

在開始回答以下問題前，先簡單描述我做為 1 ~ 4 題的比較對象之基礎 model。該 model 先將 user 與 movie id 分別做 384 維的 embedding，做內積後再加上 user, movie id 另外 embed 出的 bias (1 維)，最後得出 rating。

該 model 的表現如右：

在 report 中 optimizer皆用 Adam(lr=5e-4)，Train 約 20 個 epochs 收斂。Valid 皆取 RMSE 最低時的結果。

Data	Train	Valid	Test
RMSE	0.562	0.859	0.86

1.請比較有無normalize(rating)的差別。並

說明如何normalize。Collaborators: 無。

延續前述 model，僅對 rating 做 standard

normalization 後、記下 train data 的 mean 與 deviation，並在 testing 時把 predict rating * deviation + mean。表現如右。

由於現在 rating 的 range 被縮小了，所以在 train 跟 valid 的 RMSE 看起來會變小很多，但在 test data 的表現其實會變差。推測是因為只對 label 做 normalization 對 training 根本沒幫助 (只是改變了 prediction 的 scale)。但卻有個風險是，如果 test data 跟 train data 的分佈不同，我卻用 train data 的 mean 與 deviation 去把 prediction scale up 回來，似乎就很可能影響表現。

Data	Train	Valid	Test
RMSE	0.476	0.783	0.890

2.比較不同的latent dimension的結果。

Collaborators: 無。

大致而言，latent dimension 越高可以得到越好的結果，不過高到一定程度後就沒什麼差別了，還會增加 training 時間。

Data	Train	Valid	Test
512 dim	0.49	0.859	0.86
128 dim	0.847	0.863	0.864
64 dim	0.862	0.883	0.885

3.比較有無 bias 的結果。Collaborators: 無。

看起來有無 bias 對結果影響其實不大。

Data	Train	Valid	Test
RMSE	0.696	0.859	0.86

4.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。Collaborators: 無。

我將 user 與 movie embed 出的 vector 各過一層 Dense, BatchNormalization, LeakyReLU 與 Dropout，Dense units 為 128, dropout_rate = 0.5，出來的結果 dot 後再加上 bias。結果如右：

可能是我 DNN tune 的不好，表現差了許多。或許可透過更改層數與 units 數改善。

Data	Train	Valid	Test
RMSE	0.81	0.886	0.887

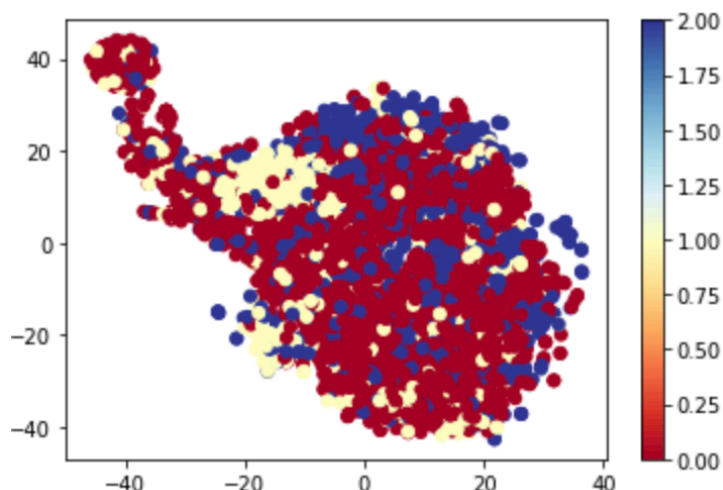
5.請試著將movie的embedding用tsne降維後，將movie category當作label來作圖。 Collaborators: b03902078 林書瑾。

我將 18 個 movie category (genre) 分成 3 大 classes，如下：

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

對每部 movie，若有多個 genre，則以其第 1 個 genre 所屬的 class 當作其 class。做出之圖形如右下：

由右圖來看似乎不太能分開這些 classes。可能是我將 genre 分類的方式並不好，也有可能是很多 movie 有多個 genre，只取第一個來賦予其 class 不恰當。當然，也可能就是 movie category 對我的 model 並不重要，所以 embedding 不會區分開來。



6.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。 Collaborators: 無。

承第 5 題，對每筆資料，我將 movie 的 category class 當作額外的 input，也 embed 出一個 bias 在 MF 的最後一步加入。

結果如右，出乎意料，就算僅用這麼粗糙的簡單做法，也提升了一些表現。早知道我就早點試試看。但由此觀察，既然 category 對

Data	Train	Valid	Test
RMSE	0.705	0.848	0.849

model 有所幫助，在第 5 題中應該要能將 movies 分開才對，這部分我有點困惑。