Active Selection of Elements with Variational Bayes for Nonnegative Matrix Factorization

Gönül Aycı Abdullatif Köksal Melih Mutlu

Advisor: Prof. Ali Taylan Cemgil

GitHub

MOTIVATION

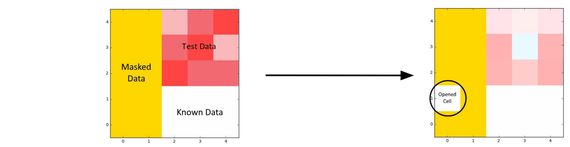
In most of the systems, collecting data is not always free. We propose an approach that learns distribution of data where information is **incomplete** or collecting it has a **cost**.

While discovering data using **Variational Bayes** for **NMF**, simultaneously we are collecting new information about data. By doing this, our main goal is **maximizing our knowledge** about data and **making the best estimation** by opening a minimum number of data.

PROBLEM

In this problem, the matrices are composed of *three different groups:* **known data** that is known and accessible at any time without any expense, **test data** that we are trying to predict, and **mask data** that is currently unknown but can be queried when desired.

When the data in the last group is queried, a cost arises. From this observation, we want to estimate the test data in the second group with the least error by making as few queries as possible.



- Redness of the cells correspond to RMSEs.

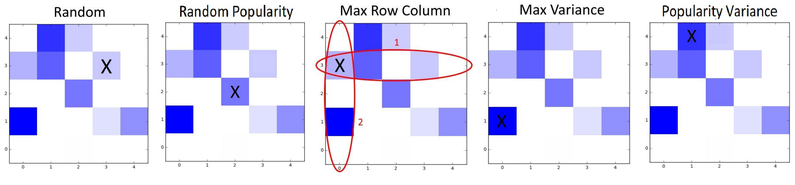
DATASET [4]

We used a popular dataset in our study which is MovieLens. The MovieLens dataset contains **100,000 ratings** on the scale of **0.5 to 5** from **671 customers** on **9066 movies**. In order to create a suitable input for the distributions that we used in our model, we rescale the ratings from **1 to 10**.

OBSERVATION SEQUENCE SELECTION STRATEGIES [1, 3]

In this project, we define six observation sequence selection methods.

- **Random**: Open a random cell in the mask matrix.
- **Random Popularity:** Choose the most popular movies, then open a cell randomly among those movies.
- **Max Row Column**: Choose the row that has the most masked cells. Then open the cell which is the most masked column index in this row.
- **Max Variance**: Open a cell which has maximum variance value.
- **Min Variance**: Open a cell which has minimum variance value.
- **Popularity Variance**: Choose the cell with maximum variance proportional to the square root of popularity of movies. This is a combined heuristic method.



- Blueness of the cells correspond to variances.

NMF (NON-NEGATIVE MATRIX FACT.) [2]

- NMF decomposes the data into two **low rank matrices: T** as the **template matrix** and **V** as the **excitation matrix.**
- The goal of NMF is to **minimize** the distance between *X* and *TV*. In this study, we use the popular **Kullback-Leibler (KL)** divergence.
- We use the **generative model** in order to describe the NMF process from a statistical perspective.
- You can see the **expectation** and **maximization** steps of the EM algorithm to calculate **MLE of T and V** which is equivalent to minimization of the information divergence.

$$X \approx TV$$

$n \times m \quad n \times r \quad r \times m$

CORPUS — SEMANTIC FEATURES — HIDDEN(SEMANTIC) VARIABLES

$$(T,V)^* = \arg\min_{T,V>0} D(X\|TV)$$

$$T \sim p(T \mid \Theta^t), \qquad V \sim p(V \mid \Theta^v),$$
$$s_{v,i,\tau} \sim \mathcal{PO}(s_{v,i,\tau}; t_{v,i} v_{i,\tau}), \qquad x_{v,\tau} = \sum_i s_{v,i,\tau}$$

E Step $\quad q(S)^{(n)} = p(S \mid X, T^{(n-1)}, V^{(n-1)}),$

M Step $\quad (T^{(n)}, V^{(n)}) = \arg\max_{T,V} \langle \log p(S, X \mid T, V) \rangle_{q(S)^{(n)}}$

VARIATIONAL INFERENCE [2]

- By assigning Gamma priors to the excitation and template, we created a more powerful hierarchical model.

$$t_{v,i} \sim \mathcal{G}\left(t_{v,i}; a^t_{v,i}, \frac{b^t_{v,i}}{a^t_{v,i}}\right), \qquad v_{i,\tau} \sim \mathcal{G}\left(v_{i,\tau}; a^v_{i,\tau}, \frac{b^v_{i,\tau}}{a^v_{i,\tau}}\right)$$

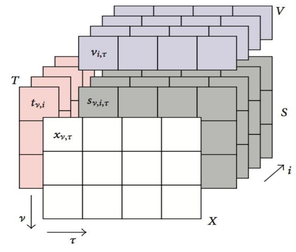
- We assume a factorized form for the in instrumental distribution as follows:

$$q(S,T,V) = q(S)q(T)q(V)$$
$$= \left(\prod_{v,\tau} q(s_{v,1:I,\tau})\right)\left(\prod_{v,i} q(t_{v,i})\right)\left(\prod_{i,\tau} q(v_{i,\tau})\right) \equiv \prod_{\alpha \in \mathcal{C}} q_\alpha$$

where $\alpha \in \mathcal{C} = \{\{S\},\{T\},\{V\}\}$.

