SGD Algorithm to predict movie ratings

In [5]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remo
unt, call drive.mount("/content/drive", force_remount=True).
```

There will be some functions that start with the word "grader" ex: grader_matrix(), grader_mean(), grader_dim() etc, you should not change those function definition.

Every Grader function has to return True.

1. Download the data from [here (https://drive.google.com/open?id=1-1z7iDB52cB6_JpO7Dqa-eOYSs-mivpq)](https://drive.google.com/open?id=1-1z7iDB52cB6_JpO7Dqa-eOYSs-mivpq)
2. The data will be of this format, each data point is represented as a triplet of user_id, movie_id and rating

user_id	movie_id	rating
77	236	3
471	208	5
641	401	4
31	298	4
58	504	5
235	727	5

Task 1

Predict the rating for a given (user_id, movie_id) pair

Predicted rating $\hat{y}_{ij}$ for user i, movied j pair is calcuated as $\hat{y}_{ij} = \mu + b_i + c_j + u_i^T v_j$, here we will be finding the best values of b_i and c_j using SGD algorithm with the optimization problem for N users and M movies is defined as

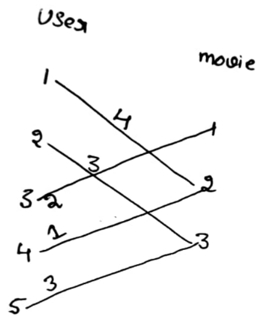
$$L = \min_{b, c, \{u_i\}_{i=1}^N, \{v_j\}_{j=1}^M} \quad \alpha\Big(\sum_{j} \sum_{k} v_{jk}^2 + \sum_{i} \sum_{k} u_{ik}^2 + \sum_{i} b_i^2 + \sum_{j} c_i^2 \Big) + \sum_{i,j \in \mathcal{I}^{\text{train}}} (y_{ij} - \mu - b_i - c_j - u_i^T v_j)^2$$

- μ : scalar mean rating
- b_i : scalar bias term for user i
- c_j : scalar bias term for movie j
- u_i : K-dimensional vector for user i
- v_j : K-dimensional vector for movie j

*. We will be giving you some functions, please write code in that functions only.

*. After every function, we will be giving you expected output, please make sure that you get that output.

1. Construct adjacency matrix with the given data, assuming its [weighted un-directed bi-partited graph](https://en.wikipedia.org/wiki/Bipartite_graph) (https://en.wikipedia.org/wiki/Bipartite_graph) and the weight of each edge is the rating given by user to the movie



the Adjacency matrix

$$
\begin{array}{c|ccc}
 & 1 & 2 & 3 \\
\hline
1 & 0 & 4 & 0 \\
2 & 0 & 0 & 3 \\
3 & 2 & 0 & 0 \\
4 & 0 & 1 & 0 \\
5 & 0 & 0 & 3
\end{array}
$$

you can construct this matrix like $A[i][j] = r_{ij}$ here i is user_id, j is movie*id and $r*{ij}$*isratinggivenbyuser*i *tothemovie*j$

Hint : you can create adjacency matrix using [csr_matrix](https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html) (https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html)

1. We will Apply SVD decomposition on the Adjaceny matrix [link1](https://stackoverflow.com/a/31528944/4084039) (https://stackoverflow.com/a/31528944/4084039), [link2](https://machinelearningmastery.com/singular-value-decomposition-for-machine-learning/) (https://machinelearningmastery.com/singular-value-decomposition-for-machine-learning/) and get three matrices $U, \sum, V$ such that
   $U \times \sum \times V^T = A,$
   if A is of dimensions $N \times M$ then
   U is of $N \times k,$
   $\sum$ is of $k \times k$ and
   V is $M \times k$ dimensions.

   *. So the matrix U can be represented as matrix representation of users, where each row u_i represents a k-dimensional vector for a user

   *. So the matrix V can be represented as matrix representation of movies, where each row v_j represents a k-dimensional vector for a movie.
2. Compute μ , μ represents the mean of all the rating given in the dataset.(write your code in def m_u())
3. For each unique user initilize a bias value B_i to zero, so if we have N users B will be a N dimensional vector, the i^{th} value of the B will corresponds to the bias term for i^{th} user (write your code in def initialize())
4. For each unique movie initilize a bias value C_j zero, so if we have M movies C will be a M dimensional vector, the j^{th} value of the C will corresponds to the bias term for j^{th} movie (write your code in def initialize())
5. Compute dL/db_i (Write you code in def derivative_db())
6. Compute dL/dc_j(write your code in def derivative_dc()

7. Print the mean squared error with predicted ratings.

