# Personalized Recommendations for Customers in E-commerce Platform Taobao by Bayesian Method

Team members: Yiwei Mai, Yizhou Zhang, Baihan Liu

**Background and Goal**

In today's digital era, e-commerce platforms like Taobao play an integral role in shaping consumer behavior and purchasing decisions. With millions of items available online and an overwhelming amount of choices, users often rely on personalized recommendations to navigate and select products. However, the challenge arises when these recommendations are based on implicit feedback – subtle, indirect signals of user preferences, rather than direct ratings and reviews.

The overarching goal of this endeavor is to design and implement a robust recommendation system capable of accurately predicting and suggesting items to users based on their historical behaviors. Such a system would not only enhance the user experience by presenting relevant products but also drive sales and user engagement for e-commerce platforms like Taobao. By harnessing the power of this dataset and leveraging advanced Bayesian analysis techniques, we aim to bridge the gap between user preferences and the vast digital marketplace, making online shopping a more tailored and intuitive experience.

**Data Description**

We found the data on a data science platform called Tianchi. (https://tianchi.aliyun.com/dataset/649) The data was released officially by Taocao, an e-commerce company in China. This is a real-world dataset, and the sample population is about a million users which were randomly sampled from the total users who had behaviors on the platform between 2017-11-25 and 2017-12-03.

There are some potential defects: firstly, the data does not contain any user demographic information, which means we cannot check the authenticity of randomization and also further grouping users based on their demographic. Secondly, the time span of the data is only 8 days, which is not wide enough. The problems are that firstly some seasonal and periodic issues may exist. For example, purchases of clothes are influenced by seasons; customers' shopping willingness may be weaker after a sales promotion campaign. Secondly, some user behaviors may happen outside of the time span mentioned above. For example, a user may view an item page in 2017-12-03 and buy the item after 2017-12-03, but in this data we cannot see that. However, such a problem may be negligible under a very large sample size.

Therefore, if we are able to collect the data, we would like to collect users' demographic information and user behaviors across a wider time span, for example, at least one year.

The data contains around a million users, 4.16 million of items and a hundred million of their online shopping behaviors, including pv(page view), buy, cart (adding an item to the cart) and fav(favorite an item). Besides, we also have user ID, item ID, item categories, timestamp of each behavior.

Although data collection processes are not included in the analysis, data transformation processes will be conducted to pivot the data into a form like: user ID, item ID, behaviors or even behavior chains further.

**Collaboration Plan**

Our regulating meeting time is 4-5pm every Wednesday.

Baihan Liu organized some meetings and wrote the problem of interest and the final goal of the problem in the proposal. In the future, her job is to do some exploratory data analysis about the data, explore other Bayesian approaches, as well as complete the report and presentation of her work.

Yizhou Zhang finished the content of the data analysis plan including statistical model description, timeline, and potential difficulties in the proposal. In the future, his job is to train the BPR model and conduct analysis and visualization of the result, and complete the report and presentation of his work.

Yiwei Mai found the data and wrote the introduction and description of the data, including data background, potential defects of the data, and facts of the data. In the future, his job is to do some background investigation, implement some baseline models and maintain the repository of this project.

We discuss the project proposal and share our ideas together in several meetings.

**Plan for data analysis**

We aim to build a recommendation system to analyze the data. Bayesian Personalized Ranking (BPR) is a bayesian algorithm frequently used in building a recommendation system. It uses matrix factorization which factorizes the user-item interaction matrix into two lower dimensional rectangular matrices. We may regard two matrices as a user matrix where each row corresponds to a user's latent factors and an item matrix where each column corresponds to an item's latent factors. The multiplication of two matrices

will give us the ranking of each item for each user. Bayesian is applied in modeling the parameters in two matrices. We start by assigning a prior distribution for two matrices, we add users' preference data and apply Bayes' Rule to get posterior distribution for two matrices.

Our timeline for this project is:
Nov 8th: Data preprocessing and exploratory data analysis
Nov 18th: Fit a preliminary model by BPR and analyze how to improve
Nov 28th: Complete training BPR model and output its performance. (May add non-bayesian model for comparison)
Dec 8th: Summarize all data analysis results, write the final report and prepare slides for presentation.

Potential Difficulties:
It might be hard to visualize the recommendation system.
The huge size of data may burden the data analysis work and slow down the modeling process. We may consider dropping a given proportion of data without hurting the generalizability.