NGUYỄN THỊ BÍCH NGỌC

PHƯƠNG PHÁP NHÂN TỬ HÓA MA TRẬN KHÔNG ÂM (NMF) VÀ ỨNG DỤNG

HÀ NỘI - 2024

ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN

NGUYỄN THỊ BÍCH NGỌC

PHƯƠNG PHÁP NHÂN TỬ HÓA MA TRẬN KHÔNG ÂM (NMF) VÀ ỨNG DỤNG

ĐỒ ÁN TỐT NGHIỆP
Chuyên ngành : Toán - Tin

HÀ NỘI, THÁNG 12/2024

ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN

ĐỒ ÁN TỐT NGHIỆP

Chuyên ngành : Toán - Tin

PHƯƠNG PHÁP NHÂN TỬ HÓA MA TRẬN KHÔNG ÂM (NMF) VÀ ỨNG DỤNG

Giảng viên hướng dẫn: TS. Phạm Thị Hoài

Sinh viên thực hiện: Nguyễn Thị Bích Ngọc

MSSV: 20185388

Lớp: Toán - Tin 01

HÀ NỘI, THÁNG 12/2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục đích và nội dung của đồ án

.....

.....

.....

.....

.....

2. Kết quả đạt được

.....

.....

.....

.....

3. Ý thức làm việc của sinh viên

.....

.....

.....

Hà Nội, ngày tháng năm 2024

Giảng viên hướng dẫn

(Ký và ghi rõ họ tên)

TS. Phạm Thị Hoài

Mục lục

Danh sách hình vẽ

Danh sách bảng

Một số ký hiệu và chữ viết tắt

NMF	Nhân tử hóa ma trận không âm
$\mathbf{x}, \mathbf{y}$	in đậm, chữ thường, là các vector
$\mathbf{A}, \mathbf{B}$	in đậm, chữ hoa, là các ma trận
$\mathbb{R}$	tập hợp các số thực
$\mathbb{R}^n$	tập hợp các vector thực có n phần tử
$\mathbb{R}^{m\times n}$	tập hợp các ma trận thực có m hàng, n cột
$\mathbb{R}_+^{m\times n}$	tập hợp các ma trận không âm cỡ $m \times n$.
$\mathbf{A}_{i:}$	hàng thứ i của ma trận $\mathbf{A}$
$\mathbf{A}_{:j}$	cột thứ j của ma trận $\mathbf{A}$
$a_{ij}, [\mathbf{A}]_{ij}$	phần tử hàng thứ i, cột thứ j của ma trận $\mathbf{A}$
$\mathbf{A}^\top$	ma trận chuyển vị của ma trận $\mathbf{A}$
$\mathbf{A}^{-1}$	ma trận nghịch đảo của ma trận khả nghịch $\mathbf{A}$
$\mathbf{I}_n$	ma trận đơn vị cấp n
$\langle \mathbf{x}, \mathbf{y} \rangle$	tích vô hướng của hai vector $\mathbf{x}, \mathbf{y}$
$\langle \mathbf{A}, \mathbf{B} \rangle$	tích vô hướng của hai ma trận $\mathbf{A}, \mathbf{B}$
$\|\mathbf{x}\|$	chuẩn Euclid của vector $\mathbf{x}$
$\|\mathbf{A}\|_F$	chuẩn Frobenius của ma trận $\mathbf{A}$
$vec(\mathbf{A})$	vector hóa của ma trận $\mathbf{A}$
$F(M, x^*)$	tập các hướng chấp nhận được của tập M tại x^*
$f'(x^*, d)$	đạo hàm theo hướng của hàm f theo hướng d tại x^*
$\nabla f(x^*)$	vector gradient của hàm f tại điểm x^*
$intM$	phần trong của tập M
$\mathbb{L}$	hàm Lagrange

Lời mở đầu

Trong thời đại hiện nay, dữ liệu đóng một vai trò vô cùng quan trọng trong mọi lĩnh vực của cuộc sống. Cứ mỗi giây trôi qua, hàng tỉ thông tin khác nhau được tạo ra và chia sẻ bởi người dùng internet, từ hình ảnh, video cho đến kinh nghiệm du lịch, mua sắm, và nhiều hơn nữa. Việc khai thác và sử dụng những thông tin này trở thành một vấn đề thu hút sự quan tâm của rất nhiều người. Một trong những phương pháp hiệu quả để khai thác dữ liệu là giảm độ phức tạp của chúng mà vẫn giữ lại những yếu tố cần thiết. Để nghiên cứu các loại dữ liệu khác nhau, người ta cũng cần áp dụng các mô hình phù hợp nhằm thu được thông tin riêng biệt từ từng loại dữ liệu.

Bài toán **nhân tử hóa ma trận không âm (Non-negative Matrix Factorization - NMF)** là một trong những bài toán quan trọng trong lĩnh vực xử lý tín hiệu và đại số tuyến tính. Nó đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau như khai thác văn bản (Paul Pauca và cộng sự, 2004) [1], nhận dạng hệ thống con (Kim Philip và Tidor, 2003) [2], phát hiện lớp ung thư (Kim và Park, 2007) [3], xử lý hình ảnh thiên văn (Richardson, 1972) [4], sinh học thần kinh - phân tách gen (Rao và Shepherd, 2004) [5] và phân tích dữ liệu - nhận dạng mẫu, phân đoạn, phân cụm, giảm chiều dữ liệu (Spratling, 2006) [6], ... Mục tiêu chính của bài toán là phân tích một ma trận không âm $\mathbf{A} \in \mathbb{R}^{m \times n}$ thành hai ma trận không âm $\mathbf{U} \in \mathbb{R}^{m \times r}$ và $\mathbf{V} \in \mathbb{R}^{r \times n}$ với số nguyên dương $r \leq \min(m, n)$ sao cho tích $\mathbf{UV}$ xấp xỉ với ma trận $\mathbf{A}$.

Kể từ khi NMF được đề xuất lần đầu bởi Paatero và Tapper (1994) [7] và đã có nhiều phương pháp khác nhau được phát triển để giải quyết bài toán này, chẳng hạn như quy tắc cập nhật nhân (Multiplicative Update Rules - MUR) do Lee và Seung giới thiệu vào năm 1999 và 2001 [8], phương pháp giảm dần độ dốc (Gradient Descent - GD) bởi Chu và cộng sự vào năm 2004 [9], cũng như bình phương tối thiểu không âm xen kẽ (Alternating Non-negative Least Squares - ANLS) do Paatero và Tapper phát triển, ...

Báo cáo đồ án này sẽ tập trung nghiên cứu quy tắc cập nhật nhân (MUR) để giải bài toán NMF và ứng dụng nó trên bộ dữ liệu Movielens 100K. Mục tiêu của việc này là phân tích và tìm ra các yếu tố ảnh hưởng đến sở thích của người dùng, từ đó xây dựng một hệ thống gợi ý phim hiệu quả.

Nội dung của đề tài gồm có ba chương được trình bày như sau:

- **Chương 1: Một số kiến thức cơ sở**
  Trình bày các khái niệm cơ bản trong đại số tuyến tính và lý thuyết tối ưu, bao gồm các loại ma trận và điều kiện tối ưu.
- **Chương 2: Thuật toán giải bài toán NMF**
  Phát biểu bài toán NMF, trình bày điều kiện cần tối ưu, tìm hiểu về quy tắc cập nhật nhân (MUR) để giải bài toán và sự hội tụ của thuật toán.
- **Chương 3: Lập trình và ứng dụng của NMF**
  Trình bày bộ dữ liệu Movielens 100K, quy trình xây dựng mô hình NMF, đánh giá mô hình và ứng dụng trong việc đưa ra danh sách các bộ phim gợi ý cho người dùng.

Nhân đây, em cũng xin chân thành cảm ơn cô **TS. Phạm Thị Hoài** đã tận tình hướng dẫn, chỉ dạy giúp đỡ em trong suốt thời gian thực hiện đề tài nghiên cứu này.

Đề tài được thực hiện trong một thời gian tương đối ngắn, nên dù em đã hết sức cố gắng hoàn thành đề tài nhưng chắc chắn sẽ không thể tránh khỏi những thiếu sót nhất định. Rất mong nhận được sự thông cảm và đóng góp những ý kiến vô cùng quý báu của các Thầy/Cô, bạn bè để hoàn thiện đề tài này hơn nữa nhằm tạo tiền đề thuận lợi cho việc phát triển đề tài trong tương lai.

Sinh viên thực hiện

Nguyễn Thị Bích Ngọc

Chương 1

Một số kiến thức cơ sở

Trong chương này sẽ trình bày lại một số khái niệm về đại số tuyến tính như tích Hadamard, chuẩn của vector, chuẩn của ma trận, ma trận không âm,.... Bên cạnh đó, cũng sẽ trình bày lại một số khái niệm và kết quả cơ bản trong lý thuyết tối ưu để phục vụ các chương sau như tập lồi và hàm lồi, điều kiện tối ưu, điều kiện Karush-Kuhn-Tucker (KKT),.... Nội dung của chương được tham khảo chủ yếu từ các tài liệu [10] và [11].

1.1 Đại số tuyến tính

1.1.1 Một số ma trận cơ bản, tích vô hướng và tích Hadamard

Cho $\mathbf{A}$ là một ma trận cỡ $m \times n$ với phần tử ở hàng thứ i và cột thứ j là a_{ij} hoặc $[\mathbf{A}]_{ij}$. Khi đó, ta viết: $\mathbf{A} = (a_{ij})_{m\times n}$ với $i = \overline{1,m}$; $j = \overline{1,n}$. Ta kí hiệu hàng thứ i của ma trận $\mathbf{A}$ bởi $\mathbf{A_{i:}}$ và cột thứ j của ma trận $\mathbf{A}$ bởi $\mathbf{A_{:j}}$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Ma trận không là ma trận có tất cả các phần tử đều bằng 0, tức $a_{ij} = 0\ \forall i, j$.

Ma trận có n hàng và n cột được gọi là ma trận vuông cấp n.

$\mathbf{A}$ là ma trận đường chéo nếu $\mathbf{A}$ là ma trận vuông có $a_{ij} = 0$ với mọi $i \neq j$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{mn} \end{bmatrix}$$

Ma trận đơn vị là ma trận chéo có các phần tử trên đường chéo đều bằng 1. Ký hiệu: $\mathbf{E}, \mathbf{E}_n$ (hoặc $\mathbf{I}, \mathbf{I}_n$).

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Ma trận chuyển vị của ma trận $\mathbf{A}$ được kí hiệu là $\mathbf{A}^\top$ và xác định $\mathbf{A}^\top = (b_{ij})_{n\times m}$ với $b_{ij} = a_{ji}$.

$$\mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

$\mathbf{A}$ được gọi là ma trận đối xứng nếu $\mathbf{A} = \mathbf{A}^\top$.

Ma trận $\mathbf{A}$ vuông cấp n được gọi là ma trận trực giao nếu $\mathbf{A}^\top\mathbf{A} = \mathbf{I_n}$.

Ma trận $\mathbf{A}$ vuông cấp n là ma trận khả nghịch nếu tồn tại ma trận $\mathbf{B}$ sao cho $\mathbf{AB} = \mathbf{BA} = \mathbf{I_n}$. Khi đó, $\mathbf{B}$ gọi là ma trận nghịch đảo của ma trận $\mathbf{A}$, kí hiệu là $\mathbf{A}^{-1}$.

Ma trận đường chéo $\mathbf{A}$ có ma trận nghịch đảo $\mathbf{A}^{-1}$ tồn tại khi và chỉ khi tất cả các phần tử trên đường chéo của nó khác không.

