Homework Report

Rihan Chen

Thursday, July 21, 2016

Introduction and data description

This problem can be taken as a recommender system problem. The rating matrix is initialized by the users' browsing history. User's features include: gender and home continient. Hotel's feature includes their star_rating.

Methods

In this project, I consider two methods: collaborative filtering and factorization machine. For both methods, I try to include as many features as possible.

For collaborative filtering method, I choose the most similar users mainly based on two criteria. The first is that the number of intersected hotels between different users. The second is the weights based on the user features. After I choose the most similar users, I calculated the weighted frequency of hotels that are considered by these most similar users. For example, for user 1, it is similar with user 2, user 2 have aleady looked at hotel 1, 2, 3. The weight between user 1 and user 2 is 1.2 (calculated by the similarity between their genders and home continents). Then when calculating the total weighted frequency of hotel 1, 2, 3, user 2 will contribute 1*1.2 to each hotel. Finally, if two hotels have the same weighted frequency, the hotel with higher star_rating will be recommended to the user. The drawback for this method is that it is very time-consuming to get all kinds of dictionaries for looks-up.

For factorization machine model, I build the model through Markov Chain Monte Carlo method. The reason is that MCMC method have fewer hyper-parameters, which can greatly save the time to tune the model. The drawback is that MCMC solver often take more time to make a prediction. For saving the time, I only use the train and test split method to tune the hyperparameter. However, the cross-validation should be taken in the future for more reasonable result. For each user, I got the scores for every hotel through FM model. Then euclidean distances between scores and 1 are calculated. I choose the hotel that have the minimum euclidean distance.

Discussion

The reason for considering two methods is that I cannot decide which method is better for this project. For collaborative filtering, it is straightforward and understandable. But the drawback is that the weights and the variants are simply decided by my intuition. For factorization machine, it is more advanced and reasonable for including user features and hotel feature. However, the problem is that the rating matrix is initialized by the users' browsing history. So the entries of the matix should be taken as the indicators instead of ratings. But I can only build a regression model for factorization machine through l2 loss, which is not reasonable somehow.