



fit@hcmus

Building A Recommendation System Based On Autoencoders

GVHD: ThS. Trần Trung Kiên

Mục Lục

1. Tổng quan
2. Mô hình VAEs cho Collaborative Filtering
3. Thực nghiệm
4. Kết luận

1. Tổng quan

1. Tổng quan

1.1 Giới thiệu

1.2 Các hướng tiếp cận

1.3 Autoencoder cho bài toán Recommendation

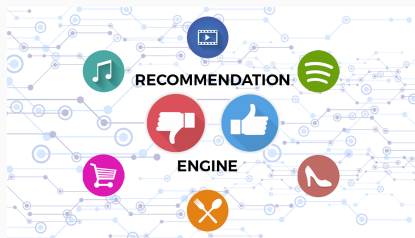
2. Mô hình VAEs cho Collaborative Filtering

3. Thực nghiệm

4. Kết luận

1.1. Giới thiệu

Hệ thống đề xuất được xây dựng để dự đoán những sản phẩm người dùng có thể thích, đặc biệt khi họ có nhiều lựa chọn.



1.1.1. Tại sao chúng ta nên xây dựng hệ thống gợi ý?

Recommendations worth a million



What is common in all these brands..?
They provide us with Recommendations...!

info@ivyproschoo.com 11 www.ivyproschoo.com


Copyright © Ivy Professional School - 2009-10 (All Rights Reserved)

- Hệ thống gợi ý sản phẩm là một lĩnh vực trong khai thác dữ liệu và máy học.
- Ngày nay, hệ thống gợi ý là một phần quan trọng không thể thiếu trong các doanh nghiệp.

Nguồn: Data Science Webinar: Recommender systems used by ecommerce companies

1.1.1. Tại sao chúng ta nên xây dựng hệ thống gợi ý?

Why Recommendation System are used?



NETFLIX -75% rented movies are from recommendation

Google -38% are more click-through are due to recommendation

amazon -35% are sales from recommendation

info@ivyproschoo.com 14 www.ivyproschoo.com

Copyright © Ivy Professional School - 2009-10 (All Rights Reserved)

1.1.2. Hệ thống gợi ý

Bài toán xây dựng hệ thống gợi ý sản phẩm (recommendation system) được phát biểu như sau:

- Cho input là lịch sử tương tác của người dùng (user) với các sản phẩm (item) hoặc có thêm các mô tả của sản phẩm.
- Yêu cầu: đưa ra tập các items (không có trong lịch sử) được dự đoán là phù hợp với người dùng.

Làm thế nào để xây dựng một hệ thống gợi ý?

1.2. Các hướng tiếp cận

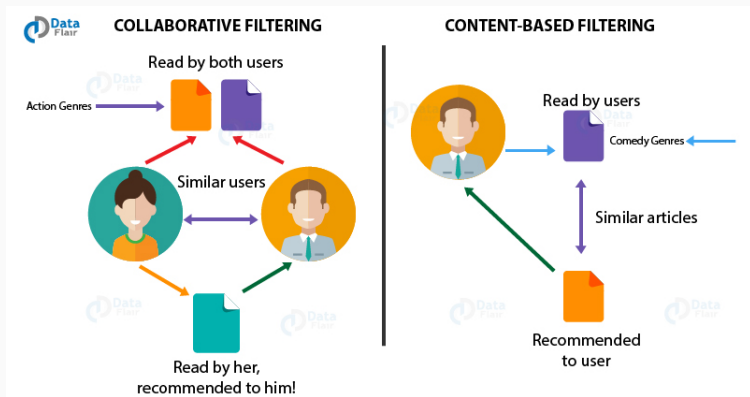


Figure: Approaches for Recommendation System

1.2. Các hướng tiếp cận

Content-based filtering

Ưu điểm

- Không cần dữ liệu từ các người dùng khác
- Mô hình có thể nắm bắt các sở thích của người dùng, có thể đề xuất các sản phẩm ít người quan tâm.

