HW6

學號 ：r06522828　系級 ： 機械碩一　姓名 ： 王榆昇
Collaborator ： r06222023 王宇昕 r06522825 董士豪

1. (1 %)請比較有無 normalize 的差別。並說明如何 normalize.

Normalize	Public	Private
Yes	0.86130	0.85367
No	0.86552	0.85586

Embedding dimension = 40， batch_size = 128，optimizer = adam

將 train.csv 裡的 Rating 讀進來後，對該資料做 gaussion normalization:

```
def normalize(x):
    x = np.array(x);
    mean = np.mean(x);
    std = np.std(x);
    return (x-mean)/std;
```

用經上面處理、正規化的資料做訓練，而在預測 test data 時，得到的結果必須還原到正規化前(使用 training data 的 mean、std):

```
pred = pred * std + mean;
```

由 kaggle 結果可見，經正規化的資料通常能幫助 model 收斂，因此也得到較好的結果。

2. (1 %)比較不同的 embedding dimension 的結果。

Embedding Dimension	Public	Privat
40	0.86130	0.85367
100	0.85883	0.85250
500	0.85860	0.85175
1000	0.87171	0.86619

batch_size = 128，optimizer = adam

當 embedding dimension 太小時，沒有足夠大的 vector 去表示該項的特性，使得相乘後的結果會聚集在一起，loss 上升；而 embedding dimension 太大時，vector 內會包含許多無意義的值，使 vector 無法準確描述該人或電影，也會導致 loss 上升，因此取到合適的中間值會有較好的結果。

3. (1 %)比較有無 bias 的結果。

Bias	Public	Private
Included	0.85583	0.85250
Not Included	0.86209	0.85413

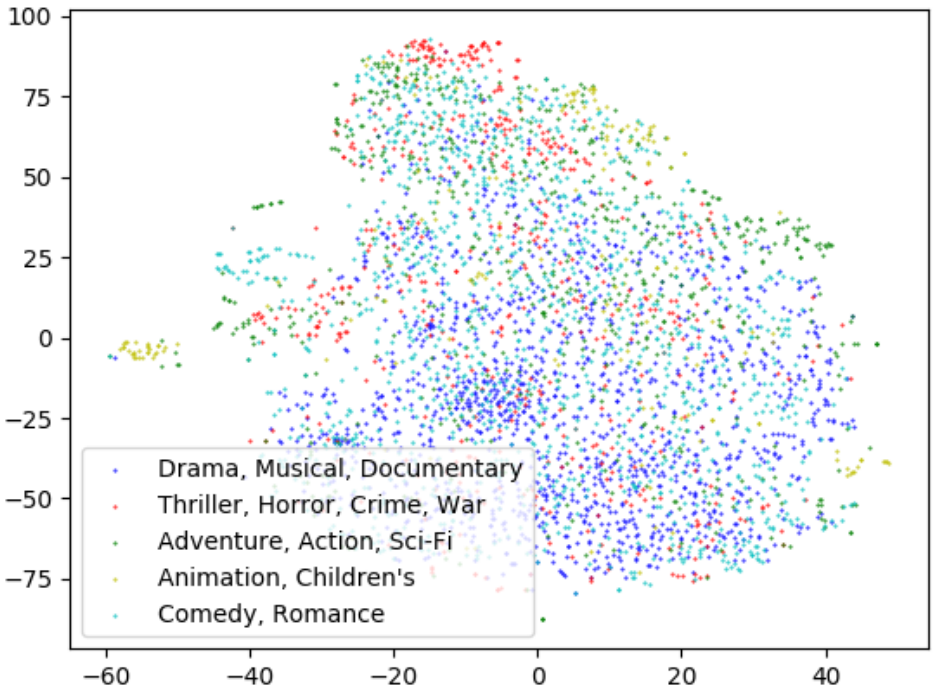
Embedding dimension = 100，batch_size = 128，optimizer = adam

