Machine Learning HW6

TAs
ntumlta2018@gmail.com

Outline

1. Task Introduction
2. Kaggle
3. Deadline and Policy
4. FAQ

Task Introduction

Matrix Factorization

Task Introduction

- Given the user's rating history on items, we want to predict the rating of unseen (user,item) pairs.
- We want you to implement matrix factorization to predict the missing value on user-item matrix.

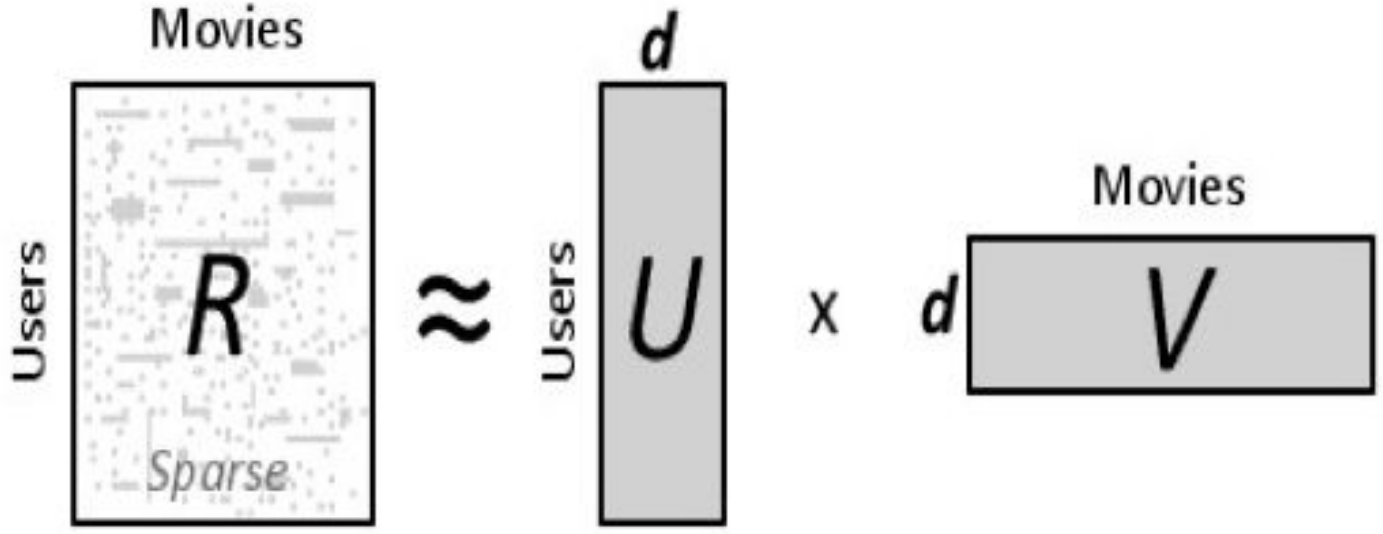
Matrix Factorization(1/4)

- Predict missing values.

	涼宮春日的憂鬱	4月是你的謊言	科學超電磁砲
大木博士	5	N/A	4
小智	N/A	3	N/A
小茂	2	N/A	2
吸盤魔偶	4	2	N/A

Matrix Factorization(2/4)

$$R \approx \hat{R} = U \cdot V^T$$



Matrix Factorization(3/4)

- Minimize loss function by gradient descent.

