EECS 349 Machine Learning

Movie Recommendation System

Ding Xiang

Oct. 21, 2017

Problem 1

A) After going through each pair of users, the mean number of movies two people have reviewed in common is 18. The median number is 10. The histogram plot is shown below.

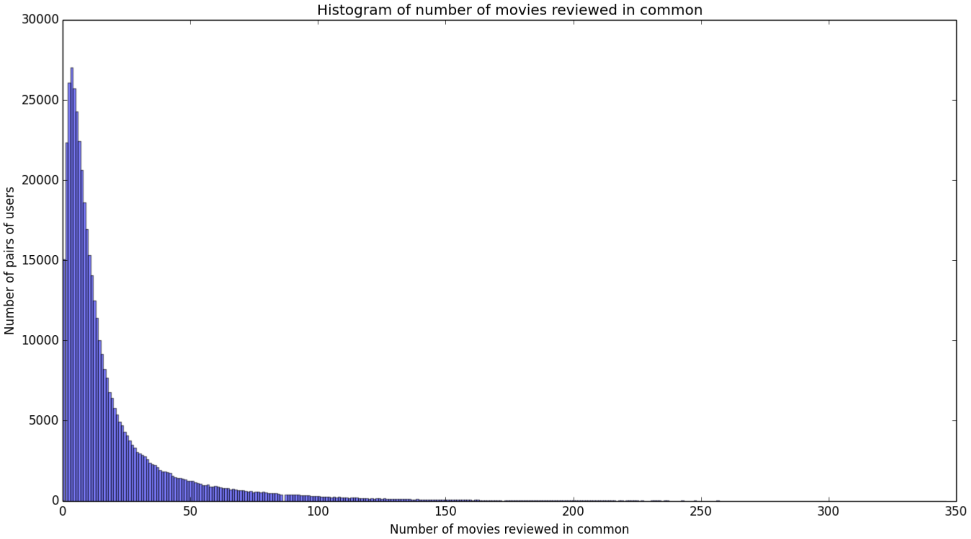


Figure 1. Histogram of number of movies reviewed in Common

In plotting this histogram, I choose the bin to be 350 since the largest number of movies a pair of users reviewed in common is 346. So this histogram can clearly reflect the trend of the data distribution.

B) After going through movies, the 50th movie had the most review. It had 583 reviews. The movies that have fewest review only get 1 review and they are 599, 677, 711, 814, 830, 852, 857, 1122, 1130, 1156, 1201, 1235, 1236, 1309, 1310, 1320, 1325, 1329, 1339, 1340, 1341, 1343, 1348, 1349, 1352, 1363, 1364, 1366, 1373, 1414, 1447, 1452, 1453, 1457, 1458, 1460, 1461, 1476, 1482, 1486, 1492, 1493, 1494, 1498, 1505, 1507, 1510, 1515, 1520, 1525, 1526, 1533, 1536, 1543, 1546, 1548, 1557, 1559, 1561, 1562, 1563, 1564, 1565, 1566, 1567, 1568, 1569, 1570, 1571, 1572, 1574, 1575, 1576, 1577, 1579, 1580, 1581, 1582, 1583, 1584, 1586, 1587, 1593, 1595, 1596, 1599, 1601, 1603, 1604, 1606, 1613, 1614, 1616, 1618, 1619, 1621, 1624, 1625, 1626, 1627, 1630, 1632, 1633, 1634, 1635, 1636, 1637, 1638, 1640, 1641, 1645, 1647, 1648, 1649, 1650, 1651, 1653, 1654, 1655, 1657, 1659, 1660, 1661, 1663, 1665, 1666, 1667, 1668, 1669, 1670, 1671, 1673, 1674, 1675, 1676, 1677, 1678, 1679, 1680, 1681, 1682.
The plot is shown below.