Ma trận $\mathbf{A}$ vuông cấp n được gọi là nửa xác định dương nếu $\mathbf{x}^\top\mathbf{Ax} \geq 0$ $\forall \mathbf{x} \in \mathbb{R}^n$. $\mathbf{A}$ được gọi là xác định dương nếu $\mathbf{x}^\top\mathbf{Ax} > 0$ $\forall \mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq 0$.

Một số λ được gọi là giá trị riêng của một ma trận vuông $\mathbf{A}$ nếu tồn tại một vector không bằng không $\mathbf{v}$ sao cho:

$$\mathbf{Av} = \lambda\mathbf{v}$$

Vector $\mathbf{v}$ tương ứng với giá trị riêng λ được gọi là vector riêng.

Vector hóa của ma trận $\mathbf{A} \in \mathbb{R}^{m\times n}$ là:

$$vec(\mathbf{A}) = \begin{pmatrix} \mathbf{A_{:1}} \\ \vdots \\ \mathbf{A_{:n}} \end{pmatrix} \in \mathbb{R}^{mn}$$

Ví dụ 1.1. *Cho ma trận* $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Vector hóa của ma trận $\mathbf{A}$ *là:* $vec(\mathbf{A}) = \begin{pmatrix} a \\ c \\ b \\ d \end{pmatrix}$.

Bằng cách vector hóa ma trận, ta có thể xem một ma trận tổng quát $\mathbf{A}$ cỡ $m \times n$ như một vector: $vec(\mathbf{A})$ với mn phần tử và có thể xác định tích vô

hướng của hai ma trận thực có cùng cỡ như sau:

$$\langle \mathbf{A}, \mathbf{B} \rangle = vec(\mathbf{A})^\top vec(\mathbf{B}) = \sum_{ij} a_{ij} b_{ij}$$

Mệnh đề 1.1. *Với* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *là các ma trận cỡ* $m \times n$ *và* $\lambda \in \mathbb{R}$*, tích vô hướng thỏa mãn:*

i) $\langle \mathbf{A}, \mathbf{B} \rangle = \langle \mathbf{B}, \mathbf{A} \rangle$

ii) $\langle \mathbf{A} + \mathbf{C}, \mathbf{B} \rangle = \langle \mathbf{A}, \mathbf{B} \rangle + \langle \mathbf{C}, \mathbf{B} \rangle$

iii) $\langle \lambda \mathbf{A}, \mathbf{B} \rangle = \lambda \langle \mathbf{A}, \mathbf{B} \rangle$

Chứng minh:

i) $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} a_{ij} b_{ij} = \sum_{ij} b_{ij} a_{ij} = \langle \mathbf{B}, \mathbf{A} \rangle$

ii) $\langle \mathbf{A} + \mathbf{C}, \mathbf{B} \rangle = \sum_{ij} (a_{ij} + c_{ij}) b_{ij} = \sum_{ij} a_{ij} b_{ij} + \sum_{ij} c_{ij} b_{ij} = \langle \mathbf{A}, \mathbf{B} \rangle + \langle \mathbf{C}, \mathbf{B} \rangle$

iii) $\langle \lambda \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} \lambda a_{ij} b_{ij} = \lambda \sum_{ij} a_{ij} b_{ij} = \lambda \langle \mathbf{A}, \mathbf{B} \rangle$ □

Tích Hadamard của hai ma trận $\mathbf{A}$ và $\mathbf{B}$ cùng cỡ $m \times n$ (kí hiệu $A \circ B$) là một ma trận cùng cỡ $\mathbf{C}$ với $c_{ij} = a_{ij} b_{ij}$.

Ví dụ 1.2. *Cho hai ma trận* $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 4 & 3 & 0 & 1 \\ 2 & 1 & 5 & 3 \end{bmatrix}$ *và* $\mathbf{B} = \begin{bmatrix} 4 & 1 & 1 & 2 \\ 5 & 4 & 3 & 1 \\ 0 & 1 & 2 & 0 \end{bmatrix}$*, thì tích Hadamard của chúng là:*

$$\mathbf{A} \circ \mathbf{B} = \begin{bmatrix} 1*4 & 2*1 & 3*1 & 0*2 \\ 4*5 & 3*4 & 0*3 & 1*1 \\ 2*0 & 1*1 & 5*2 & 3*0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 3 & 0 \\ 20 & 12 & 0 & 1 \\ 0 & 1 & 10 & 0 \end{bmatrix}$$

Mệnh đề 1.2. *Với* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *là các ma trận cỡ* $m \times n$*, tích Hadamard thỏa mãn:*

i) $\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$

ii) $\mathbf{A} \circ (\mathbf{B} \circ \mathbf{C}) = (\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C}$

iii) $\mathbf{A} \circ (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \circ \mathbf{B}) + (\mathbf{A} \circ \mathbf{C})$

iv) $\mathbf{A}^\top \circ \mathbf{B}^\top = (\mathbf{A} \circ \mathbf{B})^\top$

Chứng minh:

i) $\mathbf{A} \circ \mathbf{B} = (a_{ij}b_{ij})_{m\times n} = (b_{ij}a_{ij})_{m\times n} = \mathbf{B} \circ \mathbf{A}$

ii) $\mathbf{A} \circ (\mathbf{B} \circ \mathbf{C}) = (a_{ij}(b_{ij}c_{ij}))_{m\times n} = ((a_{ij}b_{ij})c_{ij})_{m\times n} = (\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C}$

iii) $\mathbf{A} \circ (\mathbf{B} + \mathbf{C}) = (a_{ij}(b_{ij} + c_{ij}))_{m\times n} = (a_{ij}b_{ij})_{m\times n} + (a_{ij}c_{ij})_{m\times n}$
$= (\mathbf{A} \circ \mathbf{B}) + (\mathbf{A} \circ \mathbf{C})$

iv) $\mathbf{A}^\top \circ \mathbf{B}^\top = (a_{ji}b_{ji})_{n\times m} = (a_{ij}b_{ij})^\top_{n\times m} = (\mathbf{A} \circ \mathbf{B})^\top$ □

1.1.2 Chuẩn

Định nghĩa 1.1. *Một chuẩn vector trên $\mathbb{R}^n$ là một hàm $f : \mathbb{R}^n \to \mathbb{R}$ thỏa mãn các tính chất sau:*

i) $f(\mathbf{x}) \geq 0$, $\forall \mathbf{x} \in \mathbb{R}^n$ *và* $f(\mathbf{x}) = 0 \Leftrightarrow \mathbf{x} = 0$

ii) $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

iii) $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$, $\forall \alpha \in \mathbb{R}$, $\forall \mathbf{x} \in \mathbb{R}^n$

Chuẩn của $\mathbf{x}$ thường được ký hiệu là $\|\mathbf{x}\|$. Cho vector $\mathbf{x} = (x_1, x_2, ..., x_n)^\top$, chuẩn của vector được xác định bởi:

$$\|\mathbf{x}\|_p := \left(|x_1|^p + |x_2|^p + ... + |x_n|^p\right)^{\frac{1}{p}}, p \geq 1$$

- Chuẩn 1 ($p = 1$):
$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + ... + |x_n|$$
- Chuẩn 2 ($p = 2$) hay còn gọi là chuẩn Euclide:
$$\|\mathbf{x}\| = \left(|x_1|^2 + |x_2|^2 + ... + |x_n|^2\right)^{\frac{1}{2}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$
- Chuẩn ∞ ($p = \infty$):
$$\|\mathbf{x}\|_\infty = \max(|x_1|; |x_2|; ...; |x_n|)$$

Ví dụ 1.3. *Cho vector* $\mathbf{x} = (1 \quad -2 \quad 4)^\top$.

Chuẩn 1 của vector $\mathbf{x}$ *là:* $\|\mathbf{x}\|_1 = |1| + |-2| + |4| = 7$.

Chuẩn Euclide của vector $\mathbf{x}$ *là:* $\|\mathbf{x}\| = (|1|^2 + |-2|^2 + |4|^2)^{\frac{1}{2}} = \sqrt{21}$

Chuẩn ∞ *của vector* $\mathbf{x}$ *là:* $\|\mathbf{x}\|_\infty = \max(1; -2; 4) = 4$

Định nghĩa 1.2. *Chuẩn ma trận trên* $\mathbb{R}^{m\times n}$ *là hàm số* $f:\mathbb{R}^{m\times n}\to\mathbb{R}$ *thỏa mãn các tính chất sau:*

i) $f(\mathbf{A})\geq 0$, $\forall\mathbf{A}\in\mathbb{R}^{m\times n}$ *và* $f(\mathbf{A})=0\Leftrightarrow\mathbf{A}=0$

ii) $f(\mathbf{A}+\mathbf{B})\leq f(\mathbf{A})+f(\mathbf{B})$, $\forall\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$

iii) $f(\alpha\mathbf{A})=|\alpha|f(\mathbf{A})$, $\forall\alpha\in\mathbb{R}$, $\forall\mathbf{A}\in\mathbb{R}^{m\times n}$

Chuẩn của $\mathbf{A}$ thường được ký hiệu là $\|\mathbf{A}\|$. Cho ma trận $\mathbf{A}=(a_{ij})_{m\times n}$, một số chuẩn ma trận thông dụng là:

- Chuẩn 1 (chuẩn cực đại theo cột):

$$\|\mathbf{A}\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|$$

- Chuẩn Frobenius. Kí hiệu: $\|\mathbf{A}\|_F$

$$\|\mathbf{A}\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}=\sqrt{\langle\mathbf{A},\mathbf{A}\rangle}$$

- Chuẩn ∞ (chuẩn cực đại theo hàng):

$$\|\mathbf{A}\|_\infty=\max_{1\leq i\leq m}\sum_{j=1}^{n}|a_{ij}|$$

Ví dụ 1.4. *Cho ma trận* $\mathbf{A}=\begin{bmatrix}4&2&3\\-1&3&-2\\0&1&5\end{bmatrix}$

a) Chuẩn 1 của ma trận $\mathbf{A}$*:*

$$\begin{aligned}\|\mathbf{A}\|_1&=\max_{1\leq j\leq 3}\sum_{i=1}^{3}|a_{ij}|=\max_{1\leq j\leq 3}(|a_{1j}|+|a_{2j}|+|a_{3j}|)\\&=\max(|a_{11}|+|a_{21}|+|a_{31}|;|a_{12}|+|a_{22}|+|a_{32}|;|a_{13}|+|a_{23}|+|a_{33}|)\\&=\max(4+1+0;2+3+1;3+2+5)=max(5;6;10)=10\end{aligned}$$

b) Chuẩn Frobenius của ma trận $\mathbf{A}$*:*

$$\begin{aligned}\|\mathbf{A}\|_F&=\sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}|a_{ij}|^2}\\&=\left(|4|^2+|2|^2+|3|^2+|-1|^2+|3|^2+|-2|^2+|0|^2+|1|^2+|5|^2\right)^{\frac{1}{2}}=\sqrt{69}\end{aligned}$$

c) Chuẩn ∞ của ma trận $\mathbf{A}$*:*

$$\begin{aligned}\|\mathbf{A}\|_\infty &= \max_{1\le i\le 3}\sum_{j=1}^{3}|a_{ij}| = \max_{1\le i\le 3}(|a_{i1}|+|a_{i2}|+|a_{i3}|)\\ &= \max(|a_{11}|+|a_{12}|+|a_{13}|;|a_{21}|+|a_{22}|+|a_{23}|;|a_{31}|+|a_{32}|+|a_{33}|)\\ &= \max(4+2+3;1+3+2;0+1+5) = max(9;6;6) = 9\end{aligned}$$

1.1.3 Ma trận không âm

Định nghĩa 1.3. *Ma trận* $\mathbf{A}\in\mathbb{R}^{m\times n}$ *có tất cả các phần tử không âm được gọi là ma trận không âm, nghĩa là* $a_{ij}\ge 0$ *với mọi* $i=\overline{1,m}$ *và* $j=\overline{1,n}$*.*

Ký hiệu: $\mathbf{A}\ge 0$. $\mathbb{R}_+^{m\times n}$ là tập hợp các ma trận không âm cỡ $m\times n$.