$$L = \sum_{i,j} (R_{ij} - U_i \cdot V_j)^2$$

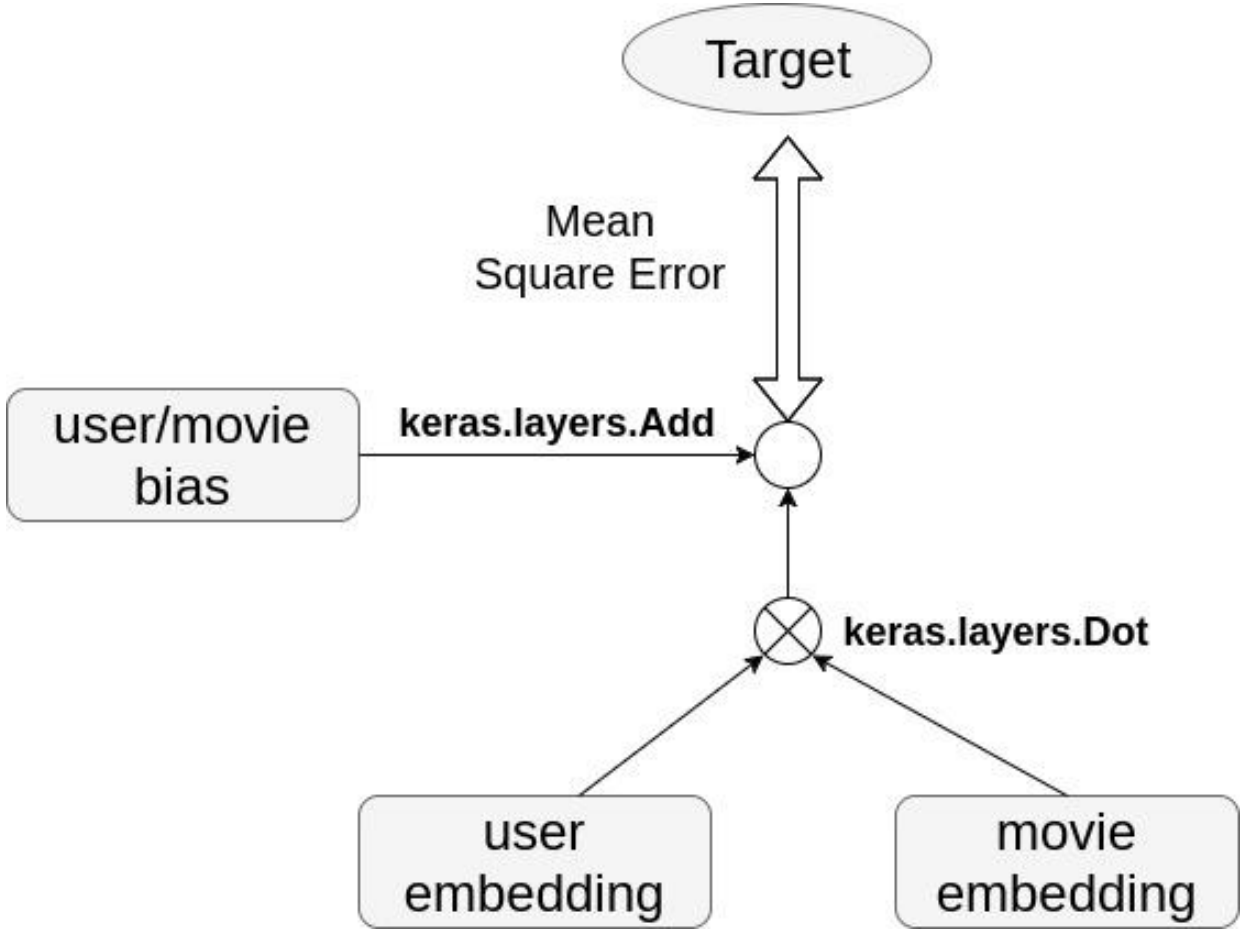
	涼宮春日的憂鬱	4月是你的謊言	科學超電磁砲
大木博士	4.7	2.7	3.9
小智	3.2	3.5	3.7
小茂	1.9	2.5	2.2
吸盤魔偶	4.1	1.8	1.2

Matrix Factorization(4/4)

- Bias term

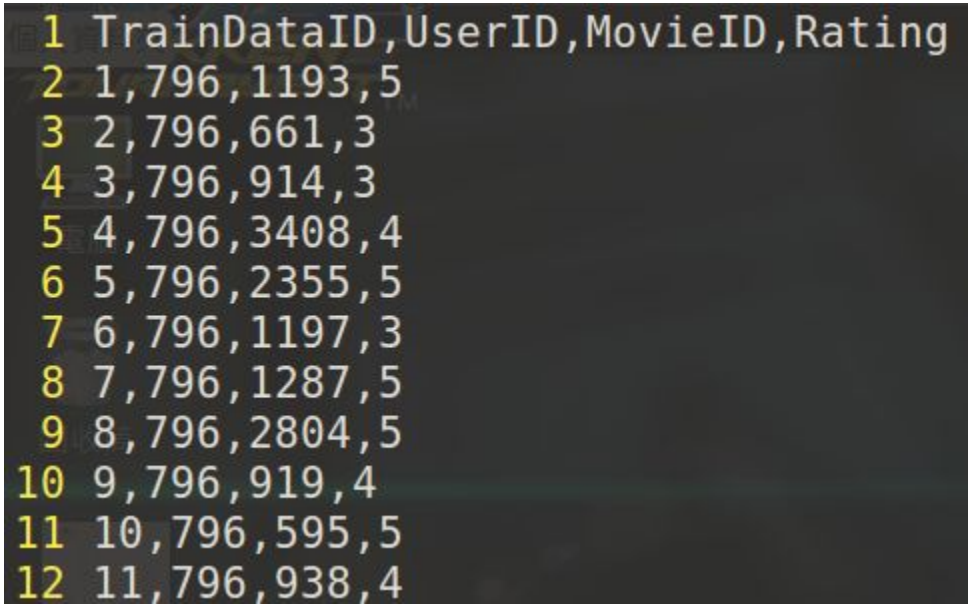
$$r_{i,j} = U_i \cdot V_j + b_i^{user} + b_j^{movie}$$

