Machine Learning HW6

學號：R04922169 系級：資工所碩二 姓名：楊智偉

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

答：

我使用 training data 的 rating 數值來做 normalize。我的做法是先算出 training rating 的 mean 以及 standard deviation，就可將 training data mapping 到一個 mean=0, standard deviation=1 的 distribution 來進行 training。

Training 完成之後我們可以得到一個 normalize 過後的 model，所以在 testing 的階段，我們 predict 出來的 rating 數值還要再乘上 training std 並且加上 training mean，才會是正確的預測。

接著進行實驗比較兩者準確率的差別，結果如下表。我們可以發現有沒有加上 rating 的 normalize 對於結果沒有太大的改變。

	Public Score	Private Score
Without Normalize 1	0.86061	0.86408
Without Normalize 2	0.86049	0.86345
Without Normalize 3	0.85789	0.86233
With Normalize 1	0.86052	0.86800
With Normalize 2	0.86140	0.86633
With Normalize 3	0.85950	0.86553

2. (1%)比較不同的 latent dimension 的結果。

答：

在 Matrix Factorization 我們可以調整 latent dimension 的大小來看準確率是否會有影響。我設計三種情況來比較，分別是 latent dimension 為 5, 10, 20 的參數設定，並將結果列於下表。

我發現在 dimension = 5 的時候，public, private score 都相對較低。反之，當 dimension 調至 10 或是 20 的時候，準確率明顯上升。但是其中 10 與 20 的差距沒那們明顯，所以也不是盲目的增大就會有持續的改良。

	Public Score	Private Score
Dimension = 5 #1	0.87786	0.88152
Dimension = 5 #2	0.88207	0.88609
Dimension = 5 #3	0.88131	0.88442
Dimension = 10 #1	0.8625	0.86491
Dimension = 10 #2	0.86375	0.86829
Dimension = 10 #3	0.85925	0.86477

Dimension = 20 #1	0.85789	0.86233
Dimension = 20 #2	0.86049	0.86345
Dimension = 20 #3	0.86061	0.86408

3. (1%)比較有無 bias 的結果。

答：

為了測試有無 bias 對準確率的影響，我利用兩個不一樣的 model 來比較。結果如下表所示，可以看出使用 bias 會讓準確率提升一些約 0.02~0.04。

可以合理的推測原因，是每個 user 會有自己的評分標準，有人給的高有人給的低。而電影也應該會有 bias，畢竟每部電影的好壞不同，大家評分的結果也會有這樣的趨勢。

	Public Score	Private Score
Without bias #1	0.86139	0.86551
Without bias #2	0.86295	0.86593
Without bias #3	0.86139	0.86566
With bias #1	0.85789	0.86233
With bias #2	0.86049	0.86345
With bias #3	0.86061	0.86408

4. (1%)請試著用 DNN 來解決這個問題，並且說明實做的方法(方法不限)。並比較 MF 和 NN 的結果，討論結果的差異。

答：

DNN 的部分我的作法是將 user embedding 以及 movie embedding 兩者 concatenate 起來得到 model，會再加上 dense layer 以及 0.1~0.25 的 droupout。下表的結果是在 1 層 dense 以及 0.25 的 dropout 所做出來的結果。

可以很明顯地看到，其實 matrix factorization 以及 DNN 的方法都還蠻容易就可以通過 strong baseline 了。但聽助教提示說如果 DNN 的參數調得好應該會有較高的準確率，但我的 model 可能沒有條道特別好的參數，所以反而沒有這麼明顯提升，效果與 MF 差不多。

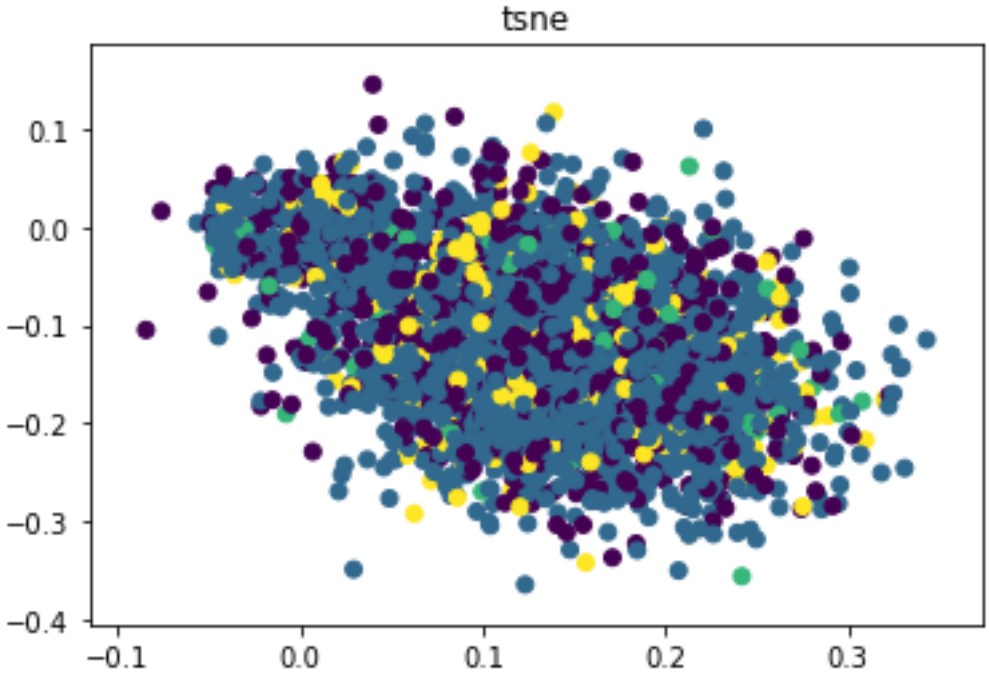
	Public Score	Private Score
MF #1	0.86139	0.86551
MF #2	0.86295	0.86593
MF #3	0.86139	0.86566
DNN #1	0.86184	0.86635
DNN #2	0.86170	0.86665
DNN #3	0.86219	0.86736

5. (1%)請試著將 movie 的 embedding 用 tsne 降維後，將 movie category 當作 label 來作圖。

答：

```
[['Thriller','Horror', 'Crime', 'Action', 'Western', 'War', 'Film-Noir'],
 ['Drama','Musical', 'Comedy', 'Romance', 'Documentary'],
 ['Animation', "Children's"],
 ['Mystery', 'Fantasy', 'Adventure', 'Sci-Fi']]
```

我將全部類別的 movie 分為四類，分類方式為以上所列。可以發現結果其實沒辦法很好的分類出來，決大部分都是互相疊合在一起的



6. (BONUS)(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。

我嘗試用過 movie 的類別來增加準確度。方法是將所有 movie 類別變成一個 18 維度的 vector 加進去一起 train。但效果差距不大，平均起來減少 0.00016 的 RMSE，不太有實質意義。