Mệnh đề 1.3. *Nếu* $\mathbf{A},\mathbf{B}$ *là hai ma trận không âm thì ta có:*

i) Với $k>0$*,* $k\mathbf{A}$ *là một ma trận không âm.*

ii) Các ma trận $\mathbf{A}+\mathbf{B},\mathbf{AB}$ *cũng đều là ma trận không âm.*

iii) Với n *là một số nguyên dương,* $\mathbf{A}^n$ *cũng là một ma trận không âm.* $\mathbf{A}^n$ *tiến dần đến ma trận hằng số khi* n *tiến đến vô cùng.*

Định lý 1.1. *Cho* $\mathbf{A}$ *là một ma trận vuông không âm. Khi đó giá trị riêng lớn nhất của* $\mathbf{A}$ *là một số không âm và tồn tại ít nhất một vector riêng không âm tương ứng với giá trị riêng lớn nhất đó.*

1.2 Lý thuyết tối ưu

1.2.1 Tập lồi và hàm lồi

Định nghĩa 1.4. *Một tập con* $M\subseteq\mathbb{R}^n$ *được gọi là tập tập lồi nếu với mọi* $x_1,x_2\in M$ *và* $0\le\lambda\le 1$ *ta có:*

$$\lambda x_1+(1-\lambda)x_2\in M$$

Ví dụ 1.5. *Tập hợp các ma trận không âm cỡ* $m\times n$ $(\mathbb{R}_+^{m\times n})$ *là một tập lồi.*

Chứng minh:
Cho hai ma trận không âm $\mathbf{A}=(a_{ij})_{m\times n}$ và $\mathbf{B}=(b_{ij})_{m\times n}$.
Xét ma trận $\mathbf{C}=\lambda\mathbf{A}+(1-\lambda)\mathbf{B}=\lambda(a_{ij})_{m\times n}+(1-\lambda)(b_{ij})_{m\times n}$ với $\lambda\in[0,1]$.
Ta cần chứng minh rằng mọi phần tử $c_{ij}=\lambda a_{ij}+(1-\lambda)b_{ij}$ đều không âm.

Vì $\mathbf{A}$ và $\mathbf{B}$ là ma trận không âm, nên $a_{ij} \geq 0$ và $b_{ij} \geq 0$ với mọi i, j.
Khi $\lambda \in [0,1]$, ta có:

$$\lambda a_{ij} \geq 0 \text{ và } (1-\lambda)b_{ij} \geq 0$$

Do đó:

$$\lambda a_{ij} + (1-\lambda)b_{ij} \geq 0$$

Vì vậy, $c_{ij} = \lambda a_{ij} + (1-\lambda)b_{ij} \geq 0$ với mọi i, j.
Như vậy, ta đã chứng minh được rằng mọi ma trận $\mathbf{C} = \lambda\mathbf{A} + (1-\lambda)\mathbf{B}$ với $\lambda \in [0,1]$ đều là ma trận không âm. $\Longrightarrow$ đpcm $\square$

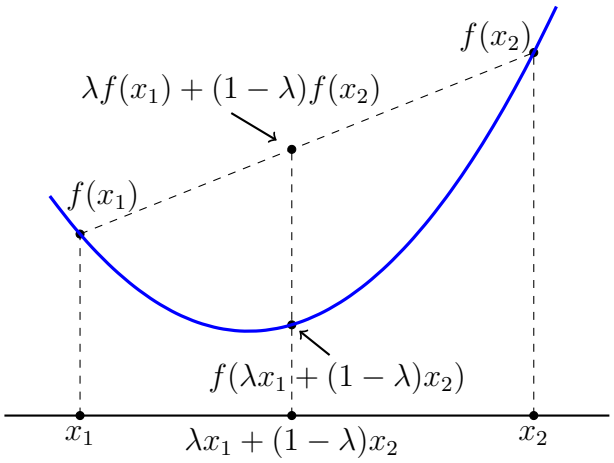
Định nghĩa 1.5. *Giả sử tập $M \subseteq \mathbb{R}^n$ là tập lồi. Hàm số $f : M \to \mathbb{R}$. Hàm f được gọi là hàm lồi xác định trên M nếu*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

với bất kỳ $x_1, x_2 \in M$ và số thực $\lambda \in [0,1]$. Ta gọi f là hàm lồi chặt trên tập lồi M nếu

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda)f(x_2)$$

với bất kỳ $x_1, x_2 \in M, x_1 \neq x_2$ và $\lambda \in (0,1)$.



Hình 1.1: Hàm lồi

Mệnh đề 1.4. *Cho $\|.\|$ là một chuẩn tương ứng với tích vô hướng $\langle .,. \rangle$ trên $\mathbb{R}^n$. Khi đó:*

i) $\|.\|$ là hàm lồi trên $\mathbb{R}^n$.

ii) $\|.\|^2$ là hàm lồi trên $\mathbb{R}^n$

Chứng minh:
i) Với mọi $x, y \in \mathbb{R}^n$ và $\lambda \in [0,1]$, ta có:

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|$$

ii) Với mọi $x, y \in \mathbb{R}^n$ và $\lambda \in [0,1]$, ta có:

$$\begin{aligned}
&\|\lambda x + (1-\lambda)y\|^2 = \langle \lambda x + (1-\lambda)y, \lambda x + (1-\lambda)y \rangle \\
&= \lambda^2 \|x\|^2 + 2\lambda(1-\lambda)\langle x, y\rangle + (1-\lambda)^2 \|y\|^2 \\
&= \lambda \|x\|^2 + (\lambda^2 - \lambda)\|x\|^2 + (1-\lambda)\|y\|^2 + \Big((1-\lambda)^2 - (1-\lambda)\Big)\|y\|^2 \\
&\qquad\qquad + 2\lambda(1-\lambda)\langle x, y\rangle \\
&= \lambda \|x\|^2 + (1-\lambda)\|y\|^2 + \lambda(\lambda-1)\|x\|^2 + \lambda(\lambda-1)\|y\|^2 + 2\lambda(\lambda-1)\langle x, y\rangle \\
&= \lambda \|x\|^2 + (1-\lambda)\|y\|^2 + \lambda(\lambda-1)\|x+y\|^2 \\
&\leq \lambda \|x\|^2 + (1-\lambda)\|y\|^2
\end{aligned}$$

□

1.2.2 Điều kiện tối ưu

Xét bài toán:

$$\min_{x \in M} f(x) \tag{1.1}$$

với $M \subseteq \mathbb{R}^n$, $f : M \to \mathbb{R}$.

Định nghĩa 1.6. *Điểm $x^* \in M$ được gọi là nghiệm tối ưu địa phương (nghiệm cực tiểu địa phương) của bài toán (1.1) nếu tồn tại một không gian hình cầu $B(x^*, \varepsilon)$ xung quanh x^* với bán kính $\varepsilon > 0$ sao cho*

$$f(x) \geq f(x^*), \forall x \in M \cap B(x^*, \varepsilon).$$

Điểm $x^* \in M$ được gọi là nghiệm tối ưu địa phương chặt (nghiệm cực tiểu địa phương chặt) của bài toán (1.1) nếu tồn tại một ε-lân cận $B(x^*, \varepsilon)$ của điểm $x^* \in M$ sao cho

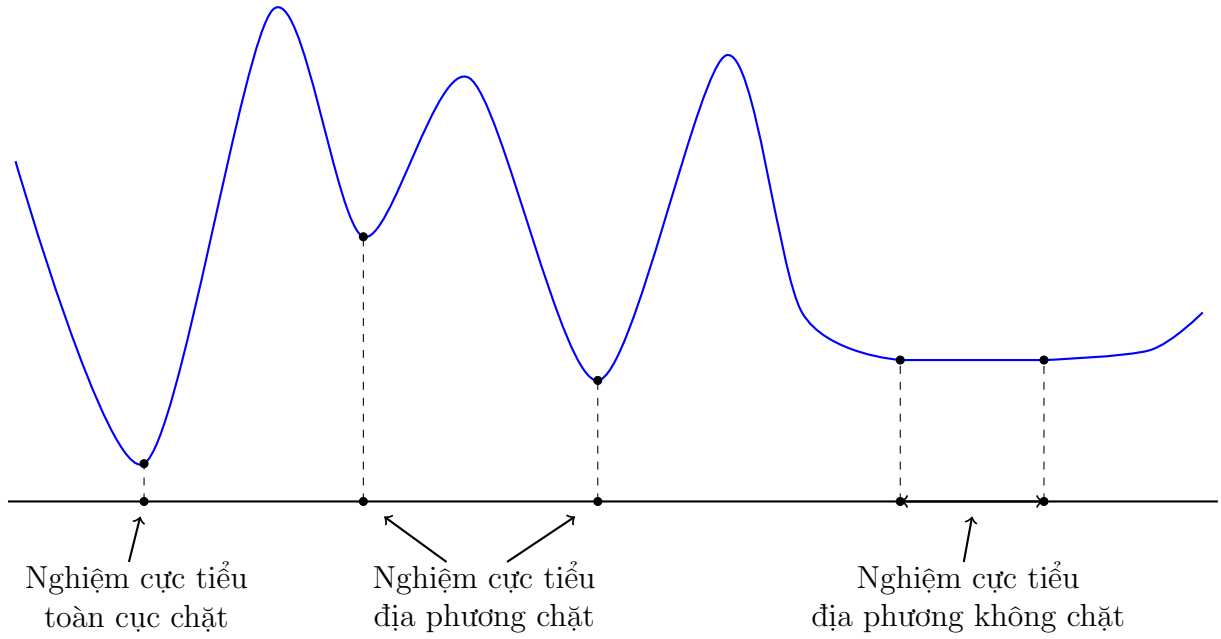
$$f(x) > f(x^*),\ \forall x \in M \cap B(x^*, \varepsilon) \text{ và } x \neq x^*.$$

Định nghĩa 1.7. *Điểm $x^* \in M$ mà*

$$f(x) \geq f(x^*), \forall x \in M$$

được gọi là nghiệm tối ưu (nghiệm tối ưu toàn cục, hoặc nghiệm cực tiểu toàn cục) của bài toán (1.1).
Điểm $x^* \in M$ được gọi là nghiệm cực tiểu toàn cục chặt của bài toán (1.1) nếu

$$f(x) > f(x^*),\ \forall x \in M \text{ và } x \neq x^*.$$



Hình 1.2: Ví dụ về nghiệm của bài toán tối ưu

Tập nghiệm tối ưu của bài toán (1.1) được ký hiệu là: $\text{Argmin}\{f(x)|x \in M\}$. Nếu bài toán chỉ có một nghiệm tối ưu x^* thì $x^* = \text{argmin}\{f(x)|x \in M\}$.

