HOMEWORK 2

Jie Tang

OCTOBER 7, 2020

Q1. Write a program to implement the following recommendation approaches. Test it on MovieLens-100K and calculate the RMSE.

a) User-User [15 marks]

b) Item-Item [15 marks]

c) Latent Factor Model [20 marks]

Please use “ua.base” for training, and “ua.test” for testing

a) The main steps of User-User approach are :

i) Consider user x
ii) Find set N of other users whose ratings are “similar” to x’s ratings
iii) Estimate x’s ratings based on ratings of users in N

In addition, we need to know the methods of determining “similar”. There are many measures, such as Jaccard Similarity, Cosine Similarity, Pearson Correlation Coefficient. I chose Cosine Similarity.
Cosine Similarity calculation:

$$\boldsymbol{sim(x, y)} = \frac{\sum_i \boldsymbol{r_{xi} \cdot r_{yi}}}{\sqrt{\sum_i \boldsymbol{r_{xi}^2}} \cdot \sqrt{\sum_i \boldsymbol{r_{yi}^2}}}$$

r_{xi}: user x’s rating on item i

For the prediction: we can use average of N other users, or weighted average of N other users’ rating on item i. Here I chose weighted average.

The RMSE I got is 10.1, which is quite high. I think there are several reasons: 1. The problem with the dataset, since the max user_id in train data is 1682, while in test data, the max user_id is 1680; 2. The User-User is quite simple, and more factors should take into consideration such as time, age.

b) The main steps of Item-Item approach are :

i) For item i, find other similar items
ii) Estimate rating for item i based on ratings for similar items
iii) Can use same similarity metrics and prediction functions as in user-user model

The RMSE I got is 13.0, which is quite high as well. I think the reasons should be similar like User-User scenario.

$$\boxed{r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}}$$

s*ij*… similarity of items ***i*** and ***j***
r*xj*…rating of user ***u*** on item ***j***
N(i;x)… set items rated by ***x*** similar to ***i***

c) The main idea of Latent Factor Model derives from collaborative filtering:

Problems/Issues:

1) Similarity measures are “arbitrary”

2) Pairwise similarities neglect interdependencies among users

3) Taking a weighted average can be restricting

Solution: Instead of sij use wij that we estimate directly from data

- **Idea:** **Let’s set values *w* such that they work well on known (user, item) ratings**
- **How to find such values *w*?**
- **Idea:** Define an objective function and solve the optimization problem

- Find $\boldsymbol{w_{ij}}$ that minimize **SSE** on **training data**!

$$J(w) = \sum_{x,i} \Big(\underbrace{\Big[b_{xi} + \sum_{j \in N(i;x)} w_{ij} \left(r_{xj} - b_{xj} \right) \Big]}_{\text{Predicted rating}} - \underbrace{r_{xi}}_{\substack{\text{True} \\ \text{rating}}} \Big)^2$$

- Think of ***w*** as a vector of numbers

The RMSE I got is 3.24, which is high but much better that the above two algorithm. I think the reasons are I consider more factor in the latent factor model, this improve the accuracy.

Notes: All the codes are in another file.