學號：R04921094 系級： 電機碩二 姓名：葉孟元

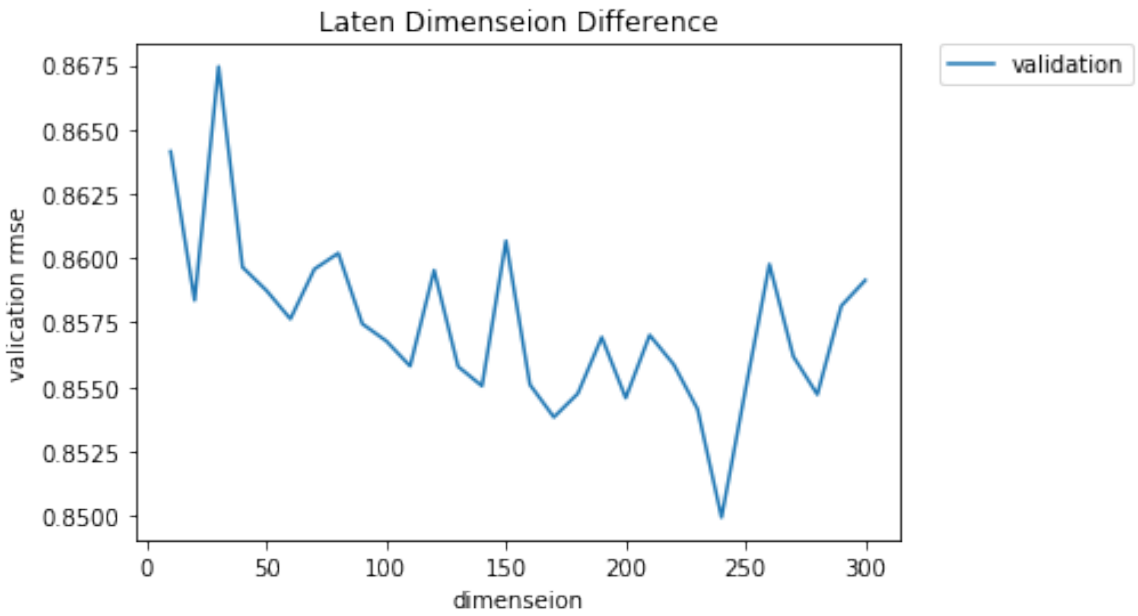
1. (1%)請比較有無 normalize(rating)的差別。並說明如何 normalize.
   答：是有所差別的，做了 normalize 之後竟然在我這邊會比較差，其實我自己不是很清楚爲什麼，我也看了助教在 fb 上的討論，我試着解釋一下我的想法，我認爲，在 feature 上做比 label 上做更有意義，在 feature 上做會讓最佳化中所說的 condition number 變小，更好做 gradient decent，但是在 label 上做就導致了 variance 變小更難被區分開。當然，還有一種可能性，因爲一個 factor 變了，包括 regularization 和 dimension 也應該跟着做對應的調整，自己沒做得比較好也只是因爲沒有找到合適的其他參數而已。

   我採用的是 sklearn 中的 StandardScaler，先做 fit_transform(), 之後再用 inverse_transform()。由於 rating 的數值改變了，所以在 learning curve 上的 rmse 和 val_rmse 並不能直接比較。我最後比較是採用一起放到 kaggle 上看他們的 public score 和 private score 的對比。這兩個模型是完全一模一樣，除了 rating 上的差別。

| | Public score | Private score |
|---|---|---|
| normalized | 0.85498 | 0.86047 |
| Non-normalized | 0.85325 | 0.85616 |

