學號：R06944049 系級： 網媒碩一 姓名：黃敬庭

1. (1%)請比較有無 normalize(rating)的差別。並說明如何 normalize.
(collaborator:)
我利用先全部減掉 mean 再除以 std 的方式對所有的 label 做 normalization, 然後在 test 的時候再反向乘以 std 然後加上 mean, 有做 normalization 時在 training 時的 loss 會小很多, 可是發現做完之後的結果並沒有比較好, 我的 mf model 在 kaggle 上達到 0.85291 但在 normalization 完之後卻只有 0.86481, 所以單就實驗的結果來看我覺得這次 normalize 並不會是一個好選擇

2. (1%)比較不同的 latent dimension 的結果。
(collaborator:)

dimension	50	200	1000
Kaggle(RMSE)	0.8517	0.8481	0.8552
Val loss(MSE)	0.7195	0.7131	0.7259

這裡比較 3 種 latent dimension 50, 200, 1000, 使用隨機 80000 筆資料做 validation, 有用上 early stopping 和 learning rate decay, 都在不做 normalization 並且加 rate = 0.5 的 dropout 下, 大概可以發現太大太小都不好, 因此看起來 200 是個不錯的選擇

3. (1%)比較有無 bias 的結果。
(collaborator:)
嘗試對 user 和 movie 各自加入 bias, bias 為一維陣列並且都預設為 0 開始, 原本 kaggle public 上的 0.85831, 可以進步到 0.85291, 因此我認為 bias 還是很幫助的, 因為他會納入每個使用者與每部電影自身的偏差, 來做出更好的模型, 而我最終最佳結果是再加入了 dropout 之後的結果

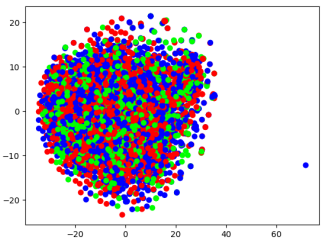
4. (1%)請試著用 DNN 來解決這個問題, 並且說明實做的方法(方法不限)。並比較 MF 和 NN 的結果, 討論結果的差異。
(collaborator:)

做法是將 user 和 movie embedding 完的 layer 直接串起來, 訓練過程使用 adamax 並且一樣使用 early stopping 與 learning rate decay 的方法, 後面接上 3 層 dense layer (128, 64, 1)最後的結果 val loss

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, None)	0	
input_2 (InputLayer)	(None, None)	0	
embedding_1 (Embedding)	(None, 1, 200)	1208000	input_1[0][0]
embedding_3 (Embedding)	(None, 1, 200)	790400	input_2[0][0]
flatten_2 (Flatten)	(None, 200)	0	embedding_1[0][0]
flatten_4 (Flatten)	(None, 200)	0	embedding_3[0][0]
dropout_1 (Dropout)	(None, 200)	0	flatten_2[0][0]
dropout_2 (Dropout)	(None, 200)	0	flatten_4[0][0]
concatenate_1 (Concatenate)	(None, 400)	0	dropout_1[0][0] dropout_2[0][0]
dense_1 (Dense)	(None, 128)	51328	concatenate_1[0][0]
dense_2 (Dense)	(None, 52)	6708	dense_1[0][0]
dense_3 (Dense)	(None, 1)	53	dense_2[0][0]

來到 0.741(MSE),然後在 kaggle public 上面來到 0.8640

5. (1%)請試著將 movie 的 embedding 用 tsne 降維後, 將 movie category 當作 label 來作圖。
   (collaborator:)
   我將比較奇幻的分在一類(blue),然後一類是比較偏劇情取向的(red),最後一類是動作戰爭之類的影片(green),最後得到這個結果(不過不太確定為什麼會有一個 outlier = =)



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:)

   我把每一部電影的類別分別轉為 0~17 的數字,然後對每一個使用者標上性別 0 為女生 1 為男生(我覺得這個比起職業年齡感覺影響更大),在 DNN 的 model 中訓練時與原本的資料串在一起送進去訓練,(也就是有 movies, user, genre, gender 4 個 input),結果出來比原本的 DNN model 好上一點點(相比第 4 題), val_lass 在 0.737, kaggle public 上則拿到 0.859