1970).
8. S. Rendle, and L. Schmidt-Thieme, “Pairwise interaction tensor factorization for personalized tag recommendation,” in *WSDM ’10: Proceedings of the third ACM international conference on Web search and data mining*, ACM, New York, NY, USA, 2010, pp. 81–90, ISBN 978-1-60558-889-6.