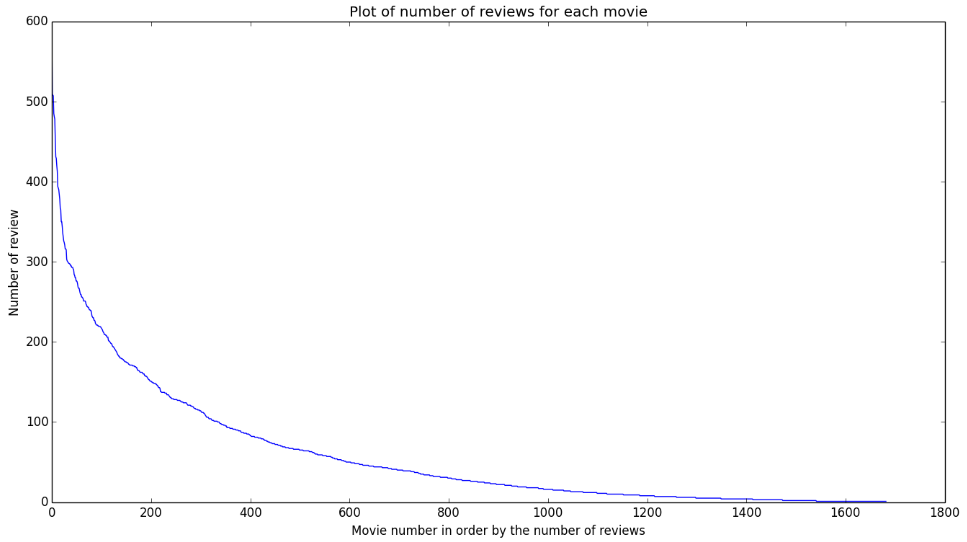


Figure 2. Number of reviews for each movie

From the figure above, I think it follows the Zipf's law because the curve looks like the Zipfian distribution.

C) Prediction is -779.14, when x=3. This is the "best" result the machine can get based on polynomial regression learning from training data with 6-fold validation. But it's not quite reliable and convincing because the training data is too different from the predicted data, as we can see that training data is within [-1, 1] while predicted data is 3. So, in order to make a good prediction it's better to use related data to train the model.

Problem 2

A) I think approach B is better. Here is a toy example. Let's say there are 3 users and 5 movies to be reviewed. The second user did not review movie 4 and 5. The rating data is shown in the following table.

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5
User 1	4	3	3	4	5
User 2	4	3	3	N/A	N/A
User 3	1	1	2	1	1

It's obivous that user2 is more similar to user 1 since they have the same rating to movie 1, 2 and 3.

But if we use approach A, we will fill 0 to "N/A". Then

$$d_A(\text{user1, user2}) = 0+0+0+4+5 = 9.$$
$$d_A(\text{user2, user3}) = 3+2+1+1+1 = 8.$$
$$d_A(\text{user2, user3}) < d_A(\text{user1, user2}).$$

This means user 2 and user 3 are more similar, which contradicts to our previous observation.

However, if we use approach B, we will fill the average 3 to "N/A". Then

$$d_B(\text{user1, user2}) = 0+0+0+1+2 = 3.$$
$$d_B(\text{user2, user3}) = 3+2+1+2+2 = 10.$$
$$d_B(\text{user1, user2}) < d_B(\text{user2, user3})$$

This means user 1 and user 2 are more similar, which is the same as our previous observation.

B) I think Euclidian distance would be better. Here is a toy example. 3 users review 3 movies.

	Movie 1	Movie 2	Movie 3
User 1	1	3	3
User 2	2	4	4
User 3	3	5	4

It’s obvious that movie 2 is more similar to movie 3 since they have the same ratings by user 1 and user 2.

If we use Pearson’s correction, then Pearson correlation coefficient between movie 2 and movie 3 is 0.866. But the coefficient between movie 2 and movie 1 is 1. This means movie 2 and movie 1 are more similar than movie 2 and movie 3, which contradicts our previous observation.

If we use Euclidian distance, then

$$d_E(\text{Movie2, Movie3}) = 1$$

$$d_E(\text{Movie2, Movie1}) = 3.464$$

$$d_E(\text{Movie2, Movie3}) < d_E(\text{Movie2, Movie1})$$

This Movie 2 is more similar to Movie 3, which is the same as our previous observation.