Nhược điểm

- Vì mô hình dựa trên sở thích của người dùng, mô hình hạn chế đối với các sản phẩm có ít thông tin.
- Cần các kiến thức nền để mô hình hóa dữ liệu.

1.2. Các hướng tiếp cận Collaborative Filtering

Ưu điểm

- Không cần các kiến thức nền tảng về lĩnh vực áp dụng
- Có thể giúp người dùng khám phá những sở thích mới
- Không cần các tính chất của sản phẩm

Nhược điểm

- Vấn đề cold-start, người dùng mới chưa có dữ liệu sẽ làm cho hệ thống khó đưa ra gợi ý
- Không thể đưa thêm vào mô hình các tính năng của sản phẩm

1.2. Các hướng tiếp cận

Maxtrix Factorization

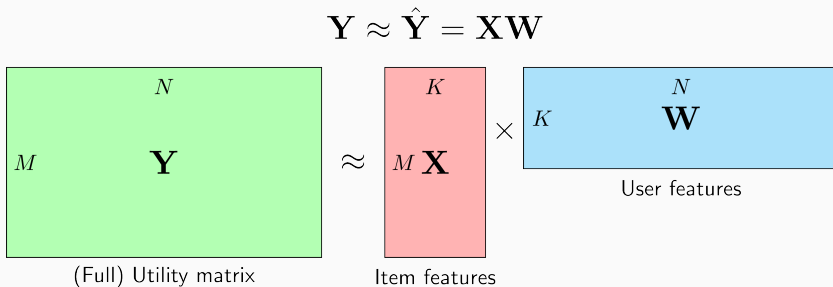


Figure: Approaches for Recommendation System

1.3. Autoencoder cho bài toán Recommendation

AutoRec

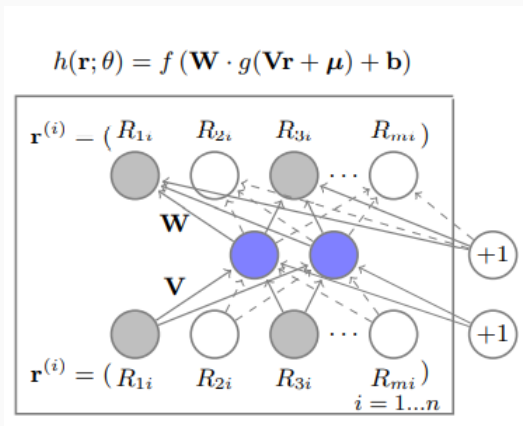


Figure: Mô hình AutoRec

1.3. Autoencoder cho bài toán Recommendation

Denoising Autoencoders: CDE

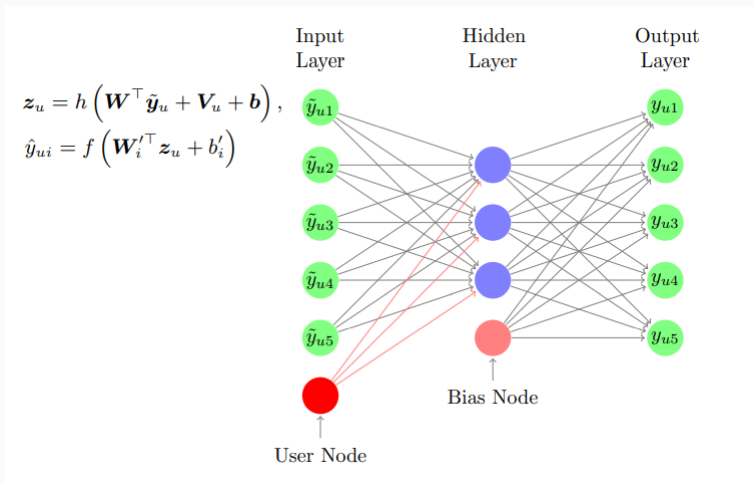


Figure: Mô hình CDAE

1.3. Autoencoder cho bài toán Recommendation

Variational Autoencoders: Mul-VAEs

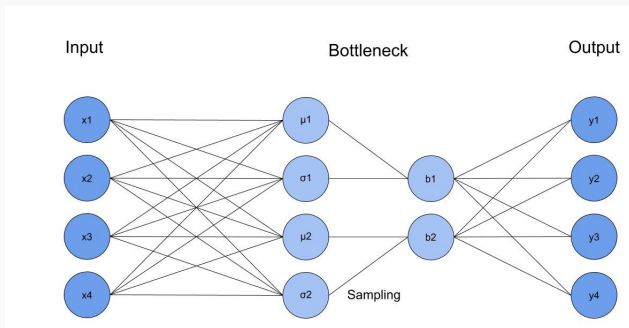


Figure: Mô hình Mul-VAE

1.3. Autoencoder cho bài toán Recommendation

Tại sao VAE phù hợp với bài toán Collaborative Filtering?

- Việc gợi ý sản phẩm được giả định trên việc tương tác của người dùng sẽ được phát sinh từ các đặc trưng ẩn.