```
for each epoch:
    for each pair of (user, movie):
        b_i =  b_i - learning_rate * dL/db_i
        c_j =  c_j - learning_rate * dL/dc_j
predict the ratings with formula
```

$\hat{y}_{ij} = \mu + b_i + c_j + \text{dot\_product}(u_i, v_j)$

1. you can choose any learning rate and regularization term in the range 10^{-3} to 10^{2}
2. **bonus**: instead of using SVD decomposition you can learn the vectors u_i, v_j with the help of SGD algo similar to b_i and c_j

Reading the csv file

In [6]:

```
import pandas as pd
data=pd.read_csv('/content/drive/MyDrive/RS/ratings_train.csv')
data.head()
```

Out[6]:

	user_id	item_id	rating
0	772	36	3
1	471	228	5
2	641	401	4
3	312	98	4
4	58	504	5

In [7]:

```
data.shape
```

Out[7]:

```
(89992, 3)
```

Create your adjacency matrix

In [8]:

```
u_i=data['user_id']
v_j=data['item_id']
```

In [9]:

```
rating=data['rating'].tolist()
user_details=u_i.tolist()
movie_details=v_j.tolist()
```

In [10]:

```
from scipy.sparse import csr_matrix
adjacency_matrix = csr_matrix((rating,(user_details, movie_details))).toarray()
# write your code of adjacency matrix here
```

In [11]:

```
adjacency_matrix.shape
```

Out[11]:

```
(943, 1681)
```

Grader function - 1

In [12]:

```
def grader_matrix(matrix):
  assert(matrix.shape==(943,1681))
  return True
grader_matrix(adjacency_matrix)
```

Out[12]:

```
True
```

SVD decompostion

Sample code for SVD decompostion

In [13]:

```
from sklearn.utils.extmath import randomized_svd
import numpy as np
matrix = np.random.random((20, 10))
U, Sigma, VT = randomized_svd(matrix, n_components=5,n_iter=5, random_state=None
)
print(U.shape)
print(Sigma.shape)
print(VT.T.shape)
```

```
(20, 5)
(5,)
(10, 5)
```

Write your code for SVD decompostion

In [14]:

```
# Please use adjacency_matrix as matrix for SVD decompostion
# You can choose n_components as your choice
U, Sigma, VT = randomized_svd(adjacency_matrix, n_components=5,n_iter=5, random_
state=None)
print(U.shape)
print(Sigma.shape)
print(VT.T.shape)
```

```
(943, 5)
(5,)
(1681, 5)
```

Compute mean of ratings

In [15]:

```
def m_u(ratings):
    '''In this function, we will compute mean for all the ratings'''
    # you can use mean() function to do this
    # check this (https://pandas.pydata.org/pandas-docs/stable/reference/api/pan
das.DataFrame.mean.html) link for more details.

    return np.mean(ratings)
```

In [16]:

```
mu=m_u(data['rating'])
print(mu)
```

```
3.529480398257623
```

Grader function -2

In [17]:

```
def grader_mean(mu):
  assert(np.round(mu,3)==3.529)
  return True
mu=m_u(data['rating'])
grader_mean(mu)
```

Out[17]:

```
True
```

Initialize B_i and C_j

Hint : Number of rows of adjacent matrix corresponds to user dimensions(B_i), number of columns of adjacent matrix corresponds to movie dimensions (C_j)

In [18]:

```
def initialize(dim):
    '''In this function, we will initialize bias value 'B' and 'C'.'''
    # initalize the value to zeros
    # return output as a list of zeros

    return np.zeros(dim)
```

In [19]:

