

**Matrix Factorization Model** (為此作業用以比較的 model，下文中以(\*) 標示)：

模型細節：

以 userID, movieID 作為 input，用 embedding layer 將其轉化為 dim=(128,)的 array，再接一層 Dropout (0.5) 防止 overfitting，最後將 user embedding 與 movie embedding 的內積值加上 user, movie bias 作為 training output.

1. (1%)請比較有無 **normalize(rating)**的差別。並說明如何 **normalize**.  
(collaborator: None)

	No normalization (*)	With normalization
Training score	0.5933	0.5144
Validation score	0.6712	0.6004
Public score	0.85548	0.85707
Private score	0.85514	0.85629

**Normalization method** : 將 training data 的 rating 做標準化調整到平均值( $\mu$ )為 0，標準差( $\sigma$ )為 1，丟入 training，predict 實在將 output (= x)做如下 transform:

$$f(x) = \sigma x + \mu$$

得到最後預測。

結果 **normalization** 卻沒有原本的好，猜測是在 **training data, testing data** 之間存在 **bias** 的狀況下，如此做 **normalize** 會有疑慮。

2. (1%)比較不同的 **latent dimension** 的結果。  
(collaborator: None)

Latent dimension	64	128 (*)	256	512
Training score	0.6208	0.5933	0.5973	0.5206
Validation score	0.6710	0.6712	0.6726	0.6755
Public score	0.85518	0.85548	0.85539	0.86172
Private score	0.85624	0.85514	0.85531	0.86361

可以觀察出 **latent dimension** 開到越大(512)，會有 **overfitting** 的現象。似乎是 128 就很夠用了。

3. (1%)比較有無 **bias** 的結果。  
(collaborator: None)

	With bias (*)	No bias
Training score	0.5933	0.8447

Validation score	0.6712	0.6958
Public score	0.85548	0.87832
Private score	0.85514	0.87735

很有可能會有以下現象：

有些慷慨的評分者，不論任何電影，給分都很高。(有較大的 **user bias**)

有些好電影，大家一致好評。(有較大的 **movie bias**)

若是 **model** 有考慮這點，引入 **bias** 項，**performance** 會比較好，實驗結果也如預期。

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。

(collaborator: None)

	MF (*)	DNN
Training score	0.5933	0.7195
Validation score	0.6712	0.7141
Public score	0.85548	0.91109
Private score	0.85514	0.91778

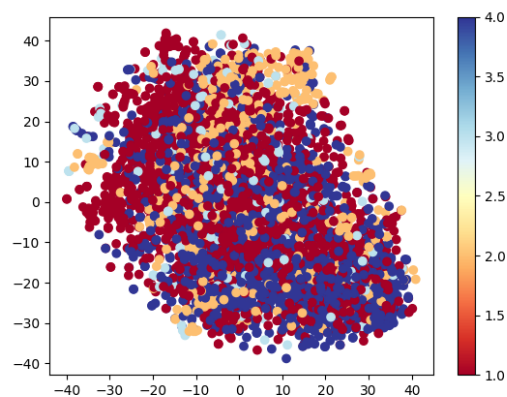
**DNN method** : user input, movie input 先過 **embedding**，再把兩者 **concatenate** 起來，一起做為 **DNN** (兩層) 的 **input**，**DNN** 的 **output** 為 **regression**。此兩 **model** 的參數量差不多。

結果是 **MF** 的方法比較好，個人認為把兩組 **embedding concatenate** 有些不直接，也許組合起來需要經過更大的 **DNN** 才能使 **performance** 變好。

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

(collaborator: None)

- 將 **animation, children, comedy, romance, musical** 分作一類，**label = red**
- 將 **crime, thriller, horror** 分作一類，**label = orange**
- 將 **fantasy, action, Sci-fi, Film-Noir, mystery** 分作一類，**label = light blue**
- 將 **drama, adventure, documentary, war** 分作一類，**label = dark blue**



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。

(collaborator: None)

	No other feature (*)	With other feature
Training score	0.5933	0.5784
Validation score	0.6712	0.6685
Public score	0.85548	0.85738
Private score	0.85514	0.85604

作法：將 **user**, **movie** 的其他 **feature** 分別通過 **NN**, 產生出個別的 **bias**, 再加上原本 **MF** 的結果, 成為最終的預測值。

**Test** 的結果是沒有加額外 **feature** 比較好。