- Thông thường người dùng chỉ tương tác với một lượng nhỏ item, do đó ta cần tìm ra những tương tác mới cho người dùng từ dữ liệu sẵn có.

2. Mô hình VAEs cho Collaborative Filtering

1. Tổng quan

2. Mô hình VAEs cho Collaborative Filtering

2.1 Mô hình Mul-VAE

2.2 Huấn luyện mô hình

3. Thực nghiệm

4. Kết luận

2.1. Mô hình Mul-VAE

- $X \in \mathbb{N}^U$ là ma trận thể hiện cho tập implicit feedback
- $x_u = [x_{u1}, \dots, x_{uI}] \in \mathbb{N}^I$ là vector tương tác của user u với tập I item.

		user			
		1	2	3	4
item	1	✓	?	?	✓
	2	✓	?	✓	✓
	3	?	✓	✓	?
	4	?	✓	?	✓

2.1. Mô hình Mul-VAE

- Generative process: Phát sinh tương tác cho người dùng

$$\mathbf{z}_u \sim \mathcal{N}(0, \mathbf{I}_K), \quad \pi(\mathbf{z}_u) \propto \exp\{f_\theta(\mathbf{z}_u)\}, \\ \mathbf{x}_u \sim \text{Mult}(N_u, \pi(\mathbf{z}_u)).$$

- Inference: "Suy diễn" đặc trưng ẩn từ dữ liệu người dùng.

$$p(\mathbf{z}|\mathbf{x}) \approx q_\theta(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu_\theta(\mathbf{x}), \sigma_\theta^2(\mathbf{x}))$$

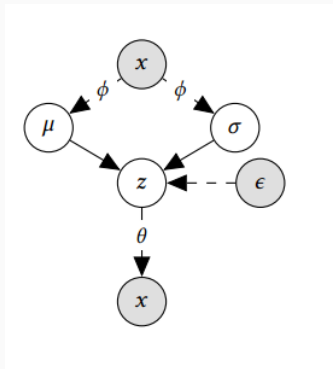


Figure: Variational Autoencoder

2.2. Huấn luyện mô hình

$$\mathcal{L}_u(\theta, \phi) = \mathbb{E}_{q_\phi(z_u|x_u)}[\log p_\theta(x_u|z_u)] - \beta \cdot KL(q_\phi(z_u|x_u) || p(z_u))$$

- Với bài toán recommendation, siêu tham số β được dùng để kiểm soát giữa việc mô hình dữ liệu và việc xấp xỉ phân phối xác suất theo giả định của mô hình.

