The Experiment Report of Machine Learning



Student ID：

201721045886, 201721045909, 201721045893

**SUBJECT:**SOFTWARE ENGINEERING

December 24, 2017

Grade:

Postgraduate

Supervisor:

Qingyao Wu

Author:

Siyuan Xiao, Li Zhang, Shengyan Wen

**SCHOOL:** SCHOOL OF SOFTWARE ENGINEERING

[[1]](#footnote-0)

Recommender System Based on Matrix Decomposition

Abstract—Nowadays, recommender system is a popular application that applying data mining techniques. In this field, Matrix Decomposition is used frequently.

# INTRODUCTION

This report will talk about the whole experiment I have made for a recommender system based on Matrix Factorization. Its content is organized as follow:

1. Section II contains the experiment steps.
2. Section III contains the code for the experiment.
3. Section IV makes conclusion for the experiment result.

# METHODS AND THEORY

Here we use the MovieLens-100k dataset which is placed in the requirement. The u1 database is used and it is already splitted into training set and validation set by the official.

Then the experiment will be performed by the following steps:

1. Load the data text into memory and do the transformation to get the specific column, and the cells that lack of data is set to 0 (or the average value).
2. For ALS method, we randomly initiate the Q matrix.
3. After that, we calculate P and Q alternatively until the iteration number is arrived or the RMSE has little change.
4. For SGD method, we randomly initiate the two matrices.
5. In each iteration, we randomly pick one row in P and Q, respectively.
6. We use these two selected rows to generate a prediction, then calculate the error and gradient for P and Q, respectively.
7. Similarly, we iterate until it converges or reaches the given number of iterations
8. Regardless of which method we use, do the whole prediction and calculate RMSE for the valid validation data.

# Experiment

Here I placed the critical code for the experiment:

1. Preprocessing:

|  |
| --- |
| 1. # created by Swain, 2017-12-19, 11:29 2. import numpy as np 3. from numpy.linalg import inv 4. import matplotlib.pyplot as plot 5. from numpy import random 6. #load u1 from MovieLens-100k dataset 7. n\_user = 943 8. n\_item = 1682 9. trainpath = './u1.base' 10. testpath = './u1.test' 11. data\_train = np.loadtxt(trainpath) 12. data\_test = np.loadtxt(testpath) 13. def make\_R(data): 14. count = 0 15. sum = 0 16. R = np.zeros([n\_user, n\_item]) 17. for i in range(0, data.shape[0]): 18. R[int(data[i, 0]) - 1, int(data[i, 1]) - 1] = int(data[i, 2]) 19. count = count + 1 20. sum = sum + int(data[i, 2]) 22. avg = sum / count 23. '''for i in range(0, R\_train.shape[0]): 24. for j in range(0, R\_train.shape[1]): 25. if R\_train[i, j] == 0: 26. R\_train[i, j] = avg''' 27. return R 29. R\_train = make\_R(data\_train) 30. R\_vali = make\_R(data\_train) 31. Np = np.sum(R\_train, axis=1) 32. Nq = np.sum(R\_train, axis=0) 33. random.seed(24) |

1. ALS:

|  |
| --- |
| 1. #ALS 2. k = 10 3. lambda\_ = 0.03 4. iteration = 1000 5. P = random.random([n\_user, k]) \* 5 6. Q = random.random([n\_item, k]) \* 5 7. Lambda = lambda\_ \* np.eye(k) 8. R = R\_train 9. for n in range(0, iteration): 10. print ("start the " + str(n + 1) + "-th iteration") 12. #calculate P matrix 13. for i in range(0, n\_user): 14. P[i] = np.dot(inv(np.dot(Q.T, Q) + Lambda), np.dot(Q.T, R[i])) 16. #calculate Q matrix 17. for i in range(0, n\_item): 18. Q[i] = np.dot(inv(np.dot(P.T, P) + Lambda), np.dot(P.T, R[:, i])) 20. #do prediction 21. R\_predict = np.dot(P, Q.T) 23. #calculate error 24. RMSE = 0 25. count = 0 26. for i in range(0, n\_user): 27. for j in range(0, n\_item): 28. if R\_vali[i, j] != 0: 29. RMSE = RMSE + np.power(R\_vali[i, j] - R\_predict[i, j], 2) 30. count = count + 1 31. RMSE = np.sqrt(RMSE / count) 32. print (" RMSE = " + str(RMSE)) 34. if n > 0 and np.abs(RMSE - last\_RMSE) < 0.0001: 35. break; 36. last\_RMSE = RMSE |

1. SGD:

|  |
| --- |
| 1. #SGD 2. k = 10 3. lr = 0.05 4. lambda\_p = 0.03 5. lambda\_q = 0.03 6. iteration = 1000000 7. #P = np.ones([n\_user, k]) 8. #Q = np.ones([n\_item, k]) 9. P = random.random([n\_user, k]) 10. Q = random.random([n\_item, k]) 11. R = R\_train 12. for n in range(0, iteration): 13. #randomly pick a row from P and Q 14. u = random.randint(0, n\_user - 1) 15. i = random.randint(0, n\_item - 1) 17. #calculate error and gradient 18. error = R[u, i] - np.dot(P[u], Q.T[:, i]) 19. deriv\_p = lambda\_p \* P[u] - error \* Q[i] 20. deriv\_q = lambda\_q \* Q[i] - error \* P[u] 22. #do updation 23. P[u] = P[u] - lr \* deriv\_p 24. Q[i] = Q[i] - lr \* deriv\_q 26. print (str(n + 1) + "-th iteration: error = " + str(error)) 28. #do prediction and calculate error 29. R\_predict = np.dot(P, Q.T) 30. RMSE = 0 31. count = 0 32. for i in range(0, n\_user): 33. for j in range(0, n\_item): 34. if R\_vali[i, j] != 0: 35. RMSE = RMSE + np.power(R\_vali[i, j] - R\_predict[i, j], 2) 36. count = count + 1 37. RMSE = np.sqrt(RMSE / count) 38. print (" RMSE = " + str(RMSE)) |

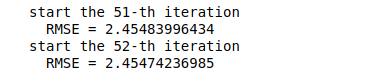
# conclusion

For ALS method, it ends with the 52-nd iteration.

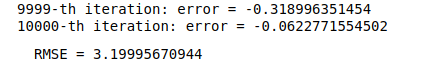
For SGD method, we do 100,000 iterations.

Here is the experiment result gained (undetermined data set to 0):

1. For ALS:



1. For SGD:



In case undetermined data set to the average value:

1. For ALS:



1. For SGD:



Then we can draw a conclusion according to the experiment:

1. ALS converges with far less iterations but may be likely to have overfitting problem.
2. SGD should run much more iterations since the data used in each iteration just generate one prediction.
3. If we set the undetermined cells of the training dataset to the average value (while the data shown is in case that setting to 0) of all training data, the RMSE of prediction can be reduced significantly.

1. [↑](#footnote-ref-0)