Problem 3

Please see code files, “user_cf.py” and “item_cf.py”

Problem 4

A) The error measure I will use is as below:

$$error = \begin{cases} 1, & if\ true\ rating \neq prediction \\ 0, & if\ (true\ rating = prediction)\ or\ (ture\ rating = 0) \end{cases}$$

For example,

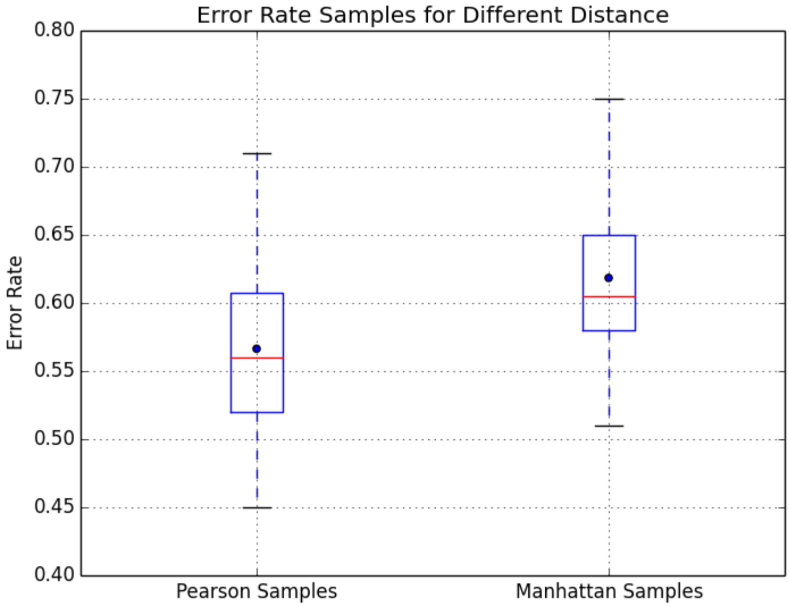
Case 1: if my predicted rating is 3 for user A on movie B, but the true rating is 2, then error is 1.

Case 2: if my predicted rating is 3 for user A on movie B, but the true rating is 3, then error is 0.

Case 3: if my predicted rating is 3 for user A on movie B, but the true rating is 0, then error is 0.
There are two reasons I define the error in this way. First, our prediction is obtained by using mode of K-nearest neighbors, not the average of K-nearest neighbors' ratings, this means the prediction doesn't reflect the average degree of ratings. So there's not enough condition or meaning to measure the degree of the error. Second, if the true rating doesn't exist (i.e. 0 value), then there's no error to measure. So in this case we just simply set it to 0.

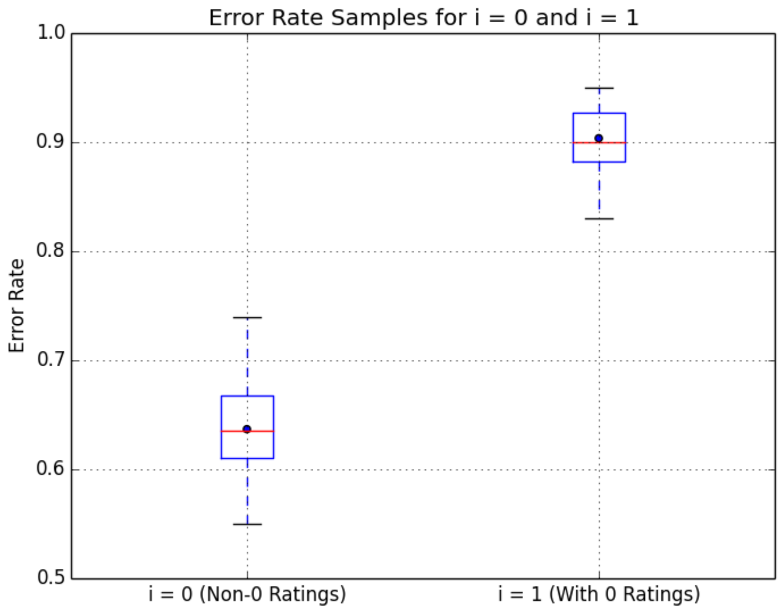
B) I will use independent samples t-test. I choose it because each sample experiment includes 100 trials, where each trial has two results, 1 or 0. So the number of successful trials could be regarded as a binomial distributed random variable. And since the number of trials is large (>30), this binomial can be approximated by a normal distribution. The null hypothesis is there is no significant difference between the two sample means. I set the confidence level to be 95%, i.e. P = 0.05.