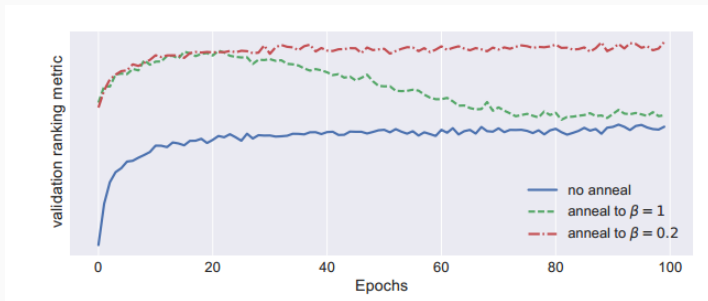
2.2.1. Multinomial Log Likelihood

Tại sao Multinomial lại hiệu quả với bài toán recommendation

$$\log p_{\theta}(\mathbf{x}_u | \mathbf{z}_u) \stackrel{c}{=} \sum_i x_{ui} \log \pi_i(\mathbf{z}_u).$$

Multinomial likelihood phù hợp với bài toán recommendation, khi đó mô hình thể hiện được xác suất được chọn giữa các item. Vì tổng $\pi(\mathbf{u})$ là 1, do đó các item sẽ phải cạnh tranh với nhau để có được xác suất được chọn cao hơn.

2.2.2. Kỹ Thuật KL-annealing



KL-annealing là một kỹ thuật lựa chọn siêu tham số dựa trên heuristic bằng cách ta sẽ tăng dần β từ 0 lên 1 trong quá trình huấn luyện để tìm ra được giá trị β tốt nhất trên tập validation.

3. Thực nghiệm

1. Tổng quan

2. Mô hình VAEs cho Collaborative Filtering

3. Thực nghiệm

3.1 Dữ liệu

3.2 Độ đo

3.3 Thực nghiệm

3.4 So sánh với các mô hình Baselines

4. Kết luận

3.1. Dữ liệu

Trong bài toán Collaborative Filtering, có 2 loại dữ liệu chính thường được dùng là:

- Explicit feedback: người dùng thể hiện sự yêu thích của họ đối với sản phẩm bằng một con số cụ thể.
- Implicit feedback: người dùng thể hiện sự yêu thích của họ đối với sản phẩm thông qua tương tác.

3.1. Dữ liệu



3.1. Dữ liệu

	MovieLens	Netflix	MSD
# of users	136,677	463,435	571,355
# of items	20,108	17,769	41,140
# of interactions	10.0M	56.9M	33.6M
% of interactions	0.36	0.69	0.14
# of held-out users	10,000	40,000	50,000

3.2. Độ đo

Recall@k

Là độ đo thể hiện tỉ lệ giữa số sản phẩm người dùng thật sự thích trong số top-k sản phẩm được gợi ý.

DCG@k

Là độ đo dùng để đánh giá chất lượng xếp hạng của mô hình, thường được dùng để đánh giá các thuật toán xếp hạng.

3.2. Độ đo

Recall

$$\text{Recall@R}(u, \omega) = \frac{\sum_{r=1}^R \mathbb{I}[\omega(r) \in I_u]}{\min(R, |I_u|)}$$

DCG@k

$$\text{DCG@R}(u, \omega) = \sum_{r=1}^R \frac{2^{\mathbb{I}[\omega(r) \in I_u]} - 1}{\log(r + 1)}$$

3.2. Độ đo

NDCG@k

$$\text{NDCG@R}(u, \omega) = \frac{\text{DCG@R}(u, \omega)}{\text{IDCG@R}}$$

where IDCG@R is ideal DCG@R.

Các siêu tham số và kiến trúc mô hình

3.3. Thực nghiệm

- Laten representation với 200 perceptron.
- Các hidden layer khác với 600 perceptron.
- hàm kích hoạt: tanh

4. Kết luận

1. Tổng quan
2. Mô hình VAEs cho Collaborative Filtering
3. Thực nghiệm
4. Kết luận