U+ 아이들 나라 추천시스템 경진대회

사용툴	python
포트폴리오 형태	경진대회
투입기간	

추천시스템 경진대회

목표

- NCF 모델 성능 개선

사용 모델

1. Neural Collaborate Filtering

- Neural Collaborative Filtering논문의 모델을 구현
- MF + MLP 로 구성된 모델
- MF 모델의 한계인 다양한 변수를 활용하지 못한 다는 점을 극복하기 위해 deeplearing 모델을 활용
- deeplearning 모델은 MLP를 사용

2. MF

- userID, itemID 를 embedding
- label은 user가 본 item을 모두 1로 지정 아닌 item에 대해서는 0으로 label을 붙임
- 일반적인 MF 모델과는 달리

3. MLP

- userID, itemID, genre, age 피쳐 활용
- userID, itemID, genre는 embedding 을 통해 수치화
- age는 이미 수치데이터 이므로 그대로 활용
- [userID, itemID, genre, age] 형태로 input 데이터 구성
- mlp층은 linear, relu, dropout을 하나의 블록으로 해서 두 개의 층을 쌓음

4. MF+MLP

- MF 모델은 원래 userID_embedding과 itemID_embedding을 adamar product함
- MF 모델 output과 MLP 모델 output을 concat 해서 linear 층에 input해서 최종 결과를 도출함

개선 사항

1. itemID에 매칭되는 keyword데이터를 활용

- item을 index, keyword를 column으로 dataframe 생성
- value는 제공된 데이터에서 item과 keyword의 관련성을 0~5사이의 수치로 표현된 값을 value로 사용
- 모든 keyword가 item과 매칭되는 것이 아니므로 결측 데이터가 발생함
- 결측치는 value가 0인 keyword와 구분하기 위해서 -1로 값을 넣어줌

	18분 이상	1세	1차/2차 성징	2분 미만	2분 이상 10분 미만	2세	3세	4세	5세	5세 추천	...	환경	환경 문제	환경과 생활	환영하기	활동	회화
album_id																	
0	-1.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	0.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
1	-1.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
2	-1.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
3	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.0	0.0	0.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
4	-1.0	0.0	-1.0	-1.0	-1.0	0.0	0.0	0.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
39870	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
39871	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
39872	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
39873	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
39874	-1.0	-1.0	-1.0	-1.0	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	...	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0

- 모든 키워드에 대해서 k-means clustering 알고리즘을 사용해 유사한 item끼리 같은 그룹으로 묶음
- 클러스터를 MLP 피쳐로 사용함

2. MLP 모델의 층을 확장

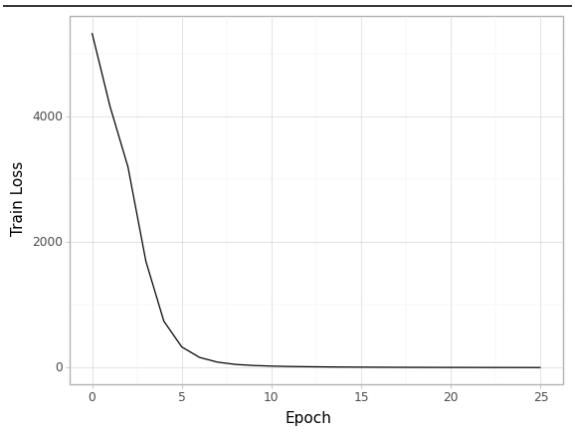
- 기존 두 개였던 층을 9개로 증축

3. 파라미터 수정

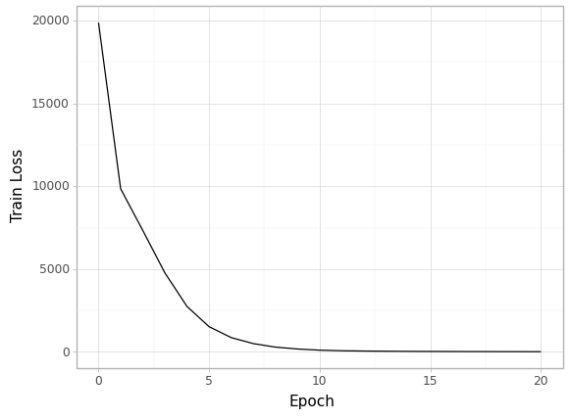
- embedding 차원을 512까지 증가
- 메모리 부족으로 batchsize는 256으로 낮춤

결과

- 기존의 베이스라인보다 높은 score가 나왔다.



baseline의 loss 그래프

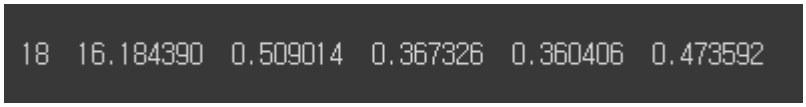


개선모델의 loss 그래프

- baseline의 validation 결과 (Epoch,Train Loss,Valid Recall@25,Valid NDCG@25,Valid Coverage,Valid Score)

```
18  4.857800  0.364911  0.257604  0.199925  0.338085
```

- 개선 모델의 validation 결과(Epoch,Train Loss,Valid Recall@25,Valid NDCG@25,Valid Coverage,Valid Score)



- loss는 조금 높지만 recall과 NDCG 점수가 확연히 높아진 것은 볼 수 있다.