C) I choose K = 15 and i = 0 (Not vary). Compare if there's significant difference between Pearson's correlation (distance = 0) and Manhattan distance (distance = 1) by using independent samples t-test. (Here I chose Welch's t-test, since we cannot guarantee that the two samples have same variances). After doing t-test, I got P-value is 1.775e-05 < 0.05. So we can reject the null hypothesis, which means there is significant difference between the samples with Pearson's distance and the samples with Manhattan distance. The box plot of two samples and the average (labeled by blue dot) are shown in the following picture.



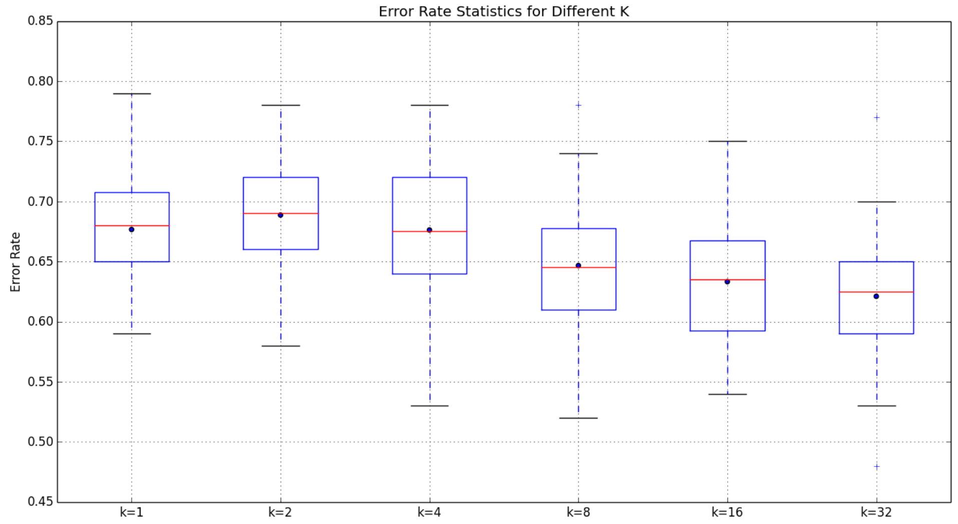
From the figure above we can also see clearly that Pearson's correlation performs better than Manhattan distance in terms of the error rate.

D) I choose K = 15 and distance = 0 (Not vary). Compare if there is difference between i = 0 (Non-0 ratings) and i = 1 (with 0 ratings) in user-based collaborative filtering. After doing Welch's t-test I got P-value is 1.997e-53 < 0.05. So we can reject the null hypothesis, which means there is significant difference between i = 0 (Non-0 ratings) and i = 1 (with 0 ratings). The box plot of two samples and the average (labeled by blue dot) are shown in the following picture.



The above figure clearly shows that the performance of non-0 ratings (i = 0) is much better than that with 0 ratings (i = 1) in terms of the error rate.

E) From C) we find that distance = 0 (Pearson's correlation) is better, and from D) we found that i = 0 is better. So here I choose distance = 0, i = 0 (Not vary). Compare different k = 1, 2, 4, 8, 16, 32 for user-based collaborative filtering. The box plot of 6 groups of samples is shown below.



