學號：R05323040　系級：經濟所碩二　姓名：田家駿

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

   此題以 128embedded dim, 100 個 epoch,超過十次 validation 沒進步就 early stopping,batch size 為 512 做比較。沒做 normalize , validation mse 停在 1.8 下不去，上傳 kaggle 的 public 成績為 0.89988, 做完 normalize 之後，validation mse 為 0.87, 上傳 kaggle 成績為 0.8649,可以看到做完標準化之後改善了不少，因訓練資料集平均數為 3.5817,標準差為 1.116 , 因此在訓練時對資料減去平均數除以標準差，預測時再把直還原回去這裡皆是 split 5 percent 的訓練資料做 validation。

2. (1 %)比較不同的 embedding dimension 的結果

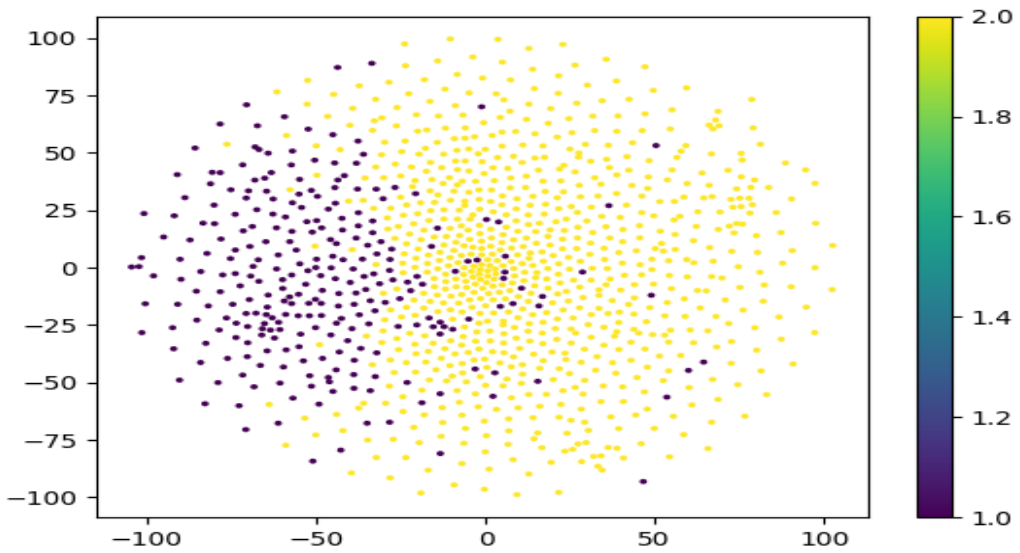
   此題比較 20,128,256 以及 512 不同的 embedded size, 其中 20 的 validation mse 為 0.87,kaggle public 成績為 0.87034, 128 validation mse 為 0.868, kaggle public 成績為 0.8649, 256 validation mse 為 0.8675, kaggle public 成績為 0.8712, 512 validation mse 為 0.872, kaggle public 成績為 0.874, 這裡皆是 split 5 percent 的訓練資料做 validation,可以看到 128 的 embbeded size 表現為最好 。

3. (1 %)比較有無 bias 的結果

   此題以 128embedded dim, 100 個 epoch,超過十次 validation 沒進步就 early stopping,batch size 為 512 做比較, 且皆是 split 5 percent 的訓練資料做 validation,若把 bias 拉掉 validation mse 為 0.95 上傳 kaggle 成績為 0.97, 加入 bias 項，validation mse 為 0.87, 上傳 kaggle 成績為 0.8649，可以看到加入 bias 項能抓到個別的 individual 的起始值, 故預測表現會較好。

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

此題我抽取同為 children 或 animation 的 movie, 在抽出同時為 Horror 或 Thriller 的電影，一共 1042 筆，下圖為降維之後的分佈圖，我們可以看到因兩種分類的 movie 差異滿大的，因此再把 embedded 的 vector 降到二維平面上分的滿開的。



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

此題我將 gender, movie genre 以及 職業分別做 one hot encoding, 維度分別為 2,18 以及 21, 我先把電影以及使用者 id 及 bias 項做一次 matrix factorization，得到 feature 維度為 1 的向量再跟 gender, movie genre, 職業,年齡 以及性別 concatenate 起來得到新的 feature，後面做了一層 hidden layer 為 64 fully connected 的 network （activation 為 elu)，最後一層在輸出一個值(activation 為 linear)，validation mse 為 0.96, 上傳 kaggle 成績為 0.94。表現不如預期增加越多 information 會變好, 原因應該為參數沒有調好。

Reference:https://github.com/qhan1028/Machine-Learning/blob/master/hw6/mf.py

Code 部份有參考之前修課同學 github 上在 validation 自己使用 keras backend 定義 mse 這一段 。