Tập các hướng chấp nhận được tại $x^* \in M$ là:

$$F(M, x^*) = \{d \in \mathbb{R}^n | \exists \lambda^* > 0 : x^* + \lambda d \in M, \forall 0 \leq \lambda \leq \lambda^*\}.$$

Định lý 1.2. *Giả sử tập $M \subset \mathbb{R}^n$ và f là một hàm khả vi trên M. Nếu x^* là nghiệm cực tiểu địa phương của f trên M thì:*

$$\langle \nabla f(x^*), d \rangle \geq 0, \forall d \in F(M, x^*).$$

Chứng minh:

Lấy $d \in F(M, x^*)$. Khi đó, tồn tại λ^* sao cho $\forall \lambda : 0 \leq \lambda \leq \lambda^*$ thì $x^* + \lambda d \in M$. x^* là nghiệm cực tiểu địa phương có nghĩa là tồn tại $\varepsilon > 0$ sao cho: $\forall x \in M \cap B(x^*, \varepsilon)$ thì $f(x) \geq f(x^*)$.
Lấy $\lambda_1 = \min\left(\lambda^*, \frac{\varepsilon}{\|d\|}\right)$. Khi đó, $\forall 0 \leq \lambda \leq \lambda_1$ ta có:

$$\begin{cases} x^* + \lambda d \in M \\ x^* + \lambda d \in B(x^*, \varepsilon) \end{cases}$$

nên

$$f(x^* + \lambda d) \geq f(x^*), \forall 0 \leq \lambda \leq \lambda_1$$

Do đó:

$$\lim_{t \to 0^+} \frac{f(x^* + td) - f(x^*)}{t} \geq 0$$

Ta lại có:

$$f^{'}(x^*, d) = \lim_{t \to 0^+} \frac{f(x^* + td) - f(x^*)}{t} = \langle \nabla f(x^*), d \rangle$$

$\Longrightarrow \langle \nabla f(x^*), d \rangle \geq 0, \forall d \in F(M, x^*)$ □

Định nghĩa 1.8. *Điểm $x^* \in M$ thỏa mãn:*

$$\langle \nabla f(x^*), d \rangle \geq 0, \forall d \in F(M, x^*)$$

được gọi là điểm dừng của bài toán (1.1).

Mệnh đề 1.5. *Giả sử $x^* \in intM$ và x^* là điểm cực tiểu địa phương của bài toán (1.1). Khi đó $\nabla f(x^*) = 0$.*

Chứng minh:

Do $x^* \in intM$ nên $F(M, x^*) = \mathbb{R}^n$. Vì x^* là điểm cực tiểu địa phương nên theo Định lý 1.2, ta có:

$$\langle \nabla f(x^*), d \rangle \geq 0, \forall d \in \mathbb{R}^n.$$

Suy ra $\nabla f(x^*) = 0$. □

Định lý 1.3. *Giả sử $M \subset \mathbb{R}^n$ là tập lồi khác rỗng và f là hàm lồi. Khi đó, nếu x^* là nghiệm tối ưu địa phương của bài toán (1.1) thì x^* cũng là nghiệm tối ưu toàn cục.*

Chứng minh:

Giả sử $x^* \in M$ là nghiệm tối ưu địa phương của bài toán (1.1). Theo Định nghĩa 1.6, tồn tại một ε-lân cận $B(x^*, \varepsilon)$ của điểm $x^* \in M$ sao cho

$$f(y) \geq f(x^*), \forall y \in B(x^*, \varepsilon) \cap M.$$

Với mọi $x \in M$, tồn tại $\lambda \in [0, 1]$ sao cho $x^* + \lambda(x - x^*) \in B(x^*, \varepsilon) \cap M$ nên

$$\overline{x} = \lambda x + (1 - \lambda)x^* = x^* + \lambda(x - x^*) \in B(x^*, \varepsilon) \cap M.$$

Do x^* là nghiệm cực tiểu địa phương và f là hàm lồi trên M nên

$$f(x^*) \leq f(\overline{x}) \leq \lambda f(x) + (1 - \lambda) f(x^*) \Rightarrow f(x) \geq f(x^*) \text{ đúng } \forall x \in M.$$

Điều đó chứng tỏ x^* là nghiệm tối ưu toàn cục của bài toán đang xét. □

1.2.3 Điều kiện Karush-Kuhn-Tucker (KKT)

Xét bài toán tối ưu:

$$\min_{x \in M} f(x) \tag{1.2}$$

trong đó $M \subset \mathbb{R}^n$ là tập nghiệm của hệ

$$\begin{cases} g_i(x) \leq 0, i = 1, 2, ..., m \\ h_j(x) = 0, j = 1, 2, ..., k \end{cases}$$

với $f, g_i(x), h_j(x)$ ($i = \overline{1,m}$ và $j = \overline{1,k}$) là các hàm số khả vi bất kỳ xác định trên $\mathbb{R}^n$ và mỗi hệ thức $g_i(x) \leq 0, i \in \{1, 2, ..., m\}$ hoặc $h_j(x) = 0, j \in \{1, 2, ..., k\}$ được gọi là một ràng buộc. Ràng buộc của bài toán (1.2) được viết trong hàm Lagrange như sau:

$$\mathbb{L}(x, \lambda_1, ..., \lambda_m, \mu_1, ..., \mu_k) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{k} \mu_j h_j(x) \tag{1.3}$$

trong đó $\lambda_i \geq 0$ ($i = \overline{1,m}$) và μ_j ($j = \overline{1,k}$) được gọi là các nhân tử Lagrange.

Định lý 1.4. *Cho x^* là nghiệm cực tiểu địa phương của bài toán (1.2). Giả sử rằng $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ là các hàm khả vi liên tục; $\nabla g_i(x^*)$ và $\nabla h_j(x^*)$ là độc lập tuyến tính. Khi đó tồn tại λ_i ($i = \overline{1,m}$) và μ_j ($j = \overline{1,k}$) thỏa mãn các điều kiện sau:*

$$i) \quad \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{k} \mu_j \nabla h_j(x^*) = 0$$

$$ii) \quad \lambda_i \geq 0 \quad (i = \overline{1,m})$$

$$iii) \quad \lambda_i g_i(x^*) = 0 \quad (i = \overline{1,m})$$

được gọi là điều kiện Karush-Kuhn-Tucker (điều kiện KKT) của bài toán (1.2).

Chương 2

Thuật toán giải bài toán NMF

Trong chương này sẽ trình bày bài toán nhân tử hóa ma trận không âm (Non-negative Matrix Factorization - NMF), điều kiện cần tối ưu, quy tắc cập nhật nhân (Multiplicative Update Rules - MUR) để giải bài toán và kiểm tra sự hội tụ của thuật toán. Nội dung của chương được tham khảo chủ yếu từ các tài liệu [8], [10], [12] và [13].

2.1 Phát biểu bài toán

Nhân tử hóa ma trận không âm (NMF) được giới thiệu lần đầu tiên vào năm 1994 bởi Paatero và Tapper [7] nhưng nó đã trở nên nổi tiếng nhờ công trình của Lee và Seung vào năm 1999 và 2001 [8]. Họ cho rằng tính không âm là rất quan trọng trong nhận thức của con người và đã đề xuất thuật toán đơn giản để tìm một biểu diễn không âm cho dữ liệu. Bài toán NMF có thể được phát biểu như sau:

Cho một ma trận dữ liệu $\mathbf{A} \in \mathbb{R}^{m \times n}$ với các phần tử không âm (tức $a_{ij} \geq 0$), tìm một phân rã sao cho:

$$\mathbf{A} \approx \mathbf{UV} \tag{2.1a}$$

trong đó $\mathbf{U}$ và $\mathbf{V}$ là các ma trận không âm có kích thước lần lượt là $m \times r$ và $r \times n$. Thông thường, giá trị r được chọn sao cho $r \leq \min(m, n)$.

Có thể dùng nhiều cách khác nhau để xác định sự khác nhau giữa ma trận dữ liệu $\mathbf{A}$ và ma trận xấp xỉ $\mathbf{UV}$ nhưng phương pháp được dùng phổ biến nhất là chuẩn Frobenius:

$$F(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{A} - \mathbf{UV}\|_F^2 = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(a_{ij} - [\mathbf{UV}]_{ij}\right)^2 \tag{2.1b}$$

còn được gọi là khoảng cách Euclide.

Mệnh đề 2.1. *Giả sử* $\mathbf{U}$ *cố định, hàm (2.1b) có thể được xem như là hợp thành của chuẩn Frobenius và một phép biến đổi tuyến tính của* $\mathbf{V}$*. Do đó,* F *là hàm lồi theo* $\mathbf{V}$*. Tương tự, nếu* $\mathbf{V}$ *cố định,* F *cũng là hàm lồi theo* $\mathbf{U}$*.*

Chứng minh:

i) Với $\mathbf{U}$ cố định, ta có:

$$\begin{aligned} F(\mathbf{U}, \lambda\mathbf{V_1} + (1-\lambda)\mathbf{V_2}) &= \frac{1}{2}\left\|\mathbf{A} - \lambda\mathbf{U}\mathbf{V_1} - (1-\lambda)\mathbf{U}\mathbf{V_2}\right\|_F^2 \\ &= \frac{1}{2}\left\|\lambda\mathbf{A} + (1-\lambda)\mathbf{A} - \lambda\mathbf{U}\mathbf{V_1} - (1-\lambda)\mathbf{U}\mathbf{V_2}\right\|_F^2 \\ &= \frac{1}{2}\left\|\lambda\left(\mathbf{A} - \mathbf{U}\mathbf{V_1}\right) + (1-\lambda)\left(\mathbf{A} - \mathbf{U}\mathbf{V_2}\right)\right\|_F^2 \end{aligned}$$

Theo Mệnh đề 1.3, $\|.\|^2$ là hàm lồi trên $\mathbb{R}^{m\times n}$, do đó:

$$\begin{aligned} F(\mathbf{U}, \lambda\mathbf{V_1} + (1-\lambda)\mathbf{V_2}) &\leq \lambda\frac{1}{2}\left\|\mathbf{A} - \mathbf{U}\mathbf{V_1}\right\|_F^2 + (1-\lambda)\frac{1}{2}\left\|\mathbf{A} - \mathbf{U}\mathbf{V_2}\right\|_F^2 \\ &= \lambda F(\mathbf{U}, \mathbf{V_1}) + (1-\lambda)F(\mathbf{U}, \mathbf{V_2}) \end{aligned}$$

Suy ra F là hàm lồi theo $\mathbf{V}$.

ii) Với $\mathbf{V}$ cố định, ta có:

$$\begin{aligned} F(\lambda\mathbf{U_1} + (1-\lambda)\mathbf{U_2}, \mathbf{V}) &= \frac{1}{2}\left\|\mathbf{A} - \lambda\mathbf{U_1}\mathbf{V} - (1-\lambda)\mathbf{U_2}\mathbf{V}\right\|_F^2 \\ &= \frac{1}{2}\left\|\lambda\mathbf{A} + (1-\lambda)\mathbf{A} - \lambda\mathbf{U_1}\mathbf{V} - (1-\lambda)\mathbf{U_2}\mathbf{V}\right\|_F^2 \\ &= \frac{1}{2}\left\|\lambda\left(\mathbf{A} - \mathbf{U_1}\mathbf{V}\right) + (1-\lambda)\left(\mathbf{A} - \mathbf{U_2}\mathbf{V}\right)\right\|_F^2 \end{aligned}$$

Theo Mệnh đề 1.3, $\|.\|^2$ là hàm lồi trên $\mathbb{R}^{m\times n}$, do đó:

$$\begin{aligned} F(\lambda\mathbf{U_1} + (1-\lambda)\mathbf{U_2}, \mathbf{V}) &\leq \lambda\frac{1}{2}\left\|\mathbf{A} - \mathbf{U_1}\mathbf{V}\right\|_F^2 + (1-\lambda)\frac{1}{2}\left\|\mathbf{A} - \mathbf{U_2}\mathbf{V}\right\|_F^2 \\ &= \lambda F(\mathbf{U_1}, \mathbf{V}) + (1-\lambda)F(\mathbf{U_2}, \mathbf{V}) \end{aligned}$$

Suy ra F là hàm lồi theo $\mathbf{U}$. □

Tóm lại, phương pháp thông thường để giải bài toán NMF là tái sắp xếp (2.1a) thành bài toán tối ưu sau đây:

$$\min_{\mathbf{U},\mathbf{V}} F(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\left\|\mathbf{A} - \mathbf{U}\mathbf{V}\right\|_{\mathbf{F}}^2 \tag{2.1c}$$

trong đó $\mathbf{A} \in \mathbb{R}_+^{m\times n}$, $\mathbf{U} \in \mathbb{R}_+^{m\times r}$ và $\mathbf{V} \in \mathbb{R}_+^{r\times n}$ với r là số nguyên dương thỏa mãn $r \leq \min(m, n)$.

