АНАЛІЗ АЛГОРИТМІВ МАТРИЧНОЇ ФАКТОРИЗАЦІЇ ДЛЯ РЕКОМЕНДАЦІЙНИХ СИСТЕМ

Онищенко М.Г., Шубін І.Ю.

Кафедра програмної інженерії,
Харківський національний університет радіоелектроніки,
Україна

E-mail: mykyta.onyshchenko@nure.ua,
igor.shubin@nure.ua

Abstract

This report presents a formal theoretical analysis of model-based collaborative filtering, focusing on the paradigm of matrix factorization. We derive and contrast the mathematical formulations of five seminal algorithms: Singular Value Decomposition (SVD) for explicit ratings, SVD++ for the integration of implicit signals, Non-negative Matrix Factorization (NMF) for interpretable factors, Alternating Least Squares (ALS) and Neural Collaborative Filtering (NCF). Furthermore, we provide a formal definition and conceptual analysis of key evaluation metrics, bifurcated into prediction accuracy (RMSE, MAE). The primary objective is to establish a conceptual foundation for the comparative study of these algorithms.

Основна суть задачі рекомендації полягає у фільтрації інформації з великого набору об'єктів для передбачення уподобань користувача. Математично це подається у вигляді матриці взаємодій «користувач–об'єкт» $R \in \mathbb{R}^{m \times n}$, де m – кількість користувачів, а n – кількість об'єктів [1]. Тоді елемент r_{ui} позначає взаємодію (наприклад, оцінку) користувача u з об'єктом i. У реальних системах ця матриця є надзвичайно великою та розрідженою, тобто переважна більшість її елементів є невідомими. Метою рекомендаційної системи є оцінка значень цих пропущених елементів.

Колаборативна фільтрація – це методологія, що формує рекомендації на основі принципу, за яким схожі користувачі мають подібні вподобання [2]. На відміну від контентного фільтрування, яке потребує явних ознак об'єктів (наприклад, жанру), колаборативна фільтрація використовує лише матрицю взаємодій «користувач-об'єкт»

Однією з головних техологій реалізації колаборативної фільтрації виступає матрична факторизація. Вона базується на припущенні, що взаємодії «користувач-об'єкт» обумовлені невеликою кількістю неявних (латентних) факторів [3]. Наприклад, факторами фільму можуть бути «серйозність проти ескапізму» або «орієнтований на чоловічу аудиторію», а факторами користувача – його схильність до тих самих концепцій.

Метою матричної факторизації є знаходження двох матриць нижчого рангу: матриці користувач–фактор $P \in R^{m \times k}$ та матриці об'єкт-фактор $Q \in R^{n \times k}$, де k – розмірність латентного простору, гіперпараметр за умови $k \ll m, n$ [1].

Рядок p_u з P є k-вимірним вектором, що представляє користувача u. Рядок q_i з Q є k-вимірним вектором, що представляє об'єкт i. Передбачувана взаємодія $\hat{r}ui$ для користувача (u) та об'єкта (i) моделюється як скалярний добуток їхніх латентних векторів:

$$\hat{r}ui = p_u \cdot q_i = p_u^T q_i = \sum_{f=1}^{k} p_{uf} q_{if}$$

Повна матриця передбачень $\hat{R}$ утворюється як добуток $\hat{R} = PQ^T$. Це обмеження низького рангу k діє як інформаційне «вузьке місце», змушуючи модель навчитися стислому, узагальненому представлення даних замість простого запам'ятовування взаємодій. Саме це примусове стиснення дозволяє передбачати пропущені елементи.

Матриці неявних факторів P та Q навчаються шляхом мінімізації функції цілі [2]. Для явного відгуку це зазвичай регуляризована сума квадратів помилок по множині спостережуваних оцінок $\mathcal{K}$. Функція цілі L має вигляд:

$$L = \sum_{(u,i)\in\mathcal{K}} \left(r_{ui} - (\mu + b_u + b_i + p_u^T q_i)\right)^2 + \lambda\left(\sum_u |p_u|^2 + \sum_i |q_i|^2 + \sum_u b_u^2 + \sum_i b_i^2\right)$$

$$L(P,Q) = \sum_{(u,i)\in\mathcal{K}} (r_{ui} - p_u^T q_i)^2 + \lambda\left(\sum_u |p_u|^2 + \sum_i |q_i|^2\right)$$

Ця функція складається з двох частин: суми квадратів помилок SSE для відомих оцінок, яка вимірює точність передбачення, та члена L_2-регуляризації (Фробеніуса), керованого параметром λ, що штрафує велику норму латентних векторів з метою запобігання перенавчанню.

