Guangdong Miao, z5094104

Recommender system using collaborative filtering
COMP9417 Machine Learning Project

Introduction
As people's reliance on the Internet has increased, the Internet has never disappointed humans. Gradually, the Internet can also understand its interests and predict its preferences by analysing users who have similar operations with certain users. In the field of information technology, it is called recommender system. In this report, the method of collaborative filtering is applied to address a user-based recommender system. The purpose of this study is to predict someone's rating for a movie and analyse the influence of the choice of the number of nearest neighbours on the forecast result.

Methodology
- Processing data

The dataset is from https://grouplens.org/datasets/movielens/100k/, 'u.data' is chosen to be the target dataset after unzipping mk-100k.zip. This file contains 4 columns represent user id, movie id, ratings and an irrelevant item. Then, we make an dictionary to store all the book ratings for each user. The dictionary looks like {user:{book: rate,…},…}. It means all the pairs of book and rating the user rates are combined.

- Pearson correlation coefficient

Pearson correlation coefficient is used to determine the similarity and relationship between two users. For two different users x and y, if they rate the same movie i, Xi is score user x rates for movie i while Yi is score user y rates for movie i. Average(X) is the average score user x rates while Average(Y) is the average score user y rates. Now we get the Pearson correlation coefficient **r** between two different users. For the positive **r**, the greater **r**, the stronger positive relevance. For the negative **r**, the smaller **r**, the stronger negative relevance. If r = 0, it means x does not have relationship with y. The formula is shown below to compute **r**.

$$r = \frac{\Sigma(X_i - \bar{X})\,(Y_i - \bar{Y})}{[\Sigma(X_i - \bar{X})^2\,\Sigma(Y_i - \bar{Y})^2]^{1/2}}.$$

- K nearest neighbours

For a given user to be predicted, computing all the **r**s between given user and all the other users. Delete zero values and sorting r list in descending order. This is all the neighbours list and the first k elements is k nearest neighbours.

- Prediction

Pearson correlation coefficient can be seen as weight for the rate from the same movie which two different users rate for. Represent s as score another people rate for the same movie with user to be predicted. Prediction can be achieved in below formula.

$$\text{Prediction} = \sum(\text{weight} * s)/\sum(s)$$

- Accuracy

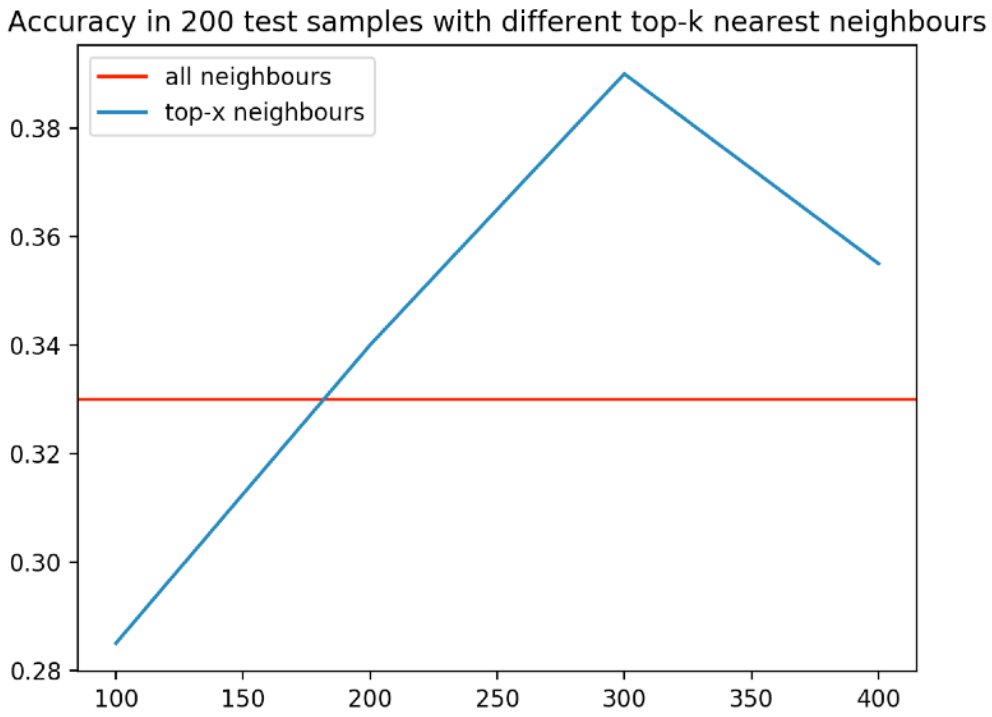
In this study, dataset is split as 80% training data and 20% test data. First, training data with all neighbours, use 200 sample tests to predict socres. Comparing the result with ratings from test and then compute the accuracy. Different number of nearest neighbours (KNN) result in different accuracy.

Result

For user 1, when we predict rate of movie 20, this system get 4.

```
Please input a user id you want to predict: 1
Please input a movie id you want to predict: 20
The user 1 and movie 20 contribute to 4 rating
```

However, this result is not always correct due to the problem of overfitting. Now I pick out 200 sample tests from test dataset to test this model. Through changing the number of nearest neighbours, we can get a series of predictions. The figure below shows the relationship between accuracy and k value. This seems to suggest that not the greater the value of k, the higher the accuracy. In the same way, the accuracy will reach a maximum at a threshold, reducing or increasing this threshold will bring about a decrease in accuracy.



Challenge

The study only uses smaller test cases, but it usually involves larger databases in real mission scenarios, which requires high scalability of the algorithm. Not only can we process data optimally, we can also use clustered servers to handle larger databases. These are all possible future work I should properly be working on to improve.

Conclusion

The purpose of this study is to achieve a system of movie scoring prediction, and study the influence of the choice of k value on the accuracy of the system. Through computing the pearson correlation coefficient, we can get nearest neighbours. Then use coefficients to calculate final ratings. Further, corresponding predictions would be generated by changing number of nearest neighbours. Through the analysis of the relation graph, you can get the connection of k value and precision. Not that the larger the value of k is, the higher the accuracy is.

Reference

[1]Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl, GroupLens: an open architecture for collaborative filtering of netnews, Computer Supported Cooperative Work, pp175-186, Chapel Hill, North Carolina, 1994.

[2]Rogers and Nicewander (1988). "Thirteen Ways to Look at the Correlation Coefficient" (PDF). *The American Statistician*. **42** (1): 59–66. doi:10.2307/2685263.