ML HW6 Report

學號：B05902052 系級：資工二 姓名：劉家維

以下的分數若未特別說明，皆是以 public/private 的形式表示。

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

沒 normalize: 0.86411 / 0.85732

有 normalize: 0.86108 / 0.85401

normalize 方法: 先算出 training data 所有評價的標準差和平均值，再將所有評價都減去平均值後除標準差進行訓練，預測時乘標準差後再加回平均值還原。雖然效果差不多，但是有 normalize 過的在 training 時，能較快達到收斂（有 normalize 過的花了 25 個 epoch，沒 normalize 的花了 45 個 epoch）。

2.(1%)比較不同的 embedding dimension 的結果。

Dimension 10: 0.87832 / 0.87192

Dimension 30: 0.86423 / 0.85675

Dimension 300: 0.86108 / 0.85401

Dimension 越大，越能對訓練資料做擬合。雖然 Dimension 越大越容易造成過擬合，但是因為有另外用 validation set 測試（我只存 validation rmse 最好的模型），因此 Dimension 300 得到的結果並沒有明顯的過擬合現象。

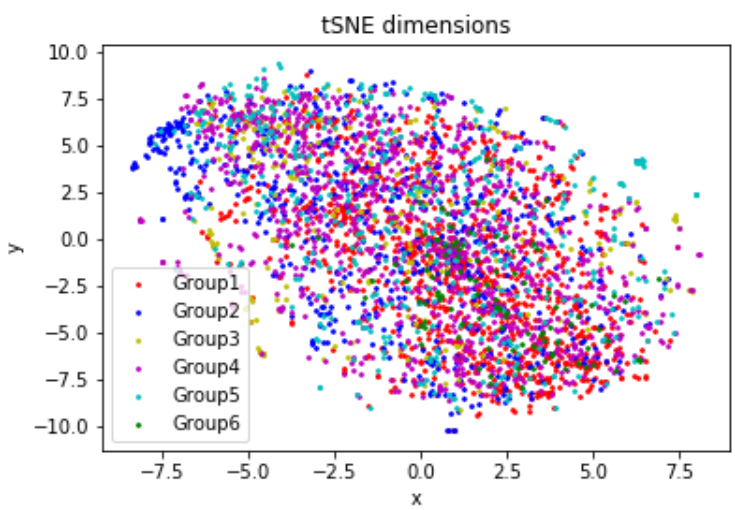
3.(1%)比較有無 bias 的結果。

沒 bias: 0.86775 / 0.86085

有 bias: 0.86108 / 0.85401

有 bias 的 model 在訓練上是有較快達到收斂（有 bias 的花了 25 個 epoch，沒 bias 的花了 30 個 epoch），但是收斂速度和結果兩者差距並沒有很大。

4.(1%)請試著將 movie 的 embedding 用 tsne 降維後，將 movie category 當作 label 來作圖。



Group1: Drama, Musical

Group2: Thriller, Horror, Crime, Film-Noir

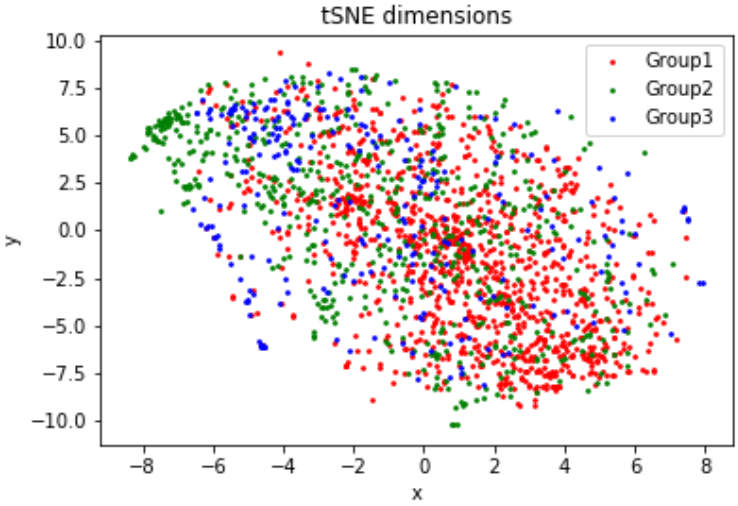
Group3: Adventure, Animation, Children's, Fantasy

Group4: Comedy, Romance, Western

Group5: Sci-Fi, Action, War, Mystery

Group6: Documentary

可以看出除了 Documentary 是只分佈在右下角一塊，其他的分類都互相混雜，也許是因為分組的方式不太好。



Group1: Drama, Musical

Group2: Thriller, Horror, Crime

Group3: Adventure, Animation, Children's

在嘗試用了助教的分組方式後，發現點的數量較少（因為有些影片不在這些分類中，我就沒畫出來），但是效果仍然沒有太好。

5.(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。

加入電影的分類進行訓練，作法是將有在 training data 出現的電影額外給出一個向量（18-dim），每個維度代表一個分類（Drama, Musical, Thriller...等等），若某電影有該分類，則該維度為 1，否則為 0。

將這個向量額外丟入一個 Dense Layer，映射出一個跟 embedding dimension 相同的 vector，將這個向量跟原本 MF 作法中 movie 單純的 embedding vector 相加後當成 movie 真正的 embedding vector。其他都跟原本的 MF 大同小異。

沒加電影分類：0.86108 / 0.85401

有加電影分類：0.86019 / 0.85013