Flowchart


Target
Mean
Square Error
user/movie
bias
keras.layers.Add
keras.layers.Dot
user
embedding
movie
embedding

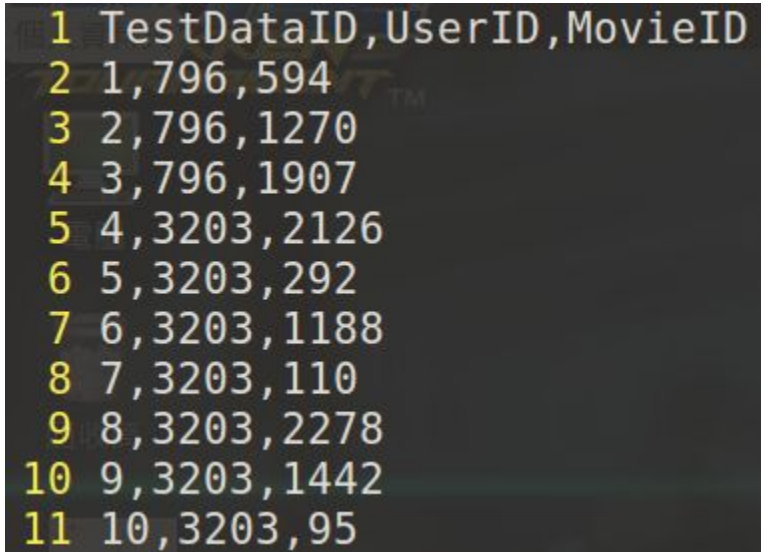
Data format(1/5)

- train.csv
- TrainDataID, UserID,MovieID,Rating



Data format(2/5)

- test.csv
- TestDataID,UserID,MovieID



Data format(3/5)

- movies.csv
- movieID::Title::Genres

```
 1 movieID::Title::Genres
 2 1::Toy Story (1995)::Animation|Children's|Comedy
 3 2::Jumanji (1995)::Adventure|Children's|Fantasy
 4 3::Grumpier Old Men (1995)::Comedy|Romance
 5 4::Waiting to Exhale (1995)::Comedy|Drama
 6 5::Father of the Bride Part II (1995)::Comedy
 7 6::Heat (1995)::Action|Crime|Thriller
 8 7::Sabrina (1995)::Comedy|Romance
 9 8::Tom and Huck (1995)::Adventure|Children's
10 9::Sudden Death (1995)::Action
11 10::GoldenEye (1995)::Action|Adventure|Thriller
12 11::American President, The (1995)::Comedy|Drama|Romance
13 12::Dracula: Dead and Loving It (1995)::Comedy|Horror
```

Data format(4/5)

- users.csv
- UserID::Gender::Age::Occupation::Zip-code

```
 1 UserID::Gender::Age::Occupation::Zip-code
 2 796::F::1::10::48067
 3 3203::M::56::16::70072
 4 4387::M::25::15::55117
 5 4771::M::45::7::02460
 6 1191::M::25::20::55455
 7 2868::F::50::9::55117
 8 1070::M::35::1::06810
 9 5074::M::25::12::11413
10 5585::M::25::17::61614
11 3402::F::35::1::95370
12 5500::F::25::1::04093
```

Data format(5/5)

