SGD Algorithm to predict movie ratings

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
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

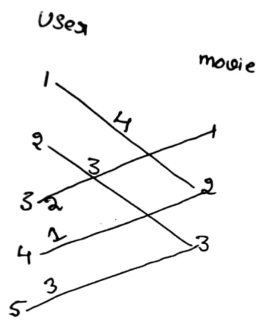
$$L = \min_{b,c,\{u_i\}_{i=1}^N,\{v_j\}_{j=1}^M} \quad \alpha\Big(\sum_j \sum_k v_{jk}^2 + \sum_i \sum_k u_{ik}^2 + \sum_i b_i^2 + \sum_j c_i^2\Big) + \sum_{i,j \in \mathcal{I}^{\text{train}}} (y_{ij} - \mu - b_i - c_j - u_i^T v_j)^2$$

- 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 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 $r{ij}isratinggivenbyuseri tothemoviej$*

Hint : you can create adjacency matrix using csr_matrix

1. We will Apply SVD decomposition on the Adjaceny matrix link1, link2 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

In [1]:
```
from google.colab import files
files = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving ratings_train.csv to ratings_train.csv
```

In [1]:
```
```

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 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

Reading the csv file

In [2]:
```
import pandas as pd
data=pd.read_csv('ratings_train.csv')
data.head()
```

Out[2]:

	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 [3]:
```
data.shape
```

Out[3]:
```
(89992, 3)
```

In [4]:
```
data.describe()
```

Out[4]:

	user_id	item_id	rating
count	89992.000000	89992.000000	89992.000000
mean	461.579151	423.584663	3.529480
std	266.720677	330.264625	1.125686
min	0.000000	0.000000	1.000000
25%	253.000000	173.000000	3.000000
50%	446.000000	319.000000	4.000000
75%	681.000000	629.000000	4.000000
max	942.000000	1680.000000	5.000000

Create your adjacency matrix

In [5]:
```
from scipy.sparse import csr_matrix
adjacency_matrix = csr_matrix((data['rating'],(data['user_id'], data['item_id'])))
```

In [6]:
```
adjacency_matrix.shape
```

Out[6]:
```
(943, 1681)
```

Grader function - 1

In [7]:
```
def grader_matrix(matrix):
  assert(matrix.shape==(943,1681))
  return True
grader_matrix(adjacency_matrix)
```

Out[7]: True

The unique items in the given csv file are 1662 only . But the id's vary from 0-1681 but they are not continuous and hence you'll get matrix of size 943x1681.

SVD decompostion

Sample code for SVD decompostion

In [8]:
```
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 [27]:
```
# 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 = 20, n_iter = 5, random_st
print(u.shape)
print(sigma.shape)
print(vt.T.shape)
```

```
(943, 20)
(20,)
(1681, 20)
```

Compute mean of ratings

In [10]:
```
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/pandas.Data


    return ratings.mean()
```

In [11]:
```
mu=m_u(data['rating'])
print(mu)
```

```
3.529480398257623
```

Grader function -2

In [12]:
```
def grader_mean(mu):
  assert(np.round(mu,3)==3.529)
  return True
mu=m_u(data['rating'])
grader_mean(mu)
```

Out[12]: 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 [13]:
```
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 [14]:
```
dim = adjacency_matrix.shape[0]  # give the number of dimensions for b_i (Here b_i corre
b_i=initialize(dim)
```

In [15]:
```
dim= adjacency_matrix.shape[1] # give the number of dimensions for c_j (Here c_j corresp
c_j=initialize(dim)
```

Grader function -3

In [16]:
```
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[16]:
```
True
```

Compute dL/db_i

In [17]:
```
def derivative_db(user_id,item_id,rating,U,V,mu,alpha):
    '''In this function, we will compute dL/db_i'''
    reg_term = 2*alpha*b_i[user_id]
    loss_term = (-2)*(rating - mu - b_i[user_id] - c_j[item_id] - np.dot(U1[user_id],V1.

    return reg_term+loss_term
```

Grader function -4

In [18]:
```
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=2
# 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)
grader_db(value)
```

Out[18]:
```
True
```

Compute dL/dc_j

In [19]:
```
def derivative_dc(user_id,item_id,rating,U,V,mu, alpha):
    '''In this function, we will compute dL/dc_j'''
    reg_term = 2*alpha*c_j[item_id]
    loss_term = (-2)*(rating - mu - b_i[user_id] - c_j[item_id] - np.dot(U1[user_id],V1.

