

데이터사이언스 Long Term Project 보고서

소프트웨어학부 2022041055 류정환

June 15, 2025

1 알고리즘 요약

본 보고서는 기존에 MovieLens 데이터셋에서 좋은 성능을 보인다고 알려진 GC-MC (Graph Convolutional Matrix Completion)¹ 알고리즘을 사용했습니다. 해당 논문은 추천 시스템의 matrix completion을 이분 그래프에서 link prediction 문제로 보고 GC-MC 모델을 제안합니다. GC-MC은 이분 사용자-아이템 그래프 위에서 미분 가능한 메시지 패싱을 기반으로 사용자 및 아이템 임베딩을 학습하는 graph auto-encoder 프레임워크를 사용하며 side information도 자연스럽게 통합합니다. 아래 그림은 해당 과정을 개념적으로 보여줍니다.

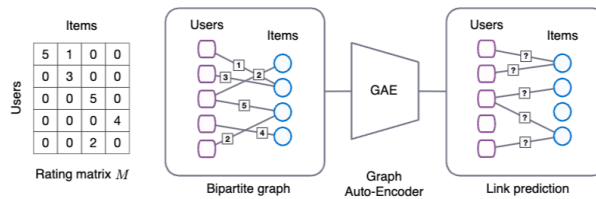


Figure 1: 왼쪽: 사용자-아이템 상호작용(1-5점 척도) 또는 결측치(0)에 해당하는 항목을 포함하는 평점 행렬 M . 오른쪽: 이분 구조를 갖는 사용자-아이템 상호작용 그래프. 엣지는 상호작용 이벤트를 나타내고, 엣지의 숫자는 사용자가 특정 아이템에 부여한 평점을 나타냅니다. 행렬 완성 작업(즉, 관찰되지 않은 상호작용에 대한 예측)은 링크 예측 문제로 간주될 수 있으며, 종단 간 학습 가능한 그래프 오토인코더를 사용하여 모델링할 수 있습니다.

또한 기존 논문에는 적용되지 않은 Learning Rate 스케줄링과 Early Stopping, Gradient Clipping과 같은 기법들을 사용하여 보다 학습이 원활하게 진행될 수 있도록 하고 성능을 높였습니다.

2 Graph Convolutional Matrix Completion Encoder

사용자와 아이템 임베딩은 그래프 컨볼루션을 통해 생성됩니다. 메시지 패싱과 누적(accumulation)을 수행하여 사용자-아이템 상호작용을 모델링합니다.

¹<https://arxiv.org/pdf/1706.02263v2>

2.1 Bilinear Decoder

논문의 “Section 2.3 (Bilinear Decoder)”에 해당하는 내용입니다. Bilinear Decoder는 사용자와 아이템 임베딩을 활용하여 평점(rating)을 예측합니다.

```
1 class BilinearDecoder(nn.Module):  
2     """Bilinear Decoder for rating prediction"""
```

- 논문의 수식 (4)와 (5)를 구현합니다.
- Bilinear 연산을 통해 평점 예측:

$$p(M_{ij} = r) = \frac{\exp(\mathbf{u}_i^\top \mathbf{Q}_r \mathbf{v}_j)}{\sum_s \exp(\mathbf{u}_i^\top \mathbf{Q}_s \mathbf{v}_j)}$$

- Basis matrices를 사용하여 가중치 공유 (수식 12).

3 GCMC 클래스

GCMC 클래스는 전체 모델 구조를 통합합니다.

```
1 class GCMC(nn.Module):  
2     """Graph Convolutional Matrix Completion Model"""
```

- Encoder와 Decoder를 통합.
- Forward pass에서 사용자-아이템 쌍의 평점을 예측.

4 GraphConvolutionalMatrixCompletion 클래스

학습 및 예측을 위한 전체 파이프라인을 구현합니다.

4.1 주요 메서드

4.1.1 create_adjacency_matrices()

- 논문의 “Section 2.5 (Vectorized implementation)”에 해당.
- 각 평점 타입별로 adjacency matrix 생성.
- Symmetric normalization 적용:

$$\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$$

4.1.2 prepare_features()

- 논문의 “Section 2.6 (Input feature representation and side information)”에 해당.
- 사용자 특징: 나이, 성별 등.
- 아이템 특징: 장르 (19차원 벡터).

4.1.3 fit()

- 논문의 “Section 2.4 (Model training)”에 해당.
- 손실 함수 (수식 6): Negative Log Likelihood.
- Node dropout 적용 (정규화).
- Mini-batching 지원.
- Learning rate scheduling 및 early stopping.

4.1.4 predict()

- 학습된 모델을 통해 사용자-아이템 쌍의 평점 예측.
- 보지 않은 사용자/아이템에 대한 기본 평점 처리.

5 데이터 로딩 함수

5.1 load_data()

- MovieLens 데이터셋의 평점 정보 로드.

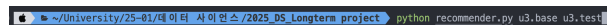
5.2 load_items()

- 아이템(영화) 정보와 장르 정보 로드.

5.3 load_user_info()

- 사용자 인구통계학적 정보 로드.

6 컴파일 방법



프로젝트 폴더에서 'python recommender.py <training data> <test data>'를 입력하시면 정상적으로 실행됩니다.