Chú ý 2.1. *Với* $\mathbf{U}$ *cố định, (2.1b) là hàm lồi theo* $\mathbf{V}$ *và với* $\mathbf{V}$ *cố định, (2.1b) là hàm lồi theo* $\mathbf{U}$*. Nhưng hàm (2.1b) không lồi theo cả* $\mathbf{U}$ *và* $\mathbf{V}$ *vì vậy bài toán (2.1c) là bài toán tối ưu không lồi. Về mặt lý thuyết, có nhiều thuật toán để tìm cực tiểu toàn cục của một bài toán tối ưu không lồi. Tuy nhiên, trong thực tế khi giải bài toán NMF, người ta thường chỉ đi tìm cực tiểu địa phương thay vì cực tiểu toàn cục.*

2.2 Điều kiện cần tối ưu

2.2.1 Hàm Lagrange

Bài toán (2.1c) tương đương với bài toán tối ưu sau đây:

$$\min_{-\mathbf{U}\leq 0,-\mathbf{V}\leq 0} F(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2 \tag{2.2}$$

Các ràng buộc $-\mathbf{U}\leq 0$ và $-\mathbf{V}\leq 0$ có nghĩa là:

$$\begin{cases} -u_{ia}\leq 0, & \forall i=\overline{1,m} \text{ và } a=\overline{1,r} \\ -v_{bj}\leq 0, & \forall b=\overline{1,r} \text{ và } j=\overline{1,n} \end{cases}$$

Ta kí hiệu các nhân tử Lagrange tương ứng là:

$$\begin{cases} \mu_{ia}\geq 0, & \forall i=\overline{1,m} \text{ và } a=\overline{1,r} \\ \nu_{bj}\geq 0, & \forall b=\overline{1,r} \text{ và } j=\overline{1,n} \end{cases}$$

Đặt $\mu=(\mu_{ia})_{m\times r}$ và $\nu=(\nu_{bj})_{r\times n}$.
Khi đó hàm Lagrange tương ứng của bài toán (2.2) là:

$$\begin{aligned}\mathbb{L}(\mathbf{U},\mathbf{V},\mu,\nu) &= \frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2 - \sum_{i=\overline{1,m}}^{a=\overline{1,r}} \mu_{ia}u_{ia} - \sum_{b=\overline{1,r}}^{j=\overline{1,n}} \nu_{bj}v_{bj} \\ &= \frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2 - \mu\circ\mathbf{U} - \nu\circ\mathbf{V}\end{aligned}$$

Chú ý 2.2. $\mu\circ\mathbf{U}$ *là tích Hadamard của* μ *và* $\mathbf{U}$ *đã định nghĩa ở Chương 1.*

2.2.2 Điều kiện cần tối ưu

Điều kiện KKT của bài toán (2.2) nói rằng nếu $(\mathbf{U},\mathbf{V})$ là điểm cực tiểu địa phương thì tồn tại $\mu\geq 0$, $\nu\geq 0$ sao cho:

- Điều kiện chấp nhận được:

$$\begin{aligned} &-\mathbf{U} \leq 0, -\mathbf{V} \leq 0 \\ \Leftrightarrow\ &\mathbf{U} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{2.3a}$$

- Đạo hàm hàm Lagrange bằng 0:

$$\nabla \mathbb{L}_{\mathbf{U}} = 0, \nabla \mathbb{L}_{\mathbf{V}} = 0$$

Vì ta có:

$$\nabla \mathbb{L}_{\mathbf{U}} = \frac{\partial \mathbb{L}}{\partial \mathbf{U}} = \frac{\partial \left(\frac{1}{2} \|\mathbf{A} - \mathbf{UV}\|_{\mathbf{F}}^2 \right)}{\partial \mathbf{U}} - \frac{\partial (\mu \circ \mathbf{U})}{\partial \mathbf{U}} = -(\mathbf{A} - \mathbf{UV})\mathbf{V}^\top - \mu$$

$$\nabla \mathbb{L}_{\mathbf{V}} = \frac{\partial \mathbb{L}}{\partial \mathbf{V}} = \frac{\partial \left(\frac{1}{2} \|\mathbf{A} - \mathbf{UV}\|_{\mathbf{F}}^2 \right)}{\partial \mathbf{V}} - \frac{\partial (\nu \circ \mathbf{V})}{\partial \mathbf{V}} = -\mathbf{U}^\top(\mathbf{A} - \mathbf{UV}) - \nu$$

Nên ta có: $\mu = \mathbf{UVV}^\top - \mathbf{AV}^\top$, $\nu = \mathbf{U}^\top\mathbf{UV} - \mathbf{U}^\top\mathbf{A}$ và với $\mu \geq 0$, $\nu \geq 0$. Khi đó điều kiện sẽ trở thành:

$$\nabla F_{\mathbf{U}} = \mathbf{UVV}^\top - \mathbf{AV}^\top \geq 0, \nabla F_{\mathbf{V}} = \mathbf{U}^\top\mathbf{UV} - \mathbf{U}^\top\mathbf{A} \geq 0 \tag{2.3b}$$

- Điều kiện bù:

$$\begin{aligned} &\mu \circ \mathbf{U} = 0, \nu \circ \mathbf{V} = 0 \\ \Leftrightarrow\ &\mathbf{U} \circ \nabla F_{\mathbf{U}} = 0, \mathbf{V} \circ \nabla F_{\mathbf{V}} = 0 \end{aligned} \tag{2.3c}$$

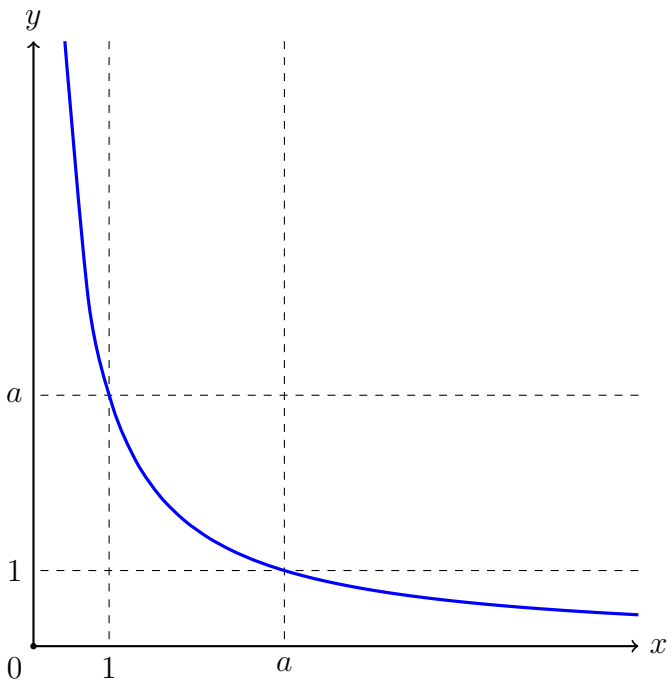
Định nghĩa 2.1. *(Điểm dừng NMF) Gọi* $(\mathbf{U}, \mathbf{V})$ *là điểm dừng của bài toán NMF khi và chỉ khi* $\mathbf{U}$ *và* $\mathbf{V}$ *thoả mãn các điều kiện KKT (2.3a), (2.3b) và (2.3c).*

Chú ý 2.3. *Ngoài ra, một điểm dừng* $(\mathbf{U}, \mathbf{V})$ *của bài toán NMF cũng có thể được định nghĩa bằng cách sử dụng điều kiện trong Định lý 1.2 trên tập lồi* $\mathbb{R}_+^{m \times r}$ *và* $\mathbb{R}_+^{n \times r}$*, tức là:*

$$\left\langle \begin{pmatrix} \nabla F_{\mathbf{U}} \\ \nabla F_{\mathbf{V}} \end{pmatrix}, \begin{pmatrix} \mathbf{X} - \mathbf{U} \\ \mathbf{Y} - \mathbf{V} \end{pmatrix} \right\rangle \geq 0, \ \forall \mathbf{X} \in \mathbb{R}_+^{m \times r}, \mathbf{Y} \in \mathbb{R}_+^{r \times n}$$

Điều kiện này tương đương với các điều kiện KKT (2.3a), (2.3b) và (2.3c).

Ví dụ 2.1. *Xét bài toán không âm đơn giản nhất trong đó ma trận* $\mathbf{A}$ *là một số* a*. Khi đó bài toán đặt ra là tìm hai số* x *và* y *sao cho tích của chúng xấp xỉ với* a*. Bài toán (2.1c) chỉ ra rằng có đúng một xấp xỉ* $(a - xy)^2 = 0$ *và có vô số các nghiệm được đưa ra bởi đồ thị* $xy = a$ *(Hình 2.1).*



Hình 2.1: Đồ thị $a = xy$

2.2.3 Đặc trưng của cực tiểu địa phương

Điều kiện (2.3c) giúp mô tả các điểm dừng của bài toán NMF. Tổng hợp tất cả các phần tử của một trong các điều kiện (2.3c), ta được:

$$0 = \sum_{ia} \left(\mathbf{U} \circ \left(\mathbf{U}\mathbf{V}\mathbf{V}^\top - \mathbf{A}\mathbf{V}^\top \right) \right)_{ia} = \left\langle \mathbf{U}, \mathbf{U}\mathbf{V}\mathbf{V}^\top - \mathbf{A}\mathbf{V}^\top \right\rangle$$
$$= \langle \mathbf{U}\mathbf{V}, \mathbf{U}\mathbf{V} - \mathbf{A} \rangle \tag{2.4}$$

Do đó, ta có một số đặc trưng cơ bản về nghiệm của bài toán NMF như sau:

Định lý 2.1. *Giả sử* $(\mathbf{U}, \mathbf{V})$ *là một điểm dừng của bài toán NMF thì ta có* $\mathbf{U}\mathbf{V} \in B\left(\frac{\mathbf{A}}{2}, \frac{1}{2}\|\mathbf{A}\|_F\right)$ *hình cầu có tâm tại* $\frac{\mathbf{A}}{2}$ *và bán kính bằng* $\frac{1}{2}\|\mathbf{A}\|_F$.