```
dim= adjacency_matrix.shape[0] # give the number of dimensions for b_i (Here b_i corresponds to users)
b_i=initialize(dim)
```

In [20]:

```
dim= adjacency_matrix.shape[1]# give the number of dimensions for c_j (Here c_j corresponds to movies)
c_j=initialize(dim)
```

Grader function -3

In [21]:

```
def grader_dim(b_i,c_j):
  assert(len(b_i)==943 and np.sum(b_i)==0)
  assert(len(c_j)==1681 and np.sum(c_j)==0)
  return True
grader_dim(b_i,c_j)
```

Out[21]:

```
True
```

Compute dL/db_i

In [22]:

```
def derivative_db(user_id,item_id,rating,U,V,mu,alpha):
    '''In this function, we will compute dL/db_i'''
    dL_db_i = alpha * 2*b_i[user_id] - (2*(rating-mu-b_i[user_id]-c_j[item_id]-np.dot(U[user_id],np.transpose(V)[item_id])))
    return dL_db_i
```

Grader function -4

In [23]:

```
def grader_db(value):
    assert(np.round(value,3)==-0.931)
    return True
U1, Sigma, V1 = randomized_svd(adjacency_matrix, n_components=2,n_iter=5, random
_state=24)
# Please don't change random state
# Here we are considering n_componets = 2 for our convinence
alpha=0.01
value=derivative_db(312,98,4,U1,V1,mu,alpha)
print (value)
grader_db(value)
```

```
-0.9308283758773337
```

Out[23]:

```
True
```

Compute dL/dc_j

In [24]:

```
def derivative_dc(user_id,item_id,rating,U,V,mu, alpha):
    '''In this function, we will compute dL/dc_j'''
    dL_dc_j = alpha * 2*c_j[item_id] - (2*(rating-mu-b_i[user_id]-c_j[item_id]-n
p.dot(U[user_id],np.transpose(V)[item_id])))
    return dL_dc_j
```

Grader function - 5

In [25]:

```
def grader_dc(value):
    assert(np.round(value,3)==-2.929)
    return True
U1, Sigma, V1 = randomized_svd(adjacency_matrix, n_components=2,n_iter=5, random
_state=24)
# Please don't change random state
# Here we are considering n_componets = 2 for our convinence
r=0.01
value=derivative_dc(58,504,5,U1,V1,mu, 0.01)

grader_dc(value)
```

Out[25]:

```
True
```

Compute MSE (mean squared error) for predicted ratings

for each epoch, print the MSE value

```
for each epoch:

    for each pair of (user, movie):

        b_i =  b_i - learning_rate * dL/db_i

        c_j =  c_j - learning_rate * dL/dc_j

predict the ratings with formula
```

$\hat{y}_{ij} = \mu + b_i + c_j + \text{dot\_product}(u_i, v_j)$

In [26]:

```
from sklearn.metrics import mean_squared_error
print(np.array(data)[0])
```

```
[772  36   3]
```

In [27]:

```
from sklearn.metrics import mean_squared_error
def compute_mse(data,U,V,mu,alpha):
  y_true = data['rating']
  mse = []
  for epoch in range(51):
    for each_point in np.array(data):
      b_i[each_point[0]] = b_i[each_point[0]] - 0.001 * derivative_db(each_point
[0],each_point[1],each_point[2],U,V,mu,alpha)
      c_j[each_point[1]] = c_j[each_point[1]] - 0.001 * derivative_dc(each_point
[0],each_point[1],each_point[2],U,V,mu,alpha)

    y_pred = []
    for point in np.array(data): #zip(data['user_id'],data['item_id']):
      y_pred.append(mu + b_i[point[0]] + c_j[point[1]] + np.dot(U[point[0]], V.T
[point[1]]))

    loss = mean_squared_error(y_true, y_pred)
    mse.append(loss)

