AutoRec: Autoencoders Meet Collaborative Filtering

https://users.cecs.anu.edu.au/~akmenon/papers/autorec/autorec-paper.pdf

0. Introduction

- 추천 시스템에서 사용자-아이템 평점 행렬의 희소성(sparsity) 문제는 주요 과제임.
- 기존 Matrix Factorization(MF) 기반 협업 필터링은 선형 관계만 학습 가능하다는 한계 존재.
- 본 논문은 Autoencoder 기반의 비선형 모델 AutoRec을 제안하여, 복잡한 사용자-아이템 관계를 모델링함.

1. Overview

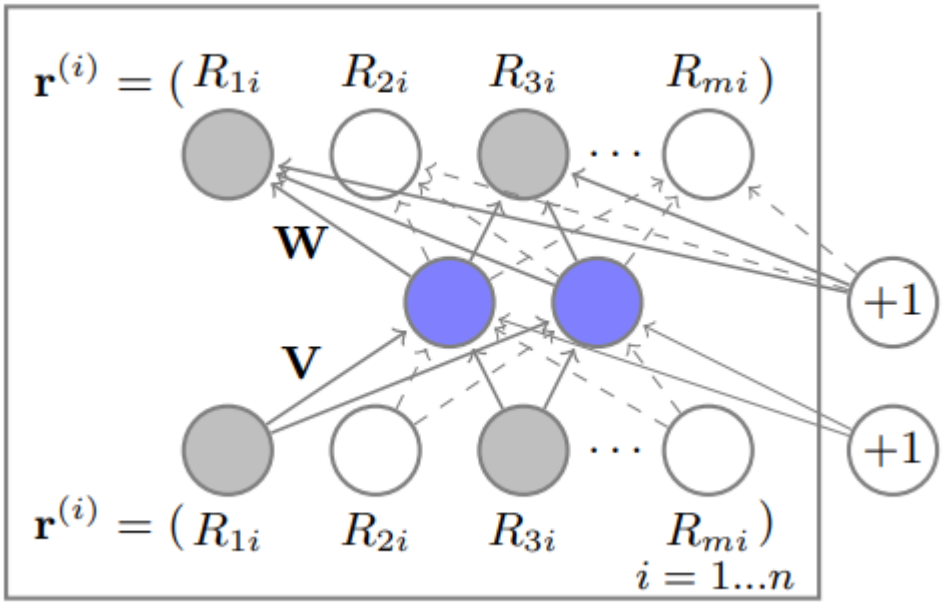
- AutoRec은 입력으로 사용자 혹은 아이템 벡터를 받아 자기 재구성(self-reconstruction) 을 수행하는 오토인코더 기반 구조임.
- 학습 목표는 입력 벡터의 관측된 부분을 기반으로 전체 벡터(누락된 값 포함) 를 재구성하는 것.
- 두 가지 변형이 존재함:
  - User-based AutoRec (U-AutoRec) : 사용자별 평점 벡터 입력
  - Item-based AutoRec (I-AutoRec) : 아이템별 평점 벡터 입력
- 기존 협업 필터링보다 RMSE 감소 및 일반화 성능 향상이 목표.

3. Challenges

- 데이터 희소성 : 대부분의 사용자-아이템 매트릭스는 95% 이상 결측치로 구성됨.
- 비선형 관계 학습의 어려움 : 기존 MF는 선형 조합으로만 예측 가능.
- 오버피팅 문제 : Autoencoder 학습 시 관측치 중심으로 과적합될 가능성 존재.

- 규모 확장성(scalability) : 대규모 데이터셋 학습 시 계산 비용이 높음.

4. Method



- AutoRec의 기본 구조:
  - 활성 함수 f, g 는 sigmoid 또는 ReLU 사용
  - 입력 :

$$r \in \mathbb{R}^N$$

  - 인코더 :

$$h = g(Vr + \mu)$$

  - 디코더 :

$$r\prime = f(Wh + b)$$

  - 손실 함수 (관측된 평점만 고려하여 계산) :

$$L = i \in observed \sum (ri - ri\prime)2 + \lambda(|| \ V \ || \ 2 + || \ W \ || \ 2)$$

- 학습 :
  - mini-batch gradient descent 기반
  - L2 정규화 + dropout 적용으로 오버피팅 방지

5. Experiments

- 데이터셋: MovieLens 1M, MovieLens 10M, Netflix
- 평가 지표: RMSE (Root Mean Square Error)
- 비교 대상 모델:
  - PMF (Probabilistic Matrix Factorization)
  - SVD++
  - RBM-CF (Restricted Boltzmann Machine CF)
- 설정:
  - hidden dimension: 500~1000
  - optimizer: SGD
  - activation: sigmoid
  - dropout: 0.1~0.5

6. Results

	ML-1M	ML-10M
U-RBM	0.881	0.823
I-RBM	0.854	0.825
U-AutoRec	0.874	0.867
I-AutoRec	**0.831**	**0.782**

(a)

$f(\cdot)$	$g(\cdot)$	RMSE
Identity	Identity	0.872
Sigmoid	Identity	0.852
Identity	Sigmoid	**0.831**
Sigmoid	Sigmoid	0.836

(b)

	ML-1M	ML-10M	Netflix
BiasedMF	0.845	0.803	0.844
I-RBM	0.854	0.825	-
U-RBM	0.881	0.823	0.845
LLORMA	0.833	**0.782**	0.834
I-AutoRec	**0.831**	**0.782**	**0.823**

(c)

- I-AutoRec 이 U-AutoRec 보다 consistently 높은 성능을 보임.
- Netflix 데이터셋 기준 RMSE:
  - PMF: 0.905
  - RBM-CF: 0.890

    - AutoRec: 0.883 (best)
- 비선형 구조 덕분에 복잡한 사용자-아이템 상호작용을 더 잘 학습함.
- hidden layer 수를 늘릴수록 초기엔 성능 향상 있으나 과적합 위험 증가.
- Ablation 실험 결과, 정규화 항 제거 시 RMSE 급상승 → regularization 중요함.

7. Insight

- AutoRec은 딥러닝을 추천 시스템에 성공적으로 도입한 초기 사례로 평가됨.
- 모델 구조가 단순하지만, 비선형 함수의 표현력으로 성능 향상을 달성.
- 이후 연구들(DeepAutoRec, CF-NADE, VAE-CF 등)의 기반이 됨.
- 실무 적용 측면에서는 데이터 희소성이 심한 환경에서 유리하지만, 대규모 학습 비용이 문제로 남음.
- 향후 발전 방향:
    - Variational Autoencoder(VAE) 기반 확장
    - Side information (메타데이터, 리뷰 등) 통합
    - 대규모 온라인 학습 환경 최적화