STAT 578 Spring 2018 Homework 1

Jifu Zhao Date: 02/18/2018

(a)

The given dataset is Movielens 100K, which contains 943 users and 1682 movies in total. The given dataset has been splitted into 5 folders, and forms 5 different training and test set, named (u1_base, u1_test), (u2_base, u2_test), (u3_base, u3_test), (u4_base, u4_test), (u5_base, u5_test). In addition to the userID, itemID, rating information, we also have the information about each user as well as each item.

For the first question, I explored two models. The first model is simply linear regression. In this model, we treat each user and each item's information as the predictor, including age, gender, item category and so on, train the model using the training set, and make predictions on the test set.

In addition to the linear regression model, we also implement KNN model using the Surprise package, which is a Python package for building and analyzing recommender systems. We choose the item-based KNN algorithm.

The performance of two algorithms is shown in Table 1

Table 1: Summary of Cross Validation

RMSE/Model	Linear Regression	KNN
CV 1	1.13332	1.15368
CV 2	1.11352	1.13066
CV 3	1.09645	1.11158
CV 4	1.09598	1.11329
CV 5	1.09997	1.11868
Mean	1.10785	1.12558
Std	0.01424	0.01556

(b)

In this part, we implement a series of visualization tools to infer the behaviour of the missing ratings.

Based on the user, item, and rating information, we can construct the utility matrix, where rows correspond to users, columns correspond to items, and the colors correspond to user's rating on specific item. For the missing ratings, we fill it with 0. The utility matrix is shown in Figure 1.

From Figure 1, we can see that, the missing ratings are not occur at random. For movies that have high ratings, there are less missing values. On the other hand, for movies that have relative low ratings, there are more missing values.

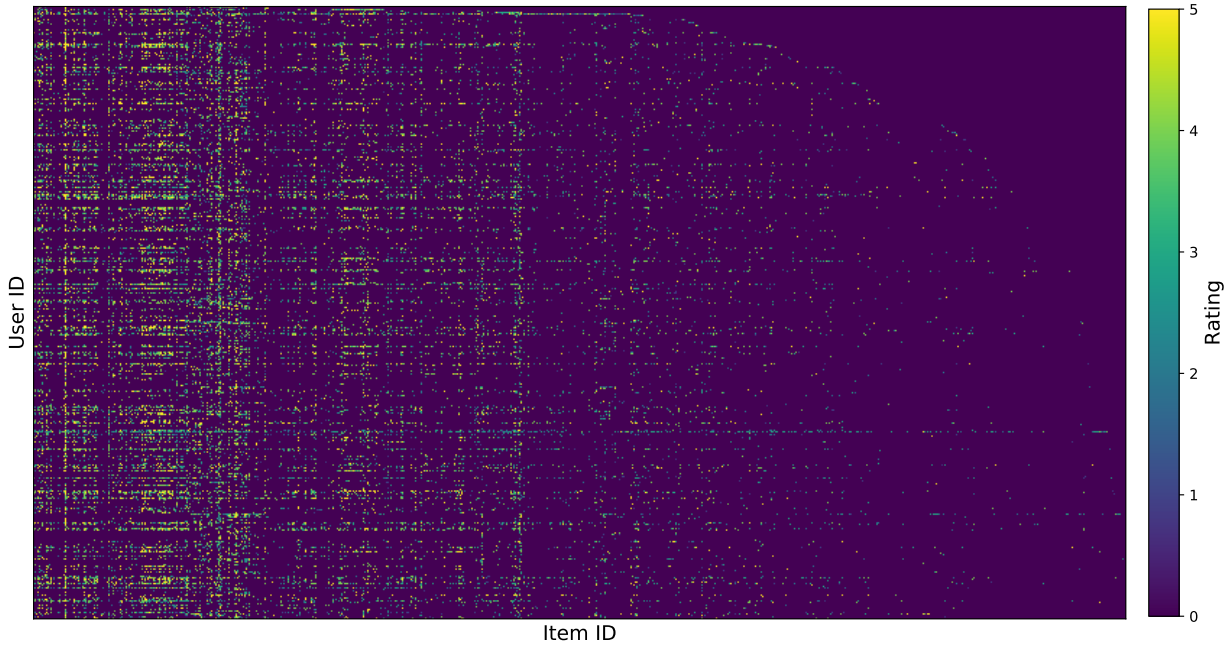
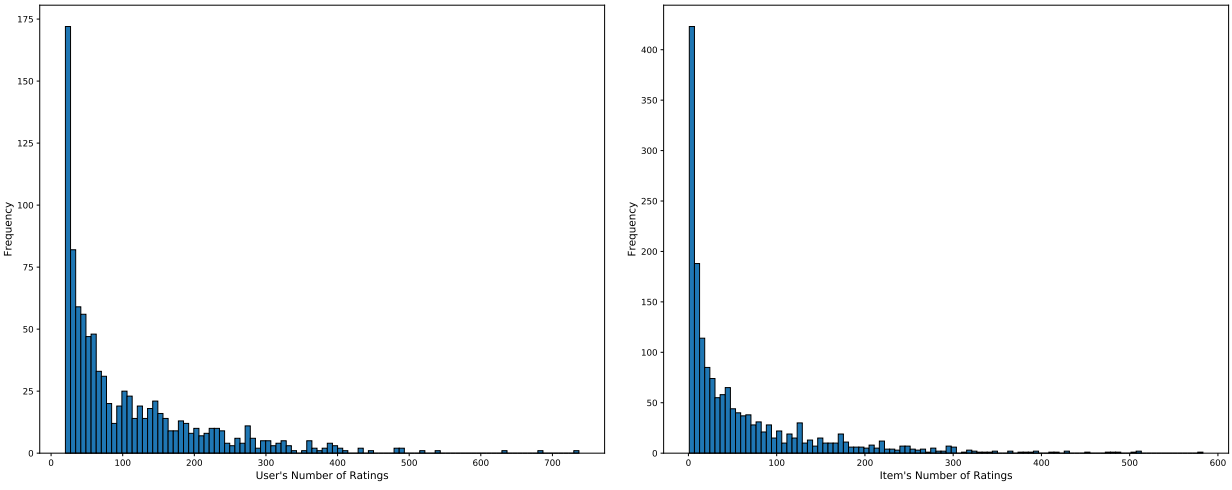