    return reg_term+loss_term
```

Grader function - 5

In [20]:
```
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=2
# Please don't change random state
# Here we are considering n_componets = 2 for our convinence
alpha=0.01
value=derivative_dc(58,504,5,U1,V1,mu, alpha)
grader_dc(value)
```

Out[20]:
```
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 [51]:
```
b_i = np.zeros(943)
c_j = np.zeros(1681)
from sklearn.metrics import mean_squared_error
alpha = 0.5
epochs = []
error = []
learning_rate = 0.0001
for e in range(30):
  epochs.append(e+1)
  preds = []
  for user_id, item_id, rating in zip(data.iloc[:,0], data.iloc[:,1], data.iloc[:,2]):
    db = derivative_db(user_id, item_id, rating, u, vt, mu, alpha)
    b_i[user_id] = b_i[user_id] - learning_rate*db
    dc = derivative_dc(user_id, item_id, rating, u, vt, mu, alpha)
    c_j[item_id] = c_j[item_id] - learning_rate*dc
  for user_id, item_id, rating in zip(data.iloc[:,0], data.iloc[:,1], data.iloc[:,2]):
    preds.append(mu+b_i[user_id]+c_j[user_id]+np.dot(u[user_id],vt.T[item_id]))
  error.append(mean_squared_error(data['rating'],np.array(preds)))
```

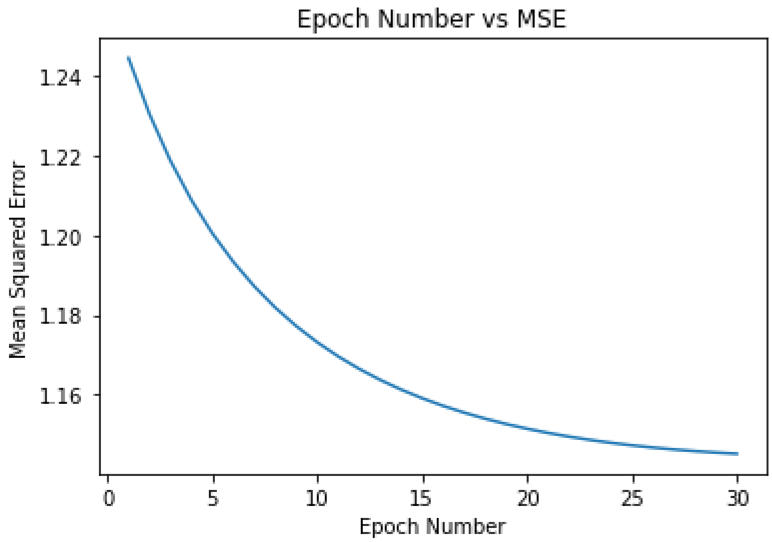
Plot epoch number vs MSE

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

In [53]:
```
import matplotlib.pyplot as plt

plt.plot(epochs, error)
plt.xlabel('Epoch Number')
```

```
plt.ylabel('Mean Squared Error')
plt.title('Epoch Number vs MSE')
plt.show()
```



Task 2

- For this task you have to consider the user_matrix U and the user_info.csv file.
- You have to consider is_male columns as output features and rest as input features. Now you have to fit a model by posing this problem as binary classification task.
- You can apply any model like Logistic regression or Decision tree and check the performance of the model.
- Do plot confusion matrix after fitting your model and write your observations how your model is performing in this task.

- Optional work- You can try scaling your U matrix.Scaling means changing the values of n_componenets while performing svd and then check your results.

In [54]:
```
from google.colab import files
files = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving user_info.csv.txt to user_info.csv.txt
```

In [55]:
```
df = pd.read_csv('user_info.csv.txt')
df.head()
```

Out[55]:

	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

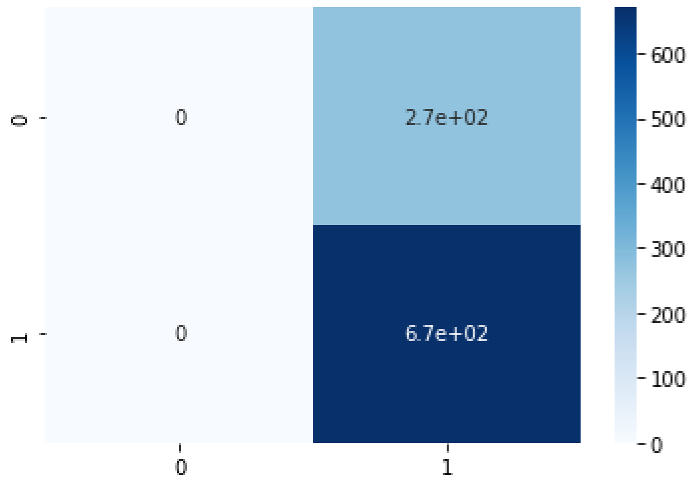
In [56]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
import seaborn as sns

model = LogisticRegression(C = 0.01).fit(u,df['is_male'])
preds = model.predict(u)

print("Accuracy score: ", accuracy_score(df['is_male'], preds))
sns.heatmap(confusion_matrix(df['is_male'], preds), annot = True, cmap = 'Blues')
```

```
Accuracy score:  0.7104984093319194
```

Out[56]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb5da87af10>
```



In [57]:
```
df['is_male'].value_counts()
```

Out[57]:
```
1    670
0    273
Name: is_male, dtype: int64
```

OBSERVATIONS:

The user matrix we got on performing SVD doesn't have data to accommodate/ consider gender as a user_info for recommendation tasks. The above logistic regression model seems to predict is_male feature as 1 always. The model achieves even a 71% accuracy score due to the imbalanced nature of the user_info dataset.

In [44]:
```

```