Probabilistic Matrix Factorization

Mert Terzihan
Department of Computer Science
Brown University
Providence, RI 02912
mert_terzihan@brown.edu

Gabriel Barth-Maron
Department of Computer Science
Brown University
Providence, RI 02912
gabriel_barth-maron@brown.edu

1 Description of the Problem

Matrix Factorization is a widely used method, especially in collaborative filtering and recommendation systems. With the rise of data available to researchers and the requirement to present personalized ads or recommendations to users led the researchers to further study of this field. As our final project, we will be dealing with Probabilistic Matrix Factorization on matrices with discrete values. In particular we will apply these methods to analyze the movie-user review data in the MovieLens 100k dataset [6].

2 Discussion of Related Work

As studied in [2] and [3], one can observe that LDA is equivalent to factorizing a matrix probabistically using a graphical model representation. Instead of dealing with a co-occurence matrix, we will try to propose a model and a learning algorithm to factorize a matrix where each cell is a rating of an item by a particular user. Therefore, in addition to LDA, we would also like to generate ratings that have been assigned by users to each item.

In [1] a method based on LDA is proposed to produce personalized recommendations. However, fLDA takes into account much more information about the user, i.e. age, gender, zipcode, etc. We are going to use a simpler model, which relies on much less personal information. We will be using only the ratings that have been given by the users to specific items.

A LDA-based probabilistic matrix factorization models is proposed in [7], however instead of using discrete distribution for generating ratings, it has placed a Gaussian distribution over the user's ratings. In our investigation we hope to relax this assumption and look into models that are not necessarily Gaussian.

3 Graphical Model

Below is the graphical model and the details about it, where $j \in U$, $i \in M$ and $t \in T$, and U is the total number of users, M is the total number of items, and T is the number of topics that we would like to extract.

$$\begin{aligned}
&\theta_j \sim Dir(\alpha) \\
&z_{ji} \sim Cat(\theta_j) \\
&\phi_g \sim Dir(\beta) \\
&x_{ji} \sim Cat(\phi_{z_{ji}}) \\
&\kappa_{tj} \sim Dir(\gamma) \\
&r_{ji} \sim Cat(\kappa_{z_{ji},j})
\end{aligned}$$

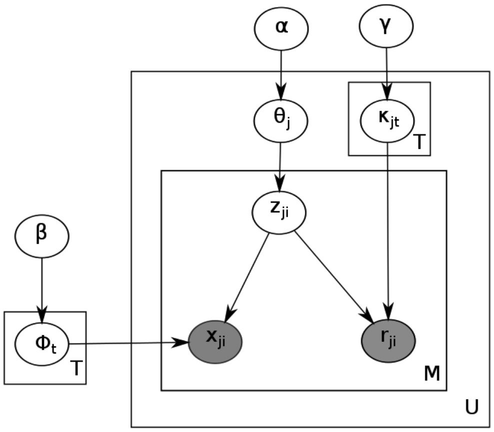

Figure 1: Representation of the Graphical Model

4 Preliminary Experiment

In our preliminary experiment we will be analyzing the data in the MovieLens 100k dataset [6]. This dataset contains 100,000 ratings from 1,000 users on 1,700 movies. It is a well studied dataset for which we should be able to obtain clean and comparable results. Our goal is to use LDA or a similar algorithm to discover latent genres among the movies. As mentioned earlier, since we will be using a generative model, we will also be able to generate ratings for movies that users have not yet rated.

[5]

5 Learning Algorithm

Because we are trying to discover hidden (latent) variables among incomplete data we will be using either MCMC or Variational Inference methods to perform learning. Generally speaking, our approach can be broken into two steps:

1. First complete the matrix and then factorize it
2. Operate on incomplete data likelihood as discussed in [4]

6 Evaluation

Because we will be using a generative model we can partition our dataset into training and testing parts. This will allow us to evaluate the reviews generated for movie-user pairs in the test set. We can use widely adopted metrics such as precision and recall to our model's effectiveness. In addition, because the MovieLens dataset is widely used we will be able to compare our performance against state of the art techniques.

Finally, to test the creation of the latent genres we can obtain expert defined genres for a subset of the movies in the dataset. We can then use a metric to analyze the distance between the predicted and expert defined genres.

7 Timeline

11/16 Models and datasets finalized and any additional related work identified.

11/21 Initial results for MovieLens dataset collected.

11/25 Comparisons of results to existing baselines completed for initial results.

12/5 Finished collecting results for model(s).

12/7 Finished evaluation and comparison to existing models.

12/8 Project presentation completed.

12/13 First draft of paper completed.

12/16 Paper completed.

References

[1] Deepak Agarwal and Bee-Chung Chen. flda: Matrix factorization through latent dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 91–100, New York, NY, USA, 2010. ACM.

[2] David M. Blei, Andrew Y. Ng, and Jordan I. Michael. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2:993–1022, 2003.

[3] Wray L. Buntine and Aleks Jakulin. Applying discrete PCA in data analysis. *CoRR*, abs/1207.4125, 2012.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal Of The Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[5] Anne-Marie Kermarrec and Afshin Moin. Data Visualization Via Collaborative Filtering. Research report, February 2012.

[6] GroupLens Research. Movielens 100k. `http://grouplens.org/datasets/movielens/`.

[7] Hanhuai Shan and Arindam Banerjee. Generalized probabilistic matrix factorizations for collaborative filtering. In Geoffrey I. Webb, Bing Liu 0001, Chengqi Zhang, Dimitrios Gunopulos, and Xindong Wu, editors, *ICDM*, pages 1025–1030. IEEE Computer Society, 2010.