Chứng minh:

Từ (2.4), ta có: $\langle \mathbf{U}\mathbf{V}, \mathbf{A} \rangle = \langle \mathbf{U}\mathbf{V}, \mathbf{U}\mathbf{V} \rangle$
Do đó:

$$\left\| \frac{\mathbf{A}}{2} - \mathbf{U}\mathbf{V} \right\|^2 = \left\langle \frac{\mathbf{A}}{2} - \mathbf{U}\mathbf{V}, \frac{\mathbf{A}}{2} - \mathbf{U}\mathbf{V} \right\rangle$$
$$= \left\langle \frac{\mathbf{A}}{2}, \frac{\mathbf{A}}{2} \right\rangle - 2\left\langle \frac{\mathbf{A}}{2}, \mathbf{U}\mathbf{V} \right\rangle + \langle \mathbf{U}\mathbf{V}, \mathbf{U}\mathbf{V} \rangle = \left\langle \frac{\mathbf{A}}{2}, \frac{\mathbf{A}}{2} \right\rangle = \frac{1}{4}\|\mathbf{A}\|_F^2$$

tức là $\mathbf{U}\mathbf{V} \in B\left(\frac{\mathbf{A}}{2}, \frac{1}{2}\|\mathbf{A}\|_F\right)$ □

Định lý 2.2. *Giả sử* $(\mathbf{U}, \mathbf{V})$ *là một điểm dừng của bài toán NMF thì*

$$\frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2 = \frac{1}{2}\left(\|\mathbf{A}\|_F^2 - \|\mathbf{UV}\|_F^2\right).$$

Chứng minh:

Từ (2.4), ta có: $\langle \mathbf{UV}, \mathbf{A}\rangle = \langle \mathbf{UV}, \mathbf{UV}\rangle$
Do đó:

$$\begin{aligned}\frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2 &= \frac{1}{2}\langle \mathbf{A}-\mathbf{UV}, \mathbf{A}-\mathbf{UV}\rangle = \frac{1}{2}\left(\|\mathbf{A}\|_F^2 - 2\langle \mathbf{UV}, \mathbf{A}\rangle + \|\mathbf{UV}\|_F^2\right)\\ &= \frac{1}{2}\left(\|\mathbf{A}\|_F^2 - \|\mathbf{UV}\|_F^2\right)\end{aligned}$$

□

Chú ý 2.4. *Định lý 2.2 cho thấy rằng tại điểm dừng* $(\mathbf{U}, \mathbf{V})$ *của bài toán NMF, ta có điều kiện sau:* $\|\mathbf{A}\|_F^2 \geq \|\mathbf{UV}\|_F^2$. *Dấu " = " của bất đẳng thức chỉ xảy ra khi có một phân tích chính xác (tức* $\mathbf{A} = \mathbf{UV}$*).*

2.3 Quy tắc cập nhật nhân (MUR)

Một trong những thuật toán phổ biến nhất để giải bài toán NMF là quy tắc cập nhật nhân (Multiplicative Update Rules - MUR) được đưa ra bởi Lee và Seung vào năm 1999 và 2001 [8].

Xét hàm mục tiêu:

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{A}-\mathbf{UV}\|_{\mathbf{F}}^2$$

Với ma trận $\mathbf{V}$ cố định, ta có gradient của $f(\mathbf{U}, \mathbf{V})$ theo $\mathbf{U}$:

$$\nabla_{\mathbf{U}} f(\mathbf{U}, \mathbf{V}) = -\left(\mathbf{A}-\mathbf{UV}\right)\mathbf{V}^\top$$

Tương tự, với ma trận $\mathbf{U}$ cố định, ta có gradient của $f(\mathbf{U}, \mathbf{V})$ theo $\mathbf{V}$:

$$\nabla_{\mathbf{V}} f(\mathbf{U}, \mathbf{V}) = -\mathbf{U}^\top\left(\mathbf{A}-\mathbf{UV}\right)$$

Các ma trận $\mathbf{U}$ và $\mathbf{V}$ được cập nhật theo quy tắc sau:

- Cập nhật $\mathbf{U}$:

$$\mathbf{U} \longleftarrow \mathbf{U} - \alpha\nabla_{\mathbf{U}} f(\mathbf{U}, \mathbf{V})$$
$$\mathbf{U} \longleftarrow \mathbf{U} + \alpha\left(\mathbf{A}-\mathbf{UV}\right)\mathbf{V}^\top$$

- Cập nhật $\mathbf{V}$:

$$\mathbf{V} \longleftarrow \mathbf{V} - \beta\nabla_{\mathbf{V}} f(\mathbf{U}, \mathbf{V})$$
$$\mathbf{V} \longleftarrow \mathbf{V} + \beta\mathbf{U}^\top\left(\mathbf{A}-\mathbf{UV}\right)$$

Trong đó, α và β là các hệ số học (learning rate) cho các ma trận $\mathbf{U}$ và $\mathbf{V}$. Lee và Seung đã chọn các hệ số học α, β lần lượt như sau:

$$\alpha = \frac{\mathbf{U}}{\mathbf{U}\mathbf{V}\mathbf{V}^\top} \quad \text{và} \quad \beta = \frac{\mathbf{V}}{\mathbf{U}^\top\mathbf{U}\mathbf{V}}$$

Khi đó, ta thu được công thức cập nhật $\mathbf{U}, \mathbf{V}$ như sau:

$$\mathbf{U} \longleftarrow \mathbf{U} \circ \frac{\mathbf{A}\mathbf{V}^\top}{\mathbf{U}\mathbf{V}\mathbf{V}^\top} \quad \text{và} \quad \mathbf{V} \longleftarrow \mathbf{V} \circ \frac{\mathbf{U}^\top\mathbf{A}}{\mathbf{U}^\top\mathbf{U}\mathbf{V}}$$

Thuật toán 1. *Quy tắc cập nhật nhân (MUR)*

Bước 1. *Khởi tạo hai ma trận không âm ban đầu là* $\mathbf{U}^{(0)}$ *và* $\mathbf{V}^{(0)}$. *Gán* $k := 0$

Bước 2.
- *Cập nhật* $\mathbf{U}, \mathbf{V}$ *bằng công thức:*

$$\mathbf{U}^{(k+1)} = \mathbf{U}^{(k)} \circ \frac{\mathbf{A}\left(\mathbf{V}^{(k)}\right)^\top}{\mathbf{U}^{(k)}\mathbf{V}^{(k)}\left(\mathbf{V}^{(k)}\right)^\top}$$

$$\mathbf{V}^{(k+1)} = \mathbf{V}^{(k)} \circ \frac{\left(\mathbf{U}^{(k+1)}\right)^\top\mathbf{A}}{\left(\mathbf{U}^{(k+1)}\right)^\top\mathbf{U}^{(k+1)}\mathbf{V}^{(k)}}$$

- *Gán* $k := k + 1$

Bước 3. *Lặp lại Bước 2 cho đến khi đạt điều kiện dừng.*

Chú ý 2.5.

- *Trong quá trình tính toán, nếu tại lần lặp thứ* k: $\mathbf{U}^{(k)}\mathbf{V}^{(k)}\left(\mathbf{V}^{(k)}\right)^\top = 0$ *hoặc* $\left(\mathbf{U}^{(k+1)}\right)^\top\mathbf{U}^{(k+1)}\mathbf{V}^{(k)} = 0$ *thì người ta thường thay nó bằng một số* $\alpha > 0$ *tương đối nhỏ.*
- *Tại mỗi bước* k, *ta luôn có* $\mathbf{U}^{(k)}, \mathbf{V}^{(k)} \geq 0$.
- *Điều kiện dừng của Thuật toán 1 có thể được lấy như sau* $\|\mathbf{A} - \mathbf{U}\mathbf{V}\|_\mathbf{F}^2 < \epsilon$ *với* ϵ *là sai số cho trước.*

2.4 Định lý hội tụ

Định lý 2.3. *Hàm mục tiêu* $\|\mathbf{A} - \mathbf{U}\mathbf{V}\|_\mathbf{F}^2$ *là không tăng khi thực hiện quy tắc cập nhật nhân theo Thuật toán 1. Cụ thể, với mỗi bước cập nhật* k, *ta có:*

$$\left\|\mathbf{A} - \mathbf{U}^{(\mathbf{k+1})}\mathbf{V}^{(\mathbf{k+1})}\right\|_\mathbf{F}^2 \leq \left\|\mathbf{A} - \mathbf{U}^{(\mathbf{k})}\mathbf{V}^{(\mathbf{k})}\right\|_\mathbf{F}^2$$

hoặc

$$(\mathbf{U}^{(\mathbf{k+1})}, \mathbf{V}^{(\mathbf{k+1})}) \leq (\mathbf{U}^{(\mathbf{k})}, \mathbf{V}^{(\mathbf{k})}).$$

Để chứng minh Định lý 2.3, ta cần sử dụng khái niệm hàm phụ và các kết quả được trình bày sau đây:

Định nghĩa 2.2. *$G(\mathbf{V}, \mathbf{V}')$ là một hàm phụ của $F(\mathbf{V})$ trên tập $M = \{\mathbf{V} \in \mathbb{R}_+^{r\times n}\}$ nếu thỏa mãn các điều kiện sau:*

$$G(\mathbf{V}, \mathbf{V}') \geq F(\mathbf{V}) \ \ \forall \mathbf{V}, \mathbf{V}' \in M \quad \text{và} \quad G(\mathbf{V}, \mathbf{V}) = F(\mathbf{V}) \ \ \forall \mathbf{V} \in M \tag{2.5}$$

Hệ quả 2.1. *Nếu G là một hàm phụ của F trên tập M thì F sẽ không tăng theo quy tắc*

$$\mathbf{V}^{(t+1)} = \arg\min_{\mathbf{V}\in M} G(\mathbf{V}, \mathbf{V}^{(t)}) \tag{2.6}$$

với $\mathbf{V}^{(t)} \in M$.

Chứng minh:

Theo điều kiện (2.5), ta có: $G(\mathbf{V}, \mathbf{V}) = F(\mathbf{V}) \ \ \forall \mathbf{V} \in M$. Với $\mathbf{V} = \mathbf{V}^{(t)}$, ta có:

$$G(\mathbf{V}^{(t)}, \mathbf{V}^{(t)}) = F(\mathbf{V}^{(t)}) \tag{2.6a}$$

Theo điều kiện (2.5), ta có: $G(\mathbf{V}, \mathbf{V}') \geq F(\mathbf{V}) \ \ \forall \mathbf{V}, \mathbf{V}' \in M$. Với $\mathbf{V} = \mathbf{V}^{(t+1)}$ và $\mathbf{V}' = \mathbf{V}^{(t)}$, ta có:

$$G(\mathbf{V}^{(t+1)}, \mathbf{V}^{(t)}) \geq F(\mathbf{V}^{(t+1)}) \tag{2.6b}$$

Theo quy tắc (2.6), ta có:

$$G(\mathbf{V}^{(t+1)}, \mathbf{V}^{(t)}) \leq G(\mathbf{V}^{(t)}, \mathbf{V}^{(t)}) \tag{2.6c}$$

Kết hợp (2.6a), (2.6b) và (2.6c), ta có:

$$F(\mathbf{V}^{(t+1)}) \leq G(\mathbf{V}^{(t+1)}, \mathbf{V}^{(t)}) \leq G(\mathbf{V}^{(t)}, \mathbf{V}^{(t)}) = F(\mathbf{V}^{(t)})$$

Vì vậy, nếu G là một hàm phụ của F trên tập M thì F sẽ không tăng theo quy tắc (2.6), tức $F(\mathbf{V}^{(t+1)}) \leq F(\mathbf{V}^{(t)})$. □

Mệnh đề 2.2. *Với $\mathbf{V}^{(t)}) > 0$, ta xét $K(\mathbf{V}^{(t)})$ là ma trận đường chéo cấp r có*

$$\left[K(\mathbf{V}^{(t)})\right]_{kk} = \frac{\left(\mathbf{U}^\top \mathbf{U}\mathbf{V}^{(t)}\right)_{k:}}{\mathbf{V}^{(t)}_{k:}} \tag{2.7}$$

với $k = \overline{1,r}$ *thì*

$$G(\mathbf{V}, \mathbf{V}^{(t)}) = F(\mathbf{V}^{(t)}) + (\mathbf{V} - \mathbf{V}^{(t)})^\top \nabla F(\mathbf{V}^{(t)}) + \frac{1}{2}(\mathbf{V} - \mathbf{V}^{(t)})^\top K(\mathbf{V}^{(t)})(\mathbf{V} - \mathbf{V}^{(t)}) \tag{2.8}$$

là một hàm phụ của

$$F(\mathbf{V}) = \frac{1}{2}\sum_{i=1}^{m}\left(A_{i:} - \sum_{k=1}^{r}\mathbf{U}_{ik}\mathbf{V}_{k:}\right)^2 \tag{2.9}$$

trên tập $\{\mathbf{V} > 0\}$.