1. Training data: 899873
2. Testing data: 100336, half private set.

Evaluation

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(\hat{y}_t - y_t)^2}{n}}$$

RMSE numpy implementation

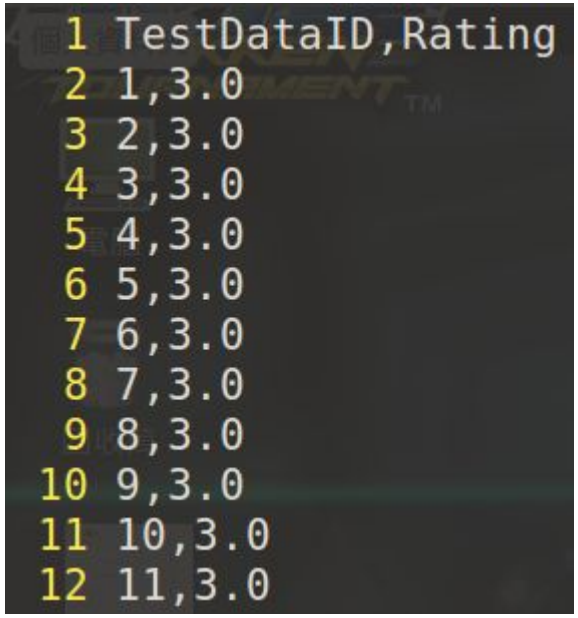
```
np.sqrt(((y_pred - y_true)**2).mean())
```

Kaggle

Kaggle

- kaggle_url: https://www.kaggle.com/t/a76e7f88bec74d9f97dbe88520a123c0
- 請使用先前使用的kaggle帳號登入。。
- Team name: 學號
- 注意！學號only！ e.g. b03901086, d03901777 ...
- 旁聽同學請勿使用類似學號開頭的隊名。
- 每日上傳上限5次。
- test set的資料將被分為兩份，一半為public，另一半為private。
- 最後的計分排名將以2筆自行選擇的結果，測試在private set上的準確率為準。
- kaggle名稱錯誤者將不會得到任何kaggle上分數。

Submission Format



```
 1 TestDataID,Rating
 2 1,3.0
 3 2,3.0
 4 3,3.0
 5 4,3.0
 6 5,3.0
 7 6,3.0
 8 7,3.0
 9 8,3.0
10 9,3.0
11 10,3.0
12 11,3.0
```

format: TestDataID,Rating

Deadline and Policy

Deadline

1. Kaggle: 6/6 23:59 (GMT+8)
2. Report and source code: 6/7 23:59 (GMT+8)

助教會在deadline一到就clone所有程式,

並且**不再重新clone任何檔案**

Note: clone指的是 git clone 不是 git lfs clone 也不是其他方法 !!!

Policy I - Repository

- github上ML2018/hw6/裡面請至少包含:
  - Report.pdf
  - hw6.sh
  - hw6_best.sh
  - your python files
  - your model files (can be loaded by your python file)
- **請不要上傳dataset**
- **hw6.sh 必須是MF的實作**
- 請將model download到與script相同的位置

Policy II – Source Code

- **Python Only**，請使用Python 3.5+, Keras 2.0.8, Tensorflow1.4.0, pytorch 0.3.0, h5py2.7.0., sklearn 0.19.1, numpy, pandas, Python Standard Lib.
- **只可使用限定的package，以及python內建的package，並且限定使用Tensorflow作為Keras的backend，不能使用 sklearn.ensemble.** 若import其他東西，或是使用不同版本，造成批改錯誤，將不接受修正。
- **請不要執行plot圖的code，並且不能在code裡import plot圖的套件。**
- 不能使用額外data來training (包括 pre-training)
- 不能call 其他線上 API (Project Oxford...)
- 請附上訓練好的model (及其參數)，hw6.sh 和 hw6_best.sh要在10分鐘內跑完

Policy II – Source Code

- 與之前作業相同，請在script中寫清楚使用python版本
- 以下的路徑，助教在跑的時候會另外指定，請保留可更改的彈性，不要寫死
  - Script usage:
    bash hw6.sh <test.csv path> <prediction file path> <movies.csv path> <users.csv path>
    bash hw6_best.sh <test.csv path> <prediction file path> <movies.csv path> <users.csv path>

Policy IV - Score

- Kaggle Rank
  - (0.8%) 超過public leaderboard的simple baseline分數
  - (0.8%) 超過public leaderboard的strong baseline分數
  - (0.8%) 超過private leaderboard的simple baseline分數
  - (0.8%) 超過private leaderboard的strong baseline分數
  - (0.8%) 5/30 23:59 (GMT+8)前超過public simple baseline
  - (BONUS) kaggle排名前五名(且願意上台跟大家分享的同學)
- 前五名排名以public平均為準，屆時助教會公布名單

小老師制度(手把手教學)