從上表來看，加入 bias 項後結果明顯會變得更好，這個 bias 可能代表該用戶喜愛看電影的程度、電影題材吸引人的程度等……，說明 bias 對於一個 MF 問題的確是非常重要的。

4. (1 %)請試著將 movie 的 embedding 用 tsne 降維後，將 movie category 當作 label 來作圖。

我先使用 PCA 將 100 維的 embedding vector 降成 30 維，再用 sklearn 內的 TSNE 降至二維，固定取 movie 的第一個 category 作為標籤:

```
movie_tsne = TSNE(n_components=2, n_iter=2500).fit_transform(pca_result);
```



如上圖，我依照類別性質的相似度分成 5 類，降維效果並不明顯，但還是能看出一切趨勢，如紅色(恐怖、暴力) 、綠色(冒險科幻)會偏圖的上方，藍色(戲劇類)會偏圖的中下方。我認為造成以上原因是，電影的類別的確會影響代表該電影的 vector，因此以類別標籤會有一定趨勢，但電影的 embedding vector 同時還會被其他許多因素所影響，如演員、導演、片長……等，因此單純考慮類別並不能很好的區分，才會造成以上 couple 在一起的圖。

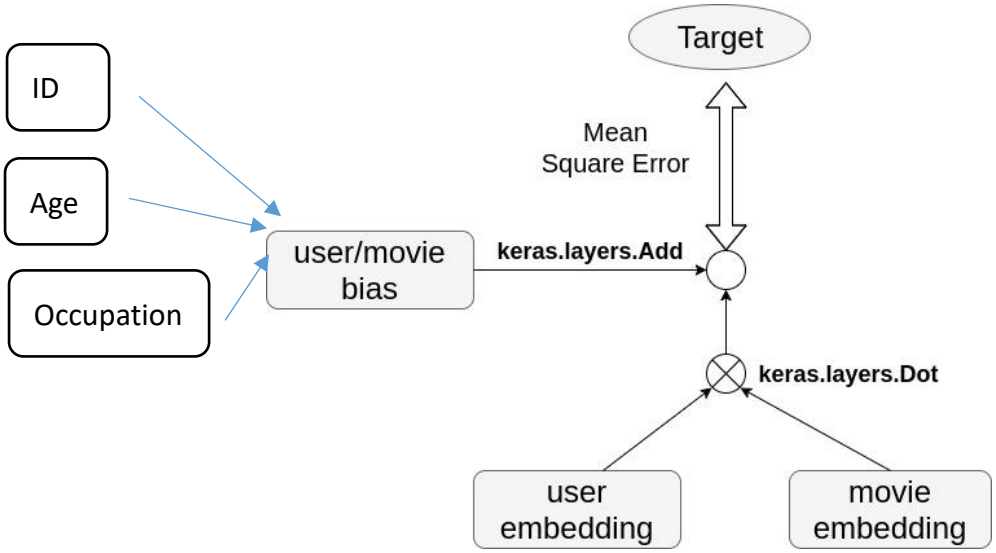
5. (1 %)試著使用除了 rating 以外的 feature, 並說明你的作法和結果，結果好壞不會影響評分。

我將 user 的 age、occupation 記錄下來成為矩陣，並一起投入學習架構中，經過 embedding 作為 bias 加入，如以下示意圖:

```
Embedding(100+1, 1, name='Age-Bias', embeddings_initializer='glorot_uniform')(age_input);
```

```
Embedding(21, 1, name='Occupation-Bias', embeddings_initializer='glorot_uniform')(occupation_input);
```

```
prod = Add()([prod,user_bias,movie_bias,age_bias,occupation_bias]);
```



	Public	Private
Original	0.85883	0.85250
With user.csv	0.85724	0.85088

embedding dimension = 100

當我考慮了 user 的 age、occupation 時，也就相當於考慮了該 user 喜歡看電影的程度，當作 bias 加進去模型裡面，不論在 public、private 的表現上都更好。