Chứng minh:

Dễ thấy $G(\mathbf{V}, \mathbf{V}) = F(\mathbf{V})$ nên ta chỉ cần chứng minh rằng $G(\mathbf{V}, \mathbf{V}^{(t)}) \geq F(\mathbf{V})$. Khai triển Taylor bậc hai của hàm F quanh điểm $\mathbf{V}^{(t)}$ ta được:

$$F(\mathbf{V}) = F(\mathbf{V}^{(t)}) + (\mathbf{V} - \mathbf{V}^{(t)})^\top \nabla F(\mathbf{V}^{(t)}) + \frac{1}{2}(\mathbf{V} - \mathbf{V}^{(t)})^\top \left(\mathbf{U}^\top\mathbf{U}\right)(\mathbf{V} - \mathbf{V}^{(t)}) \tag{2.9a}$$

So sánh (2.8) và (2.9a), ta thấy rằng $G(\mathbf{V}, \mathbf{V}^{(t)}) \geq F(\mathbf{V})$ với mọi $\mathbf{V}^{(t)} > 0$ khi và chỉ khi

$$(\mathbf{V} - \mathbf{V}^{(t)})^\top \left(K(\mathbf{V}^{(t)}) - \mathbf{U}^\top\mathbf{U}\right)(\mathbf{V} - \mathbf{V}^{(t)}) \geq 0 \quad \forall \mathbf{V}^{(t)} > 0 \tag{2.9b}$$

Để chứng minh (2.9b) thì ta chỉ cần chứng minh rằng ma trận $K(\mathbf{V}^{(t)}) - \mathbf{U}^\top\mathbf{U}$ là nửa xác định dương.

Giả sử $\mathbf{V} - \mathbf{V}^{(t)} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r1} & x_{r2} & \cdots & x_{rn} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1:} \\ \mathbf{x}_{2:} \\ \vdots \\ \mathbf{x}_{r:} \end{pmatrix}$ nên với vector $\mathbf{x} \in \mathbb{R}^r$ ta có thể coi $\mathbf{x} = (\mathbf{x}_{1:} \quad \mathbf{x}_{2:} \quad \cdots \quad \mathbf{x}_{r:})^\top$. Khi đó ta có:

$$\mathbf{x}^\top \left(K(\mathbf{V}^{(t)}) - \mathbf{U}^\top\mathbf{U}\right)\mathbf{x}$$

$$= \mathbf{x}^\top \left[\begin{pmatrix} \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{1:}}{\mathbf{V}^{(t)}_{1:}} & 0 & \cdots & 0 \\ 0 & \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{2:}}{\mathbf{V}^{(t)}_{2:}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{r:}}{\mathbf{V}^{(t)}_{r:}} \end{pmatrix} - \mathbf{U}^\top\mathbf{U}\right]\mathbf{x}$$

$$
\begin{aligned}
&= \mathbf{x}^\top \begin{pmatrix} \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{1:}}{\mathbf{V}^{(t)}_{1:}} & 0 & \cdots & 0 \\ 0 & \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{2:}}{\mathbf{V}^{(t)}_{2:}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{r:}}{\mathbf{V}^{(t)}_{r:}} \end{pmatrix} \mathbf{x} - \mathbf{x}^\top\mathbf{U}^\top\mathbf{U}\mathbf{x} \\
&= \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{1:}}{\mathbf{V}^{(t)}_{1:}}\mathbf{x}^2_{1:} + \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{2:}}{\mathbf{V}^{(t)}_{2:}}\mathbf{x}^2_{2:} + \cdots + \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{r:}}{\mathbf{V}^{(t)}_{r:}}\mathbf{x}^2_{r:} - \mathbf{x}^\top\mathbf{U}^\top\mathbf{U}\mathbf{x} \\
&= \sum_{i=1}^{r} \frac{\left(\mathbf{U}^\top\mathbf{U}\mathbf{V}^{(t)}\right)_{i:}}{\mathbf{V}^{(t)}_{i:}}\mathbf{x}^2_{i:} - \mathbf{x}^\top\mathbf{U}^\top\mathbf{U}\mathbf{x} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} \frac{\left(\mathbf{U}^\top\mathbf{U}\right)_{ij}\mathbf{V}^{(t)}_{j:}}{\mathbf{V}^{(t)}_{i:}}\mathbf{x}^2_{i:} - \sum_{i=1}^{r}\sum_{j=1}^{r}\mathbf{x}_{i:}\left(\mathbf{U}^\top\mathbf{U}\right)_{ij}\mathbf{x}_{j:} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r}\left(\mathbf{U}^\top\mathbf{U}\right)_{ij}\left(\frac{\mathbf{V}^{(t)}_{j:}}{\mathbf{V}^{(t)}_{i:}}\mathbf{x}^2_{i:} - \mathbf{x}_{i:}\mathbf{x}_{j:}\right) \\
&= \sum_{i=\overline{1,r}}^{j=\overline{1,r}} \frac{1}{2}\left(\mathbf{U}^\top\mathbf{U}\right)_{ij}\left(\frac{\mathbf{V}^{(t)}_{j:}}{\mathbf{V}^{(t)}_{i:}}\mathbf{x}^2_{i:} - \mathbf{x}_{i:}\mathbf{x}_{j:} + \frac{\mathbf{V}^{(t)}_{i:}}{\mathbf{V}^{(t)}_{j:}}\mathbf{x}^2_{j:} - \mathbf{x}_{i:}\mathbf{x}_{j:}\right) \\
&= \sum_{i=\overline{1,r}}^{j=\overline{1,r}} \frac{1}{2}\left(\mathbf{U}^\top\mathbf{U}\right)_{ij}\left(\sqrt{\frac{\mathbf{V}^{(t)}_{j:}}{\mathbf{V}^{(t)}_{i:}}}\mathbf{x}_{i:} - \sqrt{\frac{\mathbf{V}^{(t)}_{i:}}{\mathbf{V}^{(t)}_{j:}}}\mathbf{x}_{j:}\right)^2 \geq 0 \qquad \forall \mathbf{V}^{(t)} > 0
\end{aligned}
$$

Vậy ta có $\mathbf{x}^\top\left(K(\mathbf{V}^{(t)}) - \mathbf{U}^\top\mathbf{U}\right)\mathbf{x} \geq 0$ nên ma trận $K(\mathbf{V}^{(t)}) - \mathbf{U}^\top\mathbf{U}$ là nửa xác định dương với mọi $\mathbf{V}^{(t)} > 0$. □

Quay lại chứng minh Định lý 2.3 đã nêu ở trên.

Chứng minh:

Từ (2.6) ta có:

$$\mathbf{V}^{(t+1)} = \underset{\mathbf{V}>0}{\operatorname{argmin}}\, G(\mathbf{V}, \mathbf{V}^{(t)}) \Rightarrow \nabla_{\mathbf{V}} G(\mathbf{V}^{(t+1)}, \mathbf{V}^{(t)}) = 0$$

Mặt khác, từ (2.8) ta có:

$$\nabla_{\mathbf{V}} G(\mathbf{V}, \mathbf{V}^{(t)}) = \nabla F(\mathbf{V}^{(t)}) + K(\mathbf{V}^{(t)})(\mathbf{V} - \mathbf{V}^{(t)})$$

Với $\mathbf{V} = \mathbf{V}^{(t+1)}$, ta có:

$$\nabla_{\mathbf{V}} G(\mathbf{V}^{(t+1)}, \mathbf{V}^{(t)}) = \nabla F(\mathbf{V}^{(t)}) + K(\mathbf{V}^{(t)})(\mathbf{V}^{(t+1)} - \mathbf{V}^{(t)}) = 0$$
$$\Leftrightarrow \mathbf{V}^{(t+1)} = \mathbf{V}^{(t)} - \left(K(\mathbf{V}^{(t)})\right)^{-1} \nabla F(\mathbf{V}^{(t)})$$

Từ (2.9) ta có:

$$\nabla F(\mathbf{V}) = \mathbf{U}^\top \mathbf{U}\mathbf{V} - \mathbf{U}^\top \mathbf{A}$$

Do đó:

$$\mathbf{V}_{k:}^{(t+1)} = \mathbf{V}_{k:}^{(t)} - \frac{\mathbf{V}_{k:}^{(t)}}{\left(\mathbf{U}^\top \mathbf{U}\mathbf{V}^{(t)}\right)_{k:}} \left[\left(\mathbf{U}^\top \mathbf{U}\mathbf{V}^{(t)}\right)_{k:} - \left(\mathbf{U}^\top \mathbf{A}\right)_{k:}\right] = \mathbf{V}_{k:}^{(t)} \frac{\left(\mathbf{U}^\top \mathbf{A}\right)_{k:}}{\left(\mathbf{U}^\top \mathbf{U}\mathbf{V}^{(t)}\right)_{k:}}$$

Vì vậy, hàm $\|\mathbf{A} - \mathbf{U}\mathbf{V}\|_{\mathbf{F}}^2$ là không tăng theo quy tắc cập nhật cho $\mathbf{V}$.

Bằng cách đổi vai trò của $\mathbf{U}$ và $\mathbf{V}$ trong Hệ quả 2.1 và Mệnh đề 2.2, ta cũng có thể khẳng định rằng $\|\mathbf{A} - \mathbf{U}\mathbf{V}\|_{\mathbf{F}}^2$ là hàm không tăng theo quy tắc cập nhật cho $\mathbf{U}$. □

Tóm lại, qua việc chứng minh rằng hàm mục tiêu $\|\mathbf{A} - \mathbf{U}\mathbf{V}\|_{\mathbf{F}}^2$ không tăng, ta có thể khẳng định rằng hàm này sẽ hội tụ về một giá trị giới hạn. Điều này có nghĩa là qua nhiều lần lặp, hàm mục tiêu sẽ tiến gần hơn đến một giá trị tối ưu (tối thiểu) và các ma trận $\mathbf{U}$, $\mathbf{V}$ khi đó sẽ phản ánh một phân tách tốt nhất cho ma trận dữ liệu $\mathbf{A}$.

Chương 3

Lập trình và ứng dụng của NMF

Trong chương này sẽ mô tả bộ cơ sở dữ liệu MovieLens 100K, lập trình xây dựng mô hình NMF dựa theo thuật toán đã được nêu ở Chương 2, đánh giá mô hình xây dựng và ứng dụng của mô hình trong việc đưa ra danh sách các bộ phim gợi ý cho người dùng. Nội dung của chương được tham khảo chủ yếu từ các tài liệu [11], [14] và [15].

3.1 Cơ sở dữ liệu MovieLens 100K

Bộ cơ sở dữ liệu MovieLens 100K (https://grouplens.org/datasets/movielens/100k/) được công bố vào năm 1998 bởi GroupLens (https://grouplens.org/). Bộ cơ sở dữ liệu này bao gồm 100000 lượt đánh giá (1-5) từ 943 người dùng cho 1682 bộ phim. Mô tả chi tiết một số file dữ liệu:

- **u.data**: chứa toàn bộ các đánh giá của 943 người dùng cho 1682 bộ phim, bao gồm: *user_id* (ID người dùng), *item_id* (ID phim), *rating* (Điểm đánh giá) và *timestamp* (Thời gian đánh giá - Số giây tính từ 00:00:00 UTC 1-1-1970). Mỗi người dùng đánh giá ít nhất 20 bộ phim.

Bảng 3.1: Điểm đánh giá phim của người dùng

	user_id	item_id	rating	timestamp	thời gian chuyển đổi
0	196	242	3	881250949	1997-12-04 15:55:49
1	186	302	3	891717742	1998-04-04 19:22:22
2	22	377	1	878887116	1997-11-07 07:18:36
3	244	51	2	880606923	1997-11-27 05:02:03
4	166	346	1	886397596	1998-02-02 05:33:16
...	...	...	...	...	...
99997	276	1090	1	874795795	1997-09-20 22:49:55
99998	13	225	2	882399156	1997-12-17 22:52:36
99999	12	203	3	879959583	1997-11-19 17:13:03

- **u.item**: chứa thông tin về 1682 bộ phim, bao gồm: *item_id* (ID phim), *title* (Tên phim), *release_date* (Ngày phát hành) và các thể loại phim.