Для удосконалення матричної факторизації використовується алгоритм сингулярної декомпозиції (SVD). Водночас слід зауважити, що мова йде не по класичний детермінований сингулярний розклад матриці з лінійної алгебри, який не визначено для розріджених матриць. Натомість йдеться про модель матричної факторизації, популяризовану Саймоном Фанком. Така модель навчається за допомогою ітеративної оптимізації (наприклад, стохастичного градієнтного спуску) [2].

Ця модель удосконалює базову матричну факторизацію шляхом введення зміщень, що враховують базові відмінності в даних. Повна модель передбачення $\hat{r}ui$ має вигляд*:*

$$\hat{r}ui = \mu + b_u + b_i + p_u^T q_i$$

де,

μ - глобальне середнє значення оцінок;

b_u - користувацьке зміщення (наприклад, «поблажливий» користувач, який ставить оцінки на 0.5 вищі за середні);

b_i - зміщення об'єкта (наприклад, «популярний» фільм, який оцінюється на 0.3 вище за середнє).

Ці зміщення відображають просту базову поведінку, тоді як член (p_u^T q_i) моделює справжню, тонку відповідність між користувачем і об'єктом. Відповідна функція цілі мінімізує помилку цього нового передбачення:

$$L = \sum_{(u,i)\in\mathcal{K}} \left(r_{ui} - (\mu + b_u + b_i + p_u^T q_i)\right)^2 + \lambda\left(\sum_u |p_u|^2 + \sum_i |q_i|^2 + \sum_u b_u^2 + \sum_i b_i^2\right)$$

Існує також алгоритм SVD++. Він є розширенням регуляризованої моделі SVD, яке покращує передбачення явних оцінок шляхом урахування сигналу від неявного відгуку [4]. Воно ґрунтується на припущенні, що сам факт взаємодії з об'єктом (незалежно від оцінки) несе інформацію про користувача.

Модель вводить другий набір латентних векторів об'єктів $Y \in R^{n \times k}$. Нехай $N(u)$ – множина об'єктів, з якими користувач (u) мав будь-яку взаємодію. Представлення користувача модифікується включенням фактора, отриманого з цієї множини.

Модель передбачення SVD++ має вигляд:

$$\hat{r}ui = \mu + b_i + b_u + q_i^T \left(p_u + |N(u)|^{-\frac{1}{2}} \sum j \in N(u)\, y_j \right)$$

У цій моделі користувач визначається не лише власним вектором явних уподобань p_u, але й цими уподобаннями, доповненими нормалізованою сумою неявних латентних факторів усіх об'єктів, з якими він взаємодіяв. Це багатша модель, що поєднує як явні, так і неявні сигнали.

Існує також алгоритм невід'ємної матричної факторизації. Він виконує ту саму факторизацію $R \approx PQ^T$, але з принциповим доповненням у вигляді обмежень невід'ємності: $P \geq 0$ та $Q \geq 0$ [5].

Це обмеження фундаментально змінює природу латентних факторів. Фактори в SVD можуть містити як додатні, так і від'ємні значення, утворюючи абстрактну «субтрактивну» модель, яку складно інтерпретувати. На відміну від цього, NMF формує репрезентацію, побудовану з частин.

Оскільки фактори можуть бути лише невід'ємними, модель є суто адитивною. Так, наприклад, фільм стає адитивною сумою своїх жанроподібних факторів (наприклад, $0.7 \times action + 0.2 \times comedy$), а користувач – сумою власних інтересів. Це робить отримані фактори високорозбірливими.

Компроміс полягає в точності передбачення: хоча NMF демонструє високу інтерпретованість і може працювати добре за умов екстремальної розрідженості матриці, методи на основі SVD зазвичай вважаються кращими з точки зору чистої точності. NMF зазвичай оптимізується шляхом мінімізації норми Фробеніуса або KL-дивергенції з використанням правил мультиплікативного оновлення. Ми не розглядатимемо їх у цій роботі [6].

Наступним розглянемо методи змінних найменших квадратів (ALS**)**. Це оптимізаційна техніка для розв'язання цільової функції матричної факторизації.Стандартна цільова функція $L(P, Q)$ не є спільно опуклою щодо P і Q. Однак, якщо фіксувати Q, ця функція перетворюється на опуклу квадратичну задачу щодо P, і навпаки [7].

ALS працює ітеративно: спочатку алгоритм фіксує Q й аналітично розв'язує задачу щодо P, мінімізуючи L, а потім фіксує оновлене P й аналітично розв'язує задачу щодо Q, мінімізуючи L.

ALS використовують для роботи з неявним зворотнім зв'язком, що розв'язує проблему «не спостережуване ≠ негативне». Ця модель вводить бінарну перевагу p_{ui} (1, якщо користувач u взаємодіяв з елементом i, 0 — інакше) та значення довіри c_{ui} для кожної пари (u, i): $c_{ui} = 1 + \alpha r_{ui}$ для спостережуваних взаємодій (довіра зростає зі збільшенням кількості взаємодій r_{ui}); $c_{ui} = w_0$ (малий базовий ваговий коефіцієнт) для всіх неспостережуваних значень [8].