Figure 1: User-item utility matrix



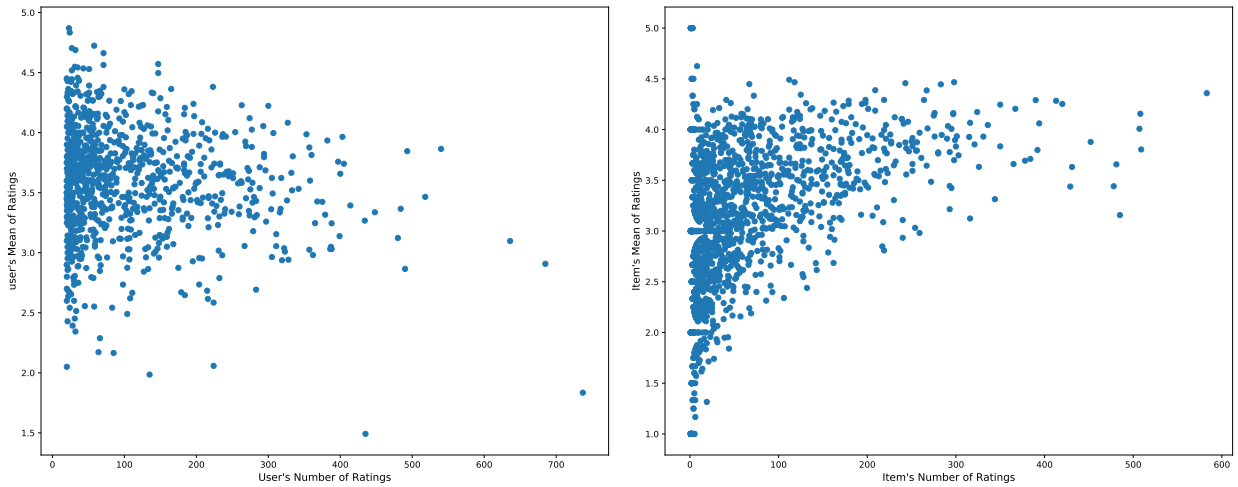
(a) Distribution of user's number of ratings

(b) Distribution of item's number of ratings

Figure 2: Distribution of User's / Item's number of ratings

To see more details, we plot the distribution of user's and item's number of ratings in Figure 2. As shown clearly, most users and most movies have low number of ratings. The distributions show long tails for high number of ratings.