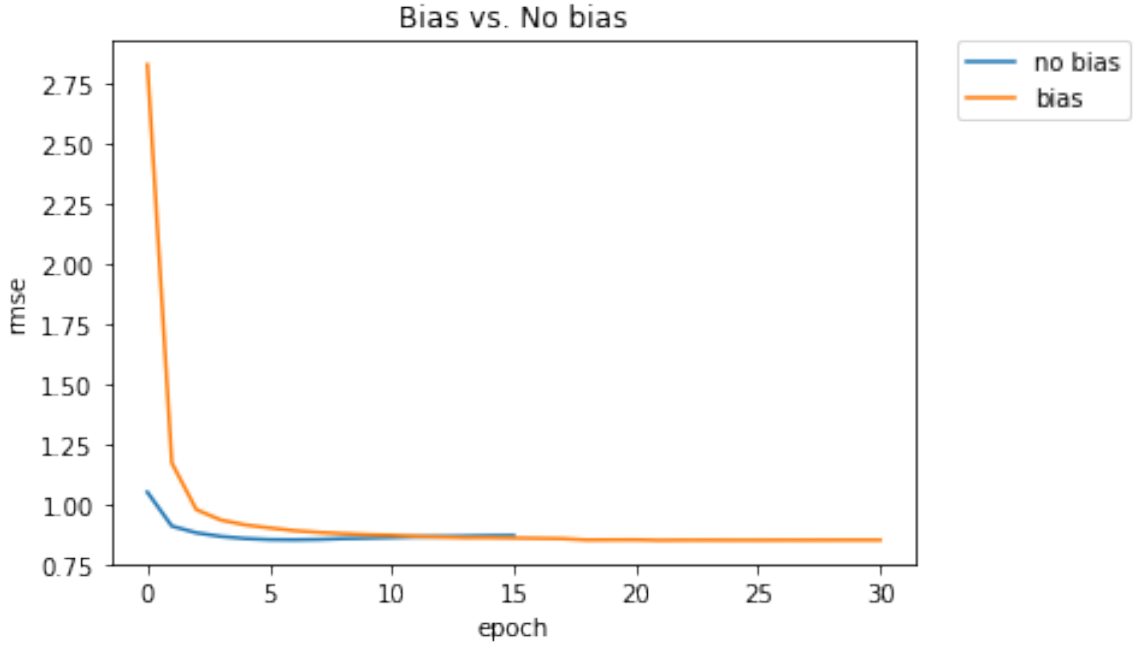
2. (1%)比較不同的 latent dimension 的結果。
   答：因爲看了 Netflix 的論文，實驗結果中表明 dimension 越大越好，也符合直覺。但是，在自己實作的時候卻發現一個有趣的現象，當 regularization 固定的時候，dimension 並不是越大越好，越大的 dimension 同時也是需要越大的 regularization，經過一個學期的學習才發現這才是對的。可是，我在做這個作業的時候才發現，可是已經過了 kaggle deadline，原本可以再調整我的模型，但是已經晚了。

   可以看到，圖上到 240 是最好的，因爲我就是按照這樣 dimension 去調整 regularization 的，也必然會發生這樣的情況。之後就開始發生 overfitting 了。可以稍微看到 240 之後趨勢是上升的。




3. (l%)比較有無 bias 的結果。

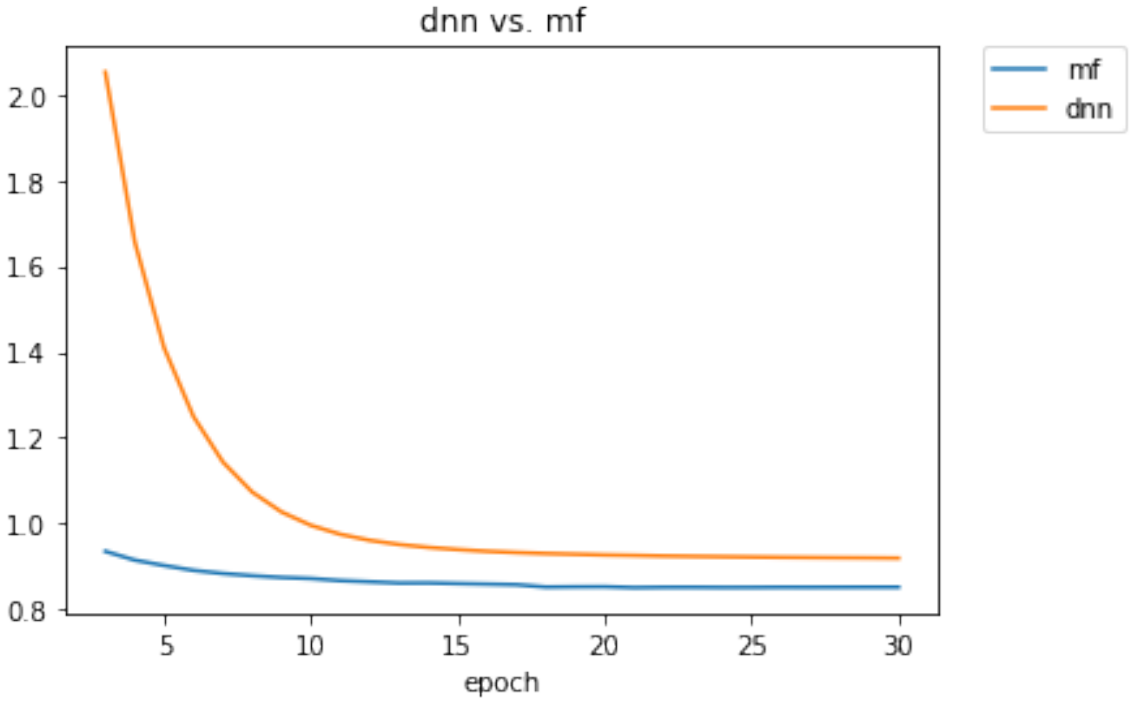
   答：有 bias 的結果會比沒有 bias 的結果明顯更好。我這邊採用的是固定的 bias，也就是不能 train 的。其實考慮過 trainable 的模型，發現結果並沒有比較好。並且在這個題目上我發現了一個有趣的現象，validation 的分數並不能說明一切，還是會有 overfitting 盡管這邊看到的沒有 bias 的模型看起來也是不錯的，可是在 kaggle 上分數還是比較差，讓我知道 k-fold 還是有必要的。下圖確實是表明了有 bias 的模型會比較好一點。但是卻沒有 kaggle 上的分數差別明顯。



