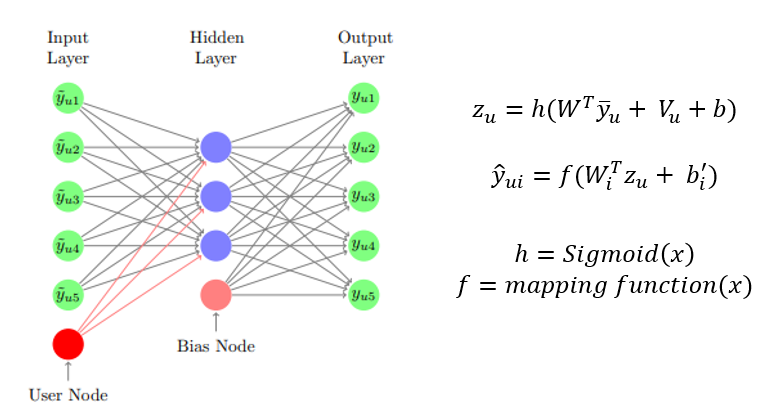
Posted by Contents&Education

2022.12.22.17:00

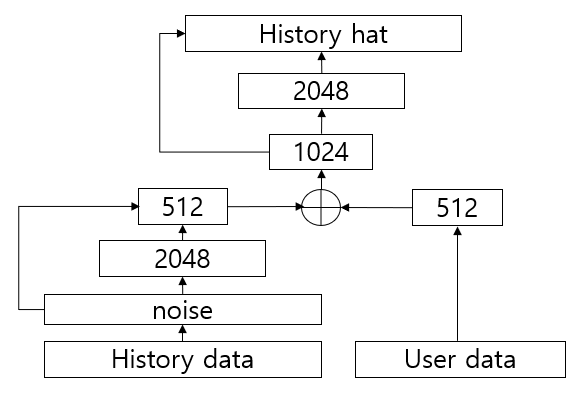
1 UP

사용한 모델

Collaborative Denoising Auto-Encoders for Top-N Recommender System 논문에 나온 구조를 커스텀하여 사용 (논문 링크)



(Collaborative Denoising Auto-Encoder)



(Custom model)

Layer 추가

Latent space : add -> concatenate

Skip connection

Layer Normalize

Activation function: Sigmoid -> Tanh

사용 데이터 및 전처리 방식

History data

history\_data.csv

중복시청 제거, 동일 log 타임 제거(base line과 동일)

유저가 시청한 item의 원핫 벡터로 변환 (25917 차원의 Sparse 벡터)

User data

Profile\_data.csv

범주형 변수 원핫 벡터에 연속형 변수를 추가하여 57 + 1 차원의 원핫 벡터로 변환(nan 값 포함)

훈련 방법

Batch size : 64

Optimizer : Adam

Learning rate : 0.0001

Epoch : 15

Dropout(noise) : 0.5

Cost function: Cross entropy

Check point : 유저를 기준으로 K-Fold를 사용하여 valid set에 대하여 평가 score가 가장 높은 것을 사용

인퍼런스

Fold별 모델(5개)에 대하여 train + valid 데이터셋을 inference함,

5개의 output에 sigmoid를 취한 값을 평균 낸 이후 높은 순서로 Top 25개를 추천 데이터로 삼음

기시청 item의 제거는 하지 않음

시나리오

일반적인 경우

유저의 모든 시청 데이터와 유저 데이터를 입력하면, 모델은 노이즈를 복원하도록 훈련되었으므로, 모든 시청 데이터가 노이즈가 있다고 판단, 이를 복원함으로 시청 했을 가능성이 높은 아이템들을 복원하게 된다.

새로운 유저가 등록되는 경우

노이즈를 통해 시청기록은 없어도 유저데이터만으로 시청이 복원된 경우가 있으므로 추천이 가능하다.

아이템이 추가되는 경우

이때에는 학습을 다시 해야함, 일정 주기에 모델을 재학습 하는 MLOPS를 고려할 필요가 있음.

다만 재학습의 코스트는 크지 않은 편 (GPU memory 4GB, 20분이내)

TODO

본 추천시스템의 목적은 일정 기간의 데이터를 가지고 이후에 볼 Top-N 개의 아이템을 추천하는 것에 있음

그렇기에 아이템의 시청 내역의 중복처리를 하지 않고 학습이 진행될 필요가 있음(동일한 것을 반복해서 보는 경우)

Noise의 경우 dropout을 통해 랜덤하게 매번 바뀌지만 그래서 모델이 잘 수렴 안되는 경우가 있음, 노이즈를 시간 별로 주는 방법을 사용해야함 (ex 시청내역의 앞에서 절반을 input으로 주고 전부를 복원하도록 학습)

시계열 예측인 부분이 있기에 LSTM Auto-Encoder를 테스트 해보아도 좋을 것 같음.