In addition to the distribution of number of ratings, we also plot the relation of user's or item's mean rating versus the number of ratings. As shown in Figure 3, for users, their mean rating are mainly normally distributed. On the other hand, for movies, the movies who have relatively higher number of ratings, tend to have relative higher mean rating.



(a) User's mean rating vs. number of ratings

(b) Item's mean rating vs. number of ratings

Figure 3: Mean ratings vs. number of ratings

(c)

Based on our analysis in Part (c), we can notice one clear phenomenon, for movies with relatively larger number of ratings, there tend to have relatively higher mean ratings. Based on this, we can use the number of ratings for each movie to indicate its popularity. In Figure 4, we fit a linear curve for mean rating and number of ratings.

To improve the performance of the prediction, we re-build the linear regression model and the item-based KNN model. This time, for linear regression model, we add another feature corresponding to popularity (in actual code, we use the fitted mean rating to represent the popularity, instead of using the raw number of ratings). For KNN model, in addition to the prediction from KNN model, we choose the average value of KNN's prediction and the mean value from fitted curve in Figure 4 as the final prediction. All the results are shown in Table 2.

As shown clearly in Table 2, compared with Table 1, both linear regression model and the KNN model give better results. More specifically, for linear regression model, the mean RMSE gets improved from 1.10785 to 1.08252, and for KNN model, the RMSE gets improved from 1.12558 to 1.09853.

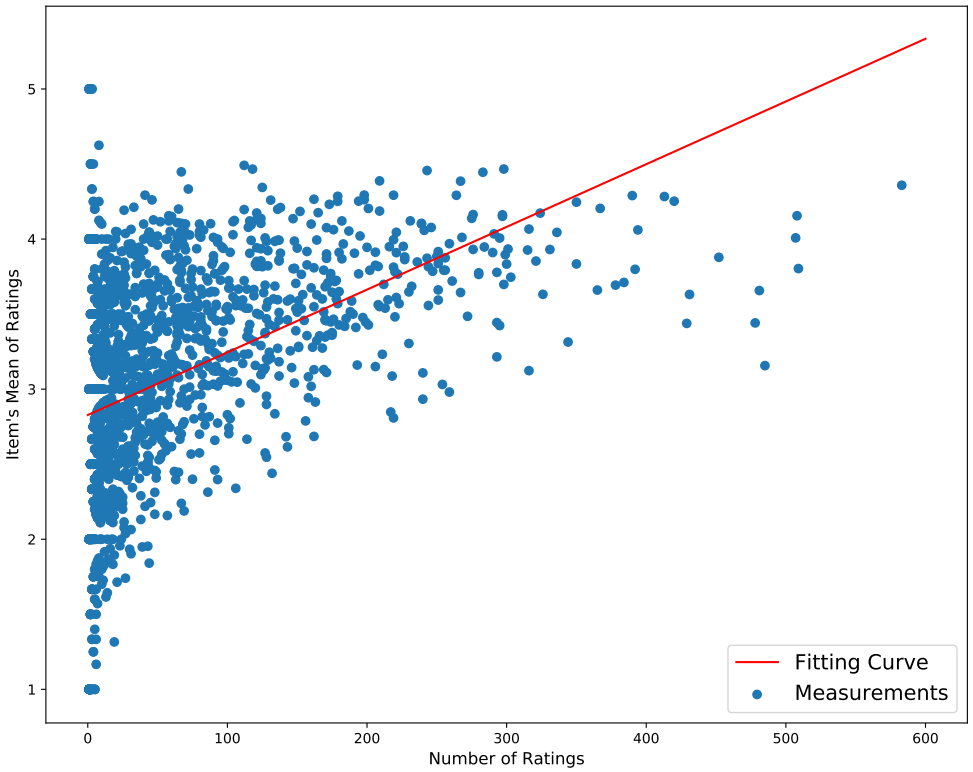


Figure 4: Item's mean of ratings vs. number of ratings

Table 2: Summary of Cross Validation

RMSE/Model	Linear Regression	KNN
CV 1	1.10348	1.12234
CV 2	1.08674	1.10264
CV 3	1.07505	1.08820
CV 4	1.07059	1.08572
CV 5	1.07674	1.09375
Mean	1.08252	1.09853
Std	0.01173	0.01325