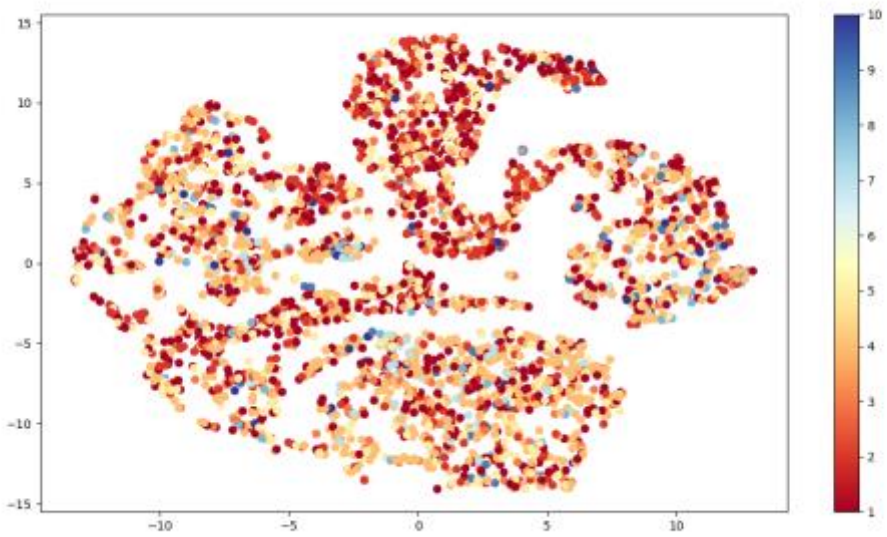
學號：R05942101 系級： 電信一 姓名：陳泓弦

1. (1%)請比較有無 normalize(rating)的差別。並說明如何 normalize.
   算出 rating 的平均值(mean)和標準差(std)，對 rating 做 normalization:
   $rating = \frac{rating-mean}{std}$，將 normalized 過的 rating 做為 training 的答案。在 testing 時再將 predict 得到的答案乘上 std 再加上 mean，得到 predict 的 rating 值。
   latent dimension 設 5，normalized 的結果在 kaggle 是 0.89766，比没有 normalized 的結果 0.88245 還差。我認為對 rating (target)做 normalize 並没有太大意義，一般做法是對 feature 做 normalize 來加快收斂的速度，而不是 target。
2. (1%)比較不同的 latent dimension 的結果。

latent dimension	RSME on Kaggle
5	0.88245
10	0.88042
15	0.88824
20	0.87867

   從上表可以發現 latent dimension 在 5～20 的結果差不多，說明其實用 5 維的 vector 就足夠找出 user 和 movie 潛在的特性。
3. (1%)比較有無 bias 的結果。
   latent dimension 設 20，做 bias 的結果是 0.87778，比没 bias 結果好了約 0.001，結果幾乎一樣。加 bias 的目的是有些人可能會傾向都給很高分或都給很低分，結果差不多的原因，我認為是在 traing 的過程中，model 就會自己學習到哪些 user 有這些特性。
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。
   取 data 中的 userID 和 movieID 做 embedding，latent dimension 設 200 ，再將 user embedding 和 movie embedding concatenate 起來再過一層 200 units 的 DNN, 得到 rating 值。在 output 有試過 regression 和分類(分 5 類)的問題來處理，regression 的結果好一點，在 kaggle 拿到 0.87160，比 MF 好，低了 0.07。

5. (1%)請試著將 movie 的 embedding 用 tsne 降維後，將 movie category 當作 label 來作圖。
   將 Animation, Children's, Comedy 視為一類，Adventure, Fantasy, Action, Sci-Fi 為一類，Drama, Musical 為一類，Crime, Thriller, Horror 為一類，其他自己一類。一電影若有多個分類時，則隨機選一個。



從上圖可以發現深紅色的點(Animation, Children's, Comedy)分佈在上方，而淺橘色的點(Drama, Musical)分佈較散，主要在左、下方，可以推測深紅色的分類還蠻正確，而淺橘色的點或許能再分更細。另外藍色的點(Documentray:7, Mystery&Film-Noir:8, Western:9, War:10)資料很少，分散的很廣，可能可以把這些類別加到深紅色或淺橘色類別中。

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。

   我使用 MF 當 model。除了 userID 和 movieID，還另外把 user 的 age 和 movie 的 genre 拿下來當 feature，做法是將 userID、age、movieID、genre 做 embedding，latent dimension 設 5，再將 userID embedding 和 age embedding concatenate 起來，movieID embedding 和 genre_embedding concatenate 起來，再將這兩個 vector 做內積，target 是 rating 值，在 kaggle 拿到 0.88298，和只拿 userID 和 movieID 當 feature 的結果 0.88245 差不多。我原本認為多拿 age 和 genre 當 feature 能做出更好的結果，因為我認為 rating 和這兩個資料會有關，例如有些人特別討厭看恐怖片而給很低的分數，但從結果來看似乎和這兩個資料沒有關係，也或許是我的 latent dimension 太低，不過因為時間關係沒辦法跑更高 dimension 的 model，之後會再試試看調高 latent dimension 是否會更好。