    if epoch%10 == 0:
      print("the mse: ",loss,"epoch: ",epoch)
  return mse

alpha=0.01
mse = compute_mse(data,U1,V1,mu,alpha)
```

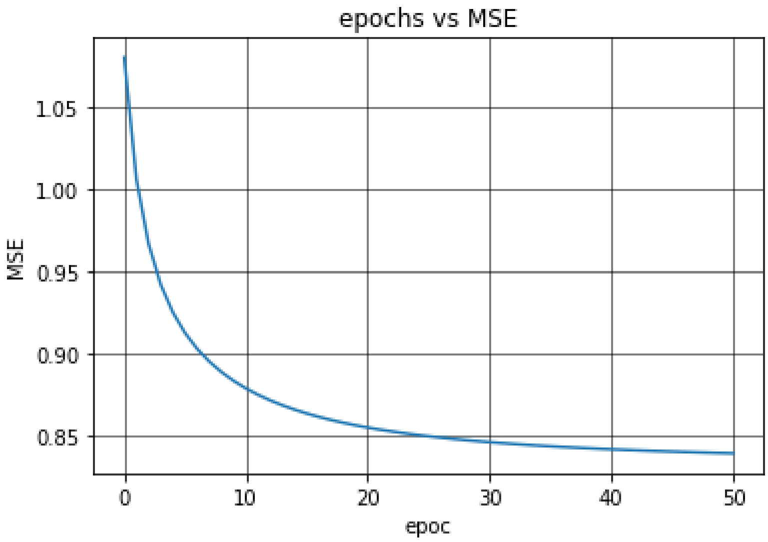
```
the mse:  1.0797989802961563 epoch:  0
the mse:  0.8790872671841912 epoch:  10
the mse:  0.8552947912172346 epoch:  20
the mse:  0.8463986335346734 epoch:  30
the mse:  0.842069050149986 epoch:  40
the mse:  0.8396161164146476 epoch:  50
```

Plot epoch number vs MSE

- epoch number on X-axis
- MSE on Y-axis

In [28]:

```
import matplotlib.pyplot as plt
epoc=np.array(list(range(51)))
plt.plot(epoc,mse)
plt.xlabel("epoc")
plt.ylabel("MSE")
plt.title("epochs vs MSE")
plt.grid(color='black', linestyle='-', linewidth=0.5)
plt.show()
```



Task 2

Task 2

As we know U is the learned matrix of user vectors, with its i-th row as the vector ui for user i. Each row of U can be seen as a "feature vector" for a particular user.

The question we'd like to investigate is this: do our computed per-user features that are optimized for predicting movie ratings contain anything to do with gender?

The provided data file [user_info.csv (https://drive.google.com/open?id=1PHFdJh_4gIPiLH5Q4UErH8GK71hTrzlY)](https://drive.google.com/open?id=1PHFdJh_4gIPiLH5Q4UErH8GK71hTrzlY) contains an is_male column indicating which users in the dataset are male. Can you predict this signal given the features U?

> **Note 1** : there is no train test split in the data, the goal of this assignment is to give an intution about how to do matrix factorization with the help of SGD and application of truncated SVD. for better understanding of the collabarative fillerting please check netflix case study.
>
> **Note 2** : Check if scaling of U, V matrices improve the metric

In [29]:

```
data1=pd.read_csv('/content/drive/MyDrive/RS/user_info.csv.txt')
data1.head()
```

Out[29]:

	user_id	age	is_male	orig_user_id
0	0	24	1	1
1	1	53	0	2
2	2	23	1	3
3	3	24	1	4
4	4	33	0	5

In [30]:

```
df2 = pd.DataFrame(data = U1,)
X=df2
```

In [31]:

```
df2['is_male']=data1['is_male']
df2.head()
y=df2['is_male']
```

In [32]:

```
X=X.drop('is_male',axis='columns')
```

In [33]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.3, random
_state=42)
```

In [34]:

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(random_state=0).fit(X_train, y_train)
```

In [35]:

```
y_pred=clf.predict(X_test)
```

In [62]:

```
from sklearn import metrics
cm = metrics.confusion_matrix(y_test,y_pred)
print(cm)

import seaborn as sns
categories = ['Zero', 'One']
sns.heatmap(cm, annot=True)
plt.title('Confusion Matrix for test data')
plt.ylabel("True Labels")
plt.xlabel("Predicted Labels")
```

```
[[  0  83]
 [  0 200]]
```

Out[62]:

```
Text(0.5, 15.0, 'Predicted Labels')
```