- The definition of parameters for the Variational NMF as follows:

$$E_t = \langle t_{v,i} \rangle \quad L_t = \exp(\langle \log t_{v,i} \rangle) \quad \Sigma_t = \sum_\tau \langle s_{v,i,\tau} \rangle$$
$$A_t = a^t_{v,i} \quad B_t = b^t_{v,i} \quad \boldsymbol{\alpha}_t = \alpha^t_{v,i} \quad \boldsymbol{\beta}_t = \beta^t_{v,i}$$
$$E_v = \langle v_{i,\tau} \rangle \quad L_v = \exp(\langle \log v_{i,\tau} \rangle) \quad \Sigma_v = \sum_v \langle s_{v,i,\tau} \rangle$$
$$A_v = a^v_{i,\tau} \quad B_v = b^v_{i,\tau} \quad \boldsymbol{\alpha}_v = \alpha^v_{i,\tau} \quad \boldsymbol{\beta}_v = \beta^v_{i,\tau}$$



- After the definition of the parameters, we initialized latent variables before iterations.
- At each iteration, we calculate source sufficient statistics to update parameters of latent variables.
- We update latent variables to decrease KL divergence between prediction and actual data.

CONCLUSION and ACKNOWLEDGEMENT

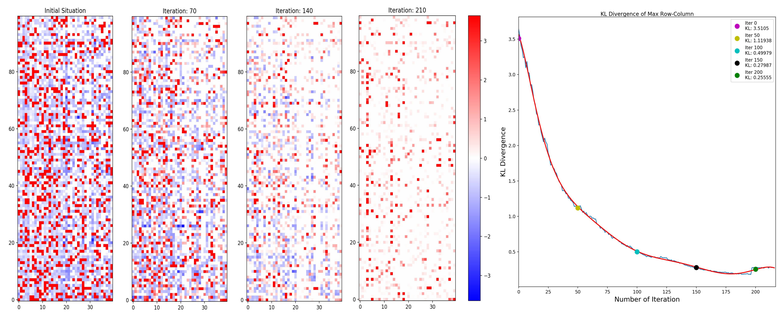
- In this study, we examined a hierarchical model with conjugate Gamma priors for Non-negative Matrix Factorization and we used Variational Bayes for inferences.
- We compared six strategies that we have defined over the Kullback-Leibler divergence and Root Mean Square Error metrics. We found that revealing cells with Maximum Row Column and Random Popularity strategies give maximum information about dense and sparse data, respectively.

- Special thanks to **Prof. A. Taylan Cemgil** and **Burak Suyunu** for their endless guidance and support.

RESULTS

We test our approach by generating synthetic data from the hierarchical model with *W = 100, K = 40,* and *I = 4.*

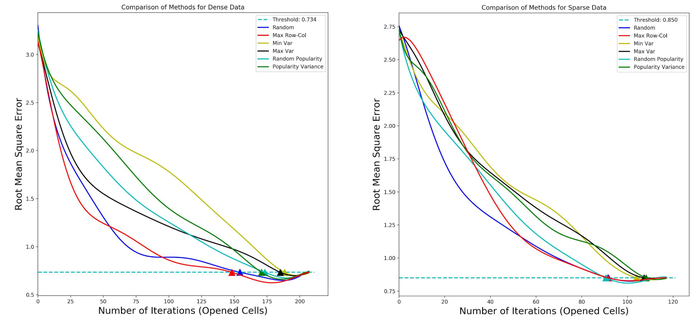
In the following figure, we have shown the results for **dense data** with **Max Row Column** cell opening strategy based on **KL divergence**. We divide our data matrix into three parts that *69%* of test, *30%* of mask, and *1%* of known out of 4000. We run Variational Inference for MAXITER = 10,000 steps following a burn-in period of 5000 steps.



- Redness and Blueness of the cells correspond to KLs and variances, respectively.

We compare our methods based on both **Kullback-Leibler** (KL) and **Root Mean Square Error** (RMSE). As we show in the following results, **Max Row Column** and **Random Popularity** are the best methods that converge to the threshold faster based on RMSE for dense and sparse data, respectively.

We have created two different experimental setups: dense and sparse. For dense setup, we used 100 of the most watched movies from MovieLens and 40 of the most watched users. For sparse setup, we selected 100 out of the 300 most watched movies, and 40 out of the 120 most watched users, randomly.



REFERENCES

1. Suyunu, Burak, Aycı, Gönül, and Cemgil, A. Taylan. "Active selection of elements for Bayesian nonnegative matrix factorization." *2018 26th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2018.
2. Cemgil, Ali Taylan. "Bayesian inference for nonnegative matrix factorisation models." *Computational Intelligence and Neuroscience* 2009 (2009).
3. Elahi, M., Ricci F., and Rubens N. "A survey of active learning in collaborative filtering recommender systems." *Computer Science Review* 20 (2016): 29-50.
4. Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." *Acm transactions on interactive intelligent systems (tiis)* 5.4 (2016): 19.