1. 在5/30以前超過simple baseline並願意在5/31在上課時間教導同學撰寫作業五程式，請填寫一下表單：https://drive.google.com/open?id=1lry7shC9Gu69H2A_l08la9SZRD8BBkSF-eMAPhe7qTw
2. 5/30將公布小老師名單在作業網頁，人數太多將以符合以下標準的同學為主:
   1. 沒有當過小老師
   2. Kaggle Public Leaderboard成績排名較高 (但請不要因此想overfit public set)
3. 小老師當次成績+1%

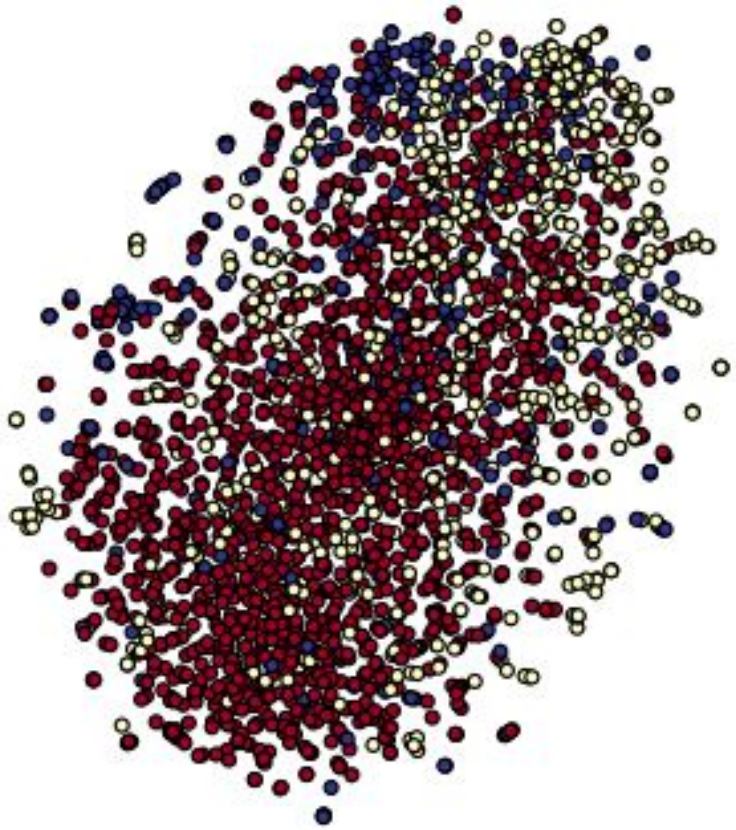
Policy IV - Score

- Report problem (PDF 3頁!)

1. (1 %)請比較有無normalize的差別。並說明如何normalize.
2. (1 %)比較不同的embedding dimension的結果。
3. (1 %)比較有無bias的結果。
4. (1 %)請試著將movie的embedding用tsne降維後，將movie category當作label來作圖。
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。

Movie category作圖範例

- 由於dataset給的movie分類有幾類滿像的，這張圖是把'Drama','Musical'作為一類，'Thriller','Horror','Crime'作為一類，Adventure,Animation,Children's做為一類所畫的圖。在畫圖時同學的類別可以自訂。**有多個分類時，可以隨機選擇一個。**
- 米色是'Drama','Musical',
  紅色是'Thriller','Horror','Crime'
  藍色是Adventure,Animation,Children's
- T-SNE教學



Score - Other Policy

- script檔名錯誤，直接0分。若是格式錯誤，請在公告時間內找助教修好，修完kaggle分數*0.7。

- script可以修改的範圍只包括:$1 $2 $3 $4的順序交換以及路徑修改。
- wget請自己維護好。
- Kaggle超過deadline直接shut down，可以繼續上傳但不計入成績。
- Github遲交一天(*0.7)，不足一天以一天計算，不得遲交超過兩天，有特殊原因請找助教。
- Github遲交表單
  :https://docs.google.com/forms/d/e/1FAIpQLSfkv3hLotNib4MK00tnySLzBDFklouHqLqY2I2g12V2stzThw/viewform(遲交才必需填寫)
  遲交請「先上傳程式」Github再填表單，助教會根據表單填寫時間當作繳交時間。

常犯錯誤！！！！

- script檔名錯誤。
- import 規定外的library
- 套件版本錯誤。
- 無法reproduce。
- 被抓到抄襲。
- kaggle多重帳號。

FAQ