535 Semester Long Project

Xianzhi Luo B00814692

1 Brief solution

In this final project, we are requested to build a recommender system. Generally, there are two types of recommender systems: Context filtering based recommender system and collaborative filtering based recommender system. Because our dataset is a dataset include user id, item id and rating value, I choose memory based collaborative filtering to build this recommender system.

The first thing to do is data preprocess. First step is to split the training and test data set and do the summarization. I use the pandas library to get the mean, total number of users’ rating, total value of users’ rating. Then I calculate the items that have been rated and fill them with default value. This step is to solve cold start issue.

To predict the rating R on item I given by user U, there are three main tasks to do. First, find all similar users that have rated on item I by cosine similarity (or other similarity algorithm). Second, I choose top 100 of these similar users and use the sum of (weights times mean rating) to get the predict rating Rp. Third, build and train the dataset by KNNwithmean and calculate the RMSE. After training and model selection, I get the optimal model for this dataset and use it to predict remain 0 values in the dataset. I also compare KNNwithmean with other algorithms or models such as Co-Clustering, SVD and so on.

2 Cold Start Issue

In collaborative filtering, there is a big issue which is called cold start problem. And in cold start, there are user cold start problem and item cold start problem which means a new user or a new item comes in, there is no relation you can use to predict this newcomer. In this dataset, among 943 users, the user with least rating number (the number of items rated by one user) is 3 while the mean of rating number is 63. Among 1628 items, there are 65 items haven’t been rated. What I do is using representative based Matrix Factorization to solve this cold start. The RBMF menas an additional constraint that m items should be represented by a linear combination of k items. So, I choose 100 most popular users by their rating number, with user 517 who has rated 517 items and so on. For those 65 items haven’t been rated, I randomly choose rating number by quantile 0.25 which is 18, mean which is 37, quantile 0.75 which is 84 as their rating number. And I choose this number from 100 most popular users, and the rating on the 65 items will be the mean of these most popular users.

3 Two real-world examples

Document-term matrix: A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. Each entry corresponds to the number of times the associated term appears in the indicated document.

Geographic Triangulation: Suppose we are given partial information about the distances between objects and would like to reconstruct the low-dimensional geometry describing their locations. For example, we may have a network of low-power wirelessly networked sensors scattered randomly across a region. We can only get a partially observed distance matrix, the row can be the place of departure and the column can be the destination. Each entry corresponds to the distance between departure and destination.

4 Mini survey