From the above picture we can see that when k = 32, the performance is the best in terms of mean value of error rate. But it's worth mentioning that when k = 32, there are several extreme cases (like noises) happen which is either way much better than k = 16 case or way much worse than k = 16 case. So the variance of k = 32 samples may not be very small.

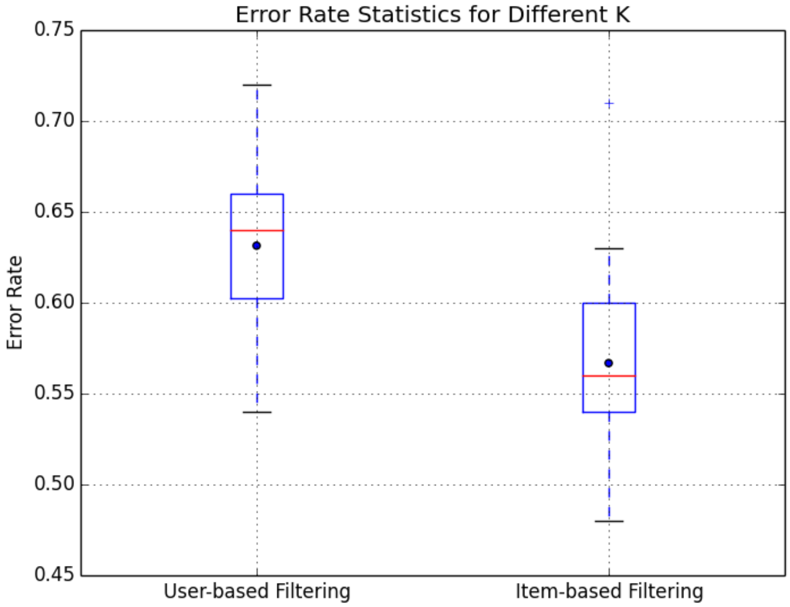
And the P-Value of Welch's t-test for samples with different k is listed in the following table.

Table 1. P-Value of T-test for Samples with Different k

P-Value	k=1	k=2	k=4	k=8	k=16	k=32
k=1	N\A	0.202	0.968	0.002	1.65e-5	2.02e-7
k=2	N\A	N\A	0.200	1.49e-5	3.35e-8	2.89e-10
k=4	N\A	N\A	N\A	0.003	3.13e-5	4.42e-7
k=8	N\A	N\A	N\A	N\A	0.159	0.011
k=16	N\A	N\A	N\A	N\A	N\A	0.218
k=32	N\A	N\A	N\A	N\A	N\A	N\A

Since the p-value for k=16 and k=32 samples (yellow block) is 0.218 > 0.05, we cannot reject the null hypothesis that the means of k=16 samples and k=32 samples are the same. So whether k=16 or k=32 is better depends on the needs of application. Here I choose k = 16 since it has less variance (0.0022) than k = 32 (with variance 0.0025).

F) From C), D) and E) we find distance = 0, i = 0 and k = 16 is better. So here we set distance = 0, i = 0 and k = 16. Compare the difference between user-based collaborative filtering and item-based one. After doing t-test, I got P-value is 5.41e-11 < 0.05. So we can reject the null hypothesis, which means there is significant difference between the samples using user-based filtering and the samples using item-based filtering. The box plot of two samples and the average are shown in the following picture.



So from the above plot and the t-test's p-value we can say that item-based filtering performed better than user-based filtering when distance = 0, i = 0 and k = 16 based on the observation from those groups of samples. Since we already set the confidence level to be 95%, P-value 5.41e-11 < 0.05 means we are confident about this conclusion. However, it is worth mentioning that the conclusion only works exactly for the randomly selected groups of samples. There are still several other assumptions of approximation from the beginning to the end, such as assume number of correct predictions in 100 rating follows binomial distribution, and assume 50 binomial distribution can approximate normal distribution so that we could do t-test etc. These factors all affect the general judgment of whether user-based or item-based filtering is better, even though in this special case (where distance = 0, i = 0, k = 16 and those selected samples) we are confident that item-based filtering performs better.