Цільова функція набуває вигляду зваженої суми квадратів похибок для всіх значень:

$$\min_{P,Q} \sum_{u,i} c_{ui} (p_{ui} - p_u^T q_i)^2 + \lambda \left(\sum_u |p_u|^2 + \sum_i |q_i|^2 \right)$$

Нарешті, розглянемо нейронну колаборативну фільтрацію (NCF). Вона виступає загальною структурою, що використовує глибокі нейронні мережі (Deep Learning) для узагальнення матричної факторизації. Його основна передумова полягає в тому, що скалярний добуток $p_u^T q_i$ є простою, фіксованою функцією, яка за своєю суттю лінійна і тому обмежує виразну здатність моделі. NCF замінює цей скалярний добуток гнучкою, нелінійною нейронною мережею, здатною вивчати довільну функцію взаємодії користувача й об'єкта [9].

Архітектура NCF складається з таких компонентів: вхідний шар – це вектори для користувача u та об'єкта i; шар вкладень – це повнозв'язний шар, що проєктує розріджені вектори у щільні латентні представлення p_u та q_i. Цей шар ідентичний стандартній матричній факторизації.

Шари в NCF – це ключове нововведення. У межах NCF запропоновано дві підмоделі, які об'єднуються в кінцевій моделі нейронної матричної факторизації: узагальнена матрична факторизація (GMF) - **це** покомпонентний добуток вкладень $p_u \odot q_i$, що по суті відтворює лінійну взаємодію, аналогічну стандартній матричній факторизації; багатошаровий перцептрон (MLP**)** - це вкладення конкатенуються, $[p_u, q_i]$, і подаються до звичайної багатошарової нейронної мережі для вивчення складних нелінійних взаємодій [9].

Але будь-яку модель матричної факторизації, навчену за допомогою кожного з розглянутих алгоритмів, необхідно якісно оцінити. Для цього традиційно використовують величину кореня середньоквадратичної похибки (RMSE). Ми також розглянемо середню абсолютну похибку (MAE) в якості доповнення.

$$RMSE = \sqrt{\frac{1}{|\mathcal{K}|}\sum_{(u,i)\in\mathcal{K}}(r_{ui} - \widehat{r_{ui}})^2}$$

$$MAE = \frac{1}{|\mathcal{K}|}\sum_{(u,i)\in\mathcal{K}}|r_{ui} - \widehat{r_{ui}}|$$

Ключова відмінність між двома метриками полягає в тому, що MAE зважує всі помилки однаково, надаючи цілісне уявлення про середню помилку. Водночас RMSE, завдяки піднесенню помилки до квадрату, сильно збільшує вагу великих помилок, тому є чутливішим до викидів [10].

Література

1. 21.3. Matrix Factorization – Dive into Deep Learning 1.0.3. [Електронний ресурс] – URL: http://d2l.ai/chapter_recommender-systems/mf.html (дата посилання: 06.11.2025).
2. Using singular value decomposition for matrix factorization [Електронний ресурс] – URL: https://robinwitte.com/wp-content/uploads/2019/10/RecommenderSystem.pdf (дата посилання: 06.11.2025).
3. Matrix Factorization for Collaborative Filtering [Електронний ресурс] – URL: https://www.cs.utexas.edu/~ans/pubs/hintz_f15.pdf (дата посилання: 06.11.2025).
4. SVD++ Recommendation Algorithm Based on Backtracking [Електронний ресурс] – URL: https://www.mdpi.com/2078-2489/11/7/369 (дата посилання: 06.11.2025).
5. Non Negative Matrix Factorization in Recommender Models [Електронний ресурс] – URL: https://www.researchgate.net/publication/397477792 (дата посилання: 06.11.2025).
6. Recommender Systems with Non-negative Matrix Factorization [Електронний ресурс] – URL: https://oulurepo.oulu.fi/bitstream/handle/10024/55049 (дата посилання: 06.11.2025).
7. ALS Implicit Collaborative Filtering [Електронний ресурс] – URL: https://medium.com/radon-dev/als-implicit-collaborative-filtering (дата посилання: 06.11.2025).
8. Fast Matrix Factorization for Online Recommendation [Електронний ресурс] – URL: https://arxiv.org/pdf/1708.05024 (дата посилання: 06.11.2025).
9. Neural Collaborative Filtering [Електронний ресурс] – URL: https://arxiv.org/abs/1708.05031 (дата посилання: 06.11.2025).
10. Machine Learning Metrics and Business [Електронний ресурс] – URL: https://neptune.ai/blog/recommender-systems-metrics (дата посилання: 06.11.2025).