| | Public score | Private Score |
|---|---|---|
| bias | 0.85325 | 0.85616 |
| No bias | 0.85975 | 0.86378 |

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：我採取的方法還是利用 embedded layer 然後將 user 和 movie vector 相加得到一個新的 vector，再接上 3 層 128 的 DNN，原本想將這個做成 classfication 的問題，可是由於是用 rmse 做最後的衡量，結果很差。就還是利用 regression 的方法來做，下圖是做出來的結果。




我認為，要利用 DNN 來做這個問題如果還是用兩個 vector 來描述這個問題並不會比較好，畢竟只有對於兩個 vector 的 inner product 才可以描述這個 ratting，而不是我這種方法。這樣必然會導致結果更差。

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

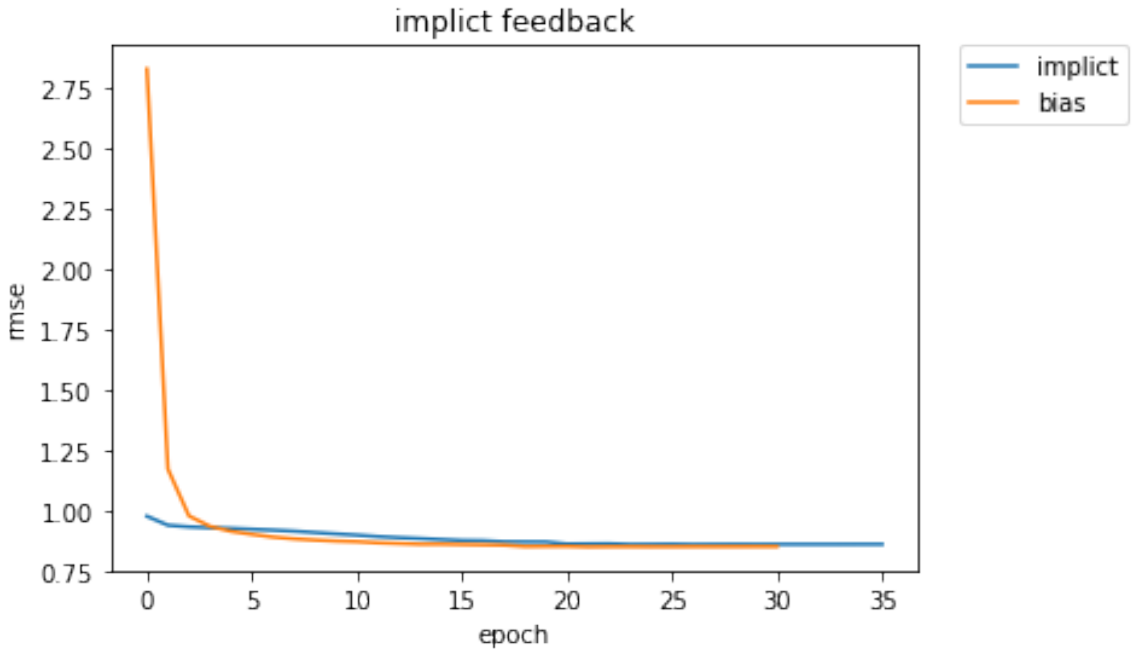
答：因為 final 時間問題，最後一天才開始寫 report，時間關係這題沒有實作。請助教扣分吧，不好意思。

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。

答：自己採用的是 Netflix prize 論文處理 implicit feedback 的做法，但是只做了 user 方面的，並沒有將 movie 的 feature 加進去。因爲我 user 做出來就沒有比之前好，於是就放棄了。將 user 的 feature 先做 preprocessing，首先先將 age 變成 age group，0 歲換爲 mean。之後將 age group, Gender, Occupation, Zip-code 都變爲一個和 user 同一維度 vector。最後將這些都加起來做爲 user 的 vector。

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T [p_u + |N(u)|^{-0.5} \sum_{i \in N(u)} x_i + \sum_{a \in A(u)} y_a]$$




下圖是比較圖，其實加上 implicit feedback 並沒有更好，當然，這確實是有可能我沒有用好的 regularization 的結果。因為自己用 bagging 的方法取得了 public 0.84467 的分數也就沒有再深究這個方法了。