Bảng 3.2: Thông tin về 1682 bộ phim

item_id	title	release_date	...	Drama	...	Romance	...
1	Toy Story (1995)	01-Jan-1995	...	0	...	0	...
2	GoldenEye (1995)	01-Jan-1995	...	0	...	0	...
3	Four Rooms (1995)	01-Jan-1995	...	0	...	0	...
4	Get Shorty (1995)	01-Jan-1995	...	1	...	0	...
5	Copycat (1995)	01-Jan-1995	...	1	...	0	...
6	Shanghai Triad (1995)	01-Jan-1995	...	1	...	0	...
...	...	...	...	...	...	...	...
1679	B. Monkey (1998)	06-Feb-1998	...	0	...	1	...
1680	Sliding Doors (1998)	01-Jan-1998	...	1	...	1	...
1681	You So Crazy (1994)	01-Jan-1994	...	0	...	0	...
1682	Scream of Stone (1991)	08-Mar-1996	...	1	...	0	...

- **u.genre**: chứa tên của 19 thể loại phim, bao gồm: không xác định (*unknown*), Hành động (*Action*), Phiêu lưu (*Adventure*), Hoạt hình (*Animation*), Trẻ em (*Children's*), Hài kịch (*Comedy*), Tội phạm (*Crime*), Phim tài liệu (Documentary), Chính kịch (*Drama*), Giả tưởng (*Fantasy*), Phim đen (*Film-Noir*), Kinh dị (*Horror*), Nhạc kịch (*Musical*), Bí ẩn (*Mystery*), Lãng mạn (*Romance*), Khoa học viễn tưởng (*Sci-Fi*), Giật gân (*Thriller*), Chiến tranh (*War*), Viễn Tây (*Western*).
- **u.user**: chứa thông tin về 943 người dùng, bao gồm: *user_id* (ID người dùng), *age* (Tuổi), *gender* (Giới tính), *occupation* (Nghề nghiệp) và *zip_code* (Mã bưu điện).

Bảng 3.3: Thông tin về 943 người dùng

user_id	age	gender	occupation	zip_code
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
...	...	...	...	...
939	26	F	student	33319
940	32	M	administrator	02215
941	20	M	student	97229
942	48	F	librarian	78209
943	22	M	student	77841

- **u.occupation**: danh sách các nghề nghiệp, bao gồm: quản trị viên (*administrator*), nghệ sĩ (*artist*), bác sĩ (*doctor*), sư phạm (*educator*), kỹ sư (*engineer*), ngành giải trí (*entertainment*), điều hành (*executive*), chăm sóc sức khỏe (*healthcare*), nội trợ (*homemaker*), luật sư (*lawyer*), thủ thư (*librarian*), tiếp thị (*marketing*), lập trình viên (*programmer*), đã nghỉ hưu (*retired*), nhân viên bán hàng (*salesman*), nhà khoa học (*scientist*), học sinh/sinh viên (*student*), kỹ thuật viên (*technician*), nhà văn (*writer*).

3.2 Xây dựng mô hình NMF

3.2.1 Tải bộ dữ liệu MovieLens 100K

Bộ dữ liệu **MovieLens 100K** chứa thông tin về các đánh giá phim từ người dùng. Bộ dữ liệu này có thể tải xuống từ trang web của MovieLens tại địa chỉ: https://files.grouplens.org/datasets/movielens/ml-100k.zip hoặc chạy đoạn code python sau:

```
import requests
import pandas as pd
import os
import zipfile

# URL của bộ dữ liệu MovieLens 100K
url = "https://files.grouplens.org/datasets/movielens/ml-100k.zip"

# Tạo thư mục để lưu dữ liệu nếu chưa có
output_dir = "movielens_100k"
os.makedirs(output_dir, exist_ok=True)

# Tải xuống bộ dữ liệu
response = requests.get(url)
zip_file_path = os.path.join(output_dir, "ml-100k.zip")

# Lưu file zip
with open(zip_file_path, 'wb') as f:
    f.write(response.content)

# Giải nén file zip
with zipfile.ZipFile(zip_file_path, 'r') as zip_ref:
    zip_ref.extractall(output_dir)
```

3.2.2 Tiền xử lý dữ liệu

- **Đọc dữ liệu:** sử dụng thư viện *pandas* để đọc file *u.data*, file này chứa thông tin về điểm đánh giá phim của người dùng.

```
# Đường dẫn đến thư mục chứa file
data_dir = "movielens_100k/ml-100k"

# Tải dữ liệu từ file u.data
ratings_file = os.path.join(data_dir, "u.data")
column_names = ['user_id', 'item_id', 'rating', 'timestamp']

# Đọc dữ liệu từ file u.data
ratings = pd.read_csv(ratings_file, sep='\t', names=column_names)
```

- **Xây dựng ma trận đánh giá A:** hàng đại diện cho người dùng, cột đại diện cho phim và các giá trị là điểm đánh giá (1-5 hoặc 0 nếu người dùng không đánh giá hay chưa xem phim). Chuyển đổi ma trận **A** thành một mảng *NumPy* để tiện cho việc xử lý và tính toán.

```
from scipy import sparse
import numpy as np

def MaTranA(X, y, shape):
    # Lấy chỉ số hàng, cột của ma trận A
    row  = X[:,0]
    col  = X[:,1]
    # Lấy giá trị đánh giá
    data = y
    # Chuyển đổi dữ liệu thành ma trận
    matrix_sparse = sparse.csr_matrix((data,(row,col)),
                                      shape=(shape[0]+1,shape[1]+1))
    A = matrix_sparse.todense()
    A = A[1:,1:]
    # Chuyển đổi ma trận thành mảng numpy
    A = np.asarray(A)
    return A
```

- **In ra thông tin của ma trận đánh giá A:**

```
n_users = len(ratings['user_id'].unique())
n_items = len(ratings['item_id'].unique())
A_shape = (n_users, n_items)
print("Kích thước của ma trận A:", A_shape)

X = ratings[['user_id', 'item_id']].values
y = ratings['rating'].values
A = MaTranA(X, y, A_shape)
print("Ma trận đánh giá A (một phần):\n", A)
```

Bảng 3.4: Thông tin về ma trận đánh giá **A**

```
Kích thước của ma trận A: (943, 1682)
Ma trận đánh giá A (một phần):
    [[5 3 4 3 3 5 4 1 5 3 2 5 5 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [4 0 0 0 0 0 0 0 0 2 0 0 4 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [0 0 0 0 0 0 0 0 0 0 4 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [4 3 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
                  ...             ...             ...
     [0 0 0 0 0 0 0 0 5 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [0 0 0 2 0 0 4 5 3 0 0 4 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [5 0 0 0 0 0 4 0 0 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]
     [0 5 0 0 0 0 0 0 3 0 4 5 0 ... 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```

3.2.3 Xây dựng mô hình NMF

NMF giúp phân tách ma trận đánh giá **A** (thường rất lớn và thưa thớt) thành các thành phần không âm. Điều này có thể giúp xác định các yếu tố cơ bản ảnh hưởng đến sự đánh giá của người dùng đối với phim.

Mục đích của NMF:

- Phân tích sở thích người dùng: NMF cho phép hiểu rõ hơn về sở thích của người dùng bằng cách phân tách các đánh giá thành các yếu tố cơ bản. Điều này giúp xác định các thành phần mà người dùng quan tâm, từ đó cá nhân hóa các gợi ý phim.
- Gợi ý phim: Một trong những ứng dụng chính của NMF là trong hệ thống gợi ý. Bằng cách sử dụng các ma trận **U** và **V**, NMF có thể dự đoán các đánh giá cho các bộ phim mà người dùng chưa xem, từ đó cung cấp các gợi ý phim có khả năng cao sẽ được yêu thích.
- Giảm độ phức tạp dữ liệu: NMF giúp giảm kích thước dữ liệu bằng cách chuyển đổi ma trận đánh giá thành các ma trận nhỏ hơn chứa các yếu tố tiềm ẩn. Điều này làm giảm độ phức tạp tính toán và cải thiện hiệu suất của các thuật toán gợi ý.
- Xác định các xu hướng: NMF có thể giúp phát hiện các xu hướng trong sở thích người dùng hoặc các đặc trưng chung của các bộ phim. Điều này có thể hữu ích trong việc phát triển nội dung mới hoặc trong việc định hướng chiến lược marketing.
- Khám phá các mối quan hệ: NMF cho phép khám phá các mối quan hệ giữa người dùng và phim, giúp nhà sản xuất hiểu rõ hơn về cách mà các yếu tố khác nhau ảnh hưởng đến sự lựa chọn phim của người xem.

NMF với ma trận đánh giá **A**, các ma trận **U** và **V** tạo thành thể hiện:

- **Ma trận U (Ma trận người dùng):** có kích thước $m \times r$ với m là số lượng người dùng và r là số lượng thành phần tiềm ẩn (đặc trưng) của các bộ phim.
  - Mỗi hàng trong ma trận **U** đại diện cho một người dùng.
  - Mỗi cột trong ma trận **U** thể hiện mức độ sở thích của người dùng đó đối với các thành phần tiềm ẩn.
  - Cụ thể, giá trị tại $[\mathbf{U}]_{ij}$ cho thấy mức độ mà người dùng i thích thành phần j.

  $\Rightarrow$ Thể hiện sở thích của người dùng với các thành phần tiềm ẩn.
- **Ma trận V (Ma trận phim):** có kích thước $r \times n$ với n là số lượng bộ phim.
  - Mỗi cột trong ma trận **V** đại diện cho một bộ phim.
  - Mỗi hàng trong ma trận **V** thể hiện mức độ liên quan của các thành phần tiềm ẩn đến bộ phim.
  - Cụ thể, giá trị tại $[\mathbf{V}]_{jk}$ cho thấy mức độ mà thành phần j góp phần vào bộ phim k.

  $\Rightarrow$ Thể hiện mức độ ảnh hưởng của các thành phần tiềm ẩn tác động vào các bộ phim.

Các thành phần tiềm ẩn là những yếu tố hoặc đặc trưng cơ bản không thể quan sát trực tiếp, nhưng có thể giải thích được sự tương tác giữa người dùng và các bộ phim. Chúng có thể đại diện cho:

- Thể loại phim: Một số yếu tố có thể liên quan đến việc người dùng thích các thể loại phim nhất định (hành động, hài hước, lãng mạn, khoa học viễn tưởng, ...).
- Phong cách làm phim: Một số yếu tố có thể phản ánh sở thích về phong cách làm phim (độc lập, bom tấn, kinh dị, thương mại, ...).
- Diễn viên và đạo diễn: Yếu tố có thể liên quan đến sự yêu thích của người dùng đối với các diễn viên hoặc đạo diễn nhất định.
- Nội dung và chủ đề: Một số yếu tố có thể đại diện cho các chủ đề cụ thể của phim (tình yêu, gia đình, phiêu lưu, ...).
- Đối tượng khán giả: Một số yếu tố có thể phản ánh sở thích của các nhóm đối tượng khán giả khác nhau (trẻ em, thanh niên, người lớn, ...).
- Cảm xúc: Các yếu tố có thể liên quan đến cảm xúc mà phim gợi lên (hài hước, hồi hộp, buồn bã, ...).

Xây dựng mô hình NMF bằng Thuật toán 1 với số lần lặp tối đa (max_iter) là 10000 và điều kiện dừng (tol) là $\|\mathbf{A} - \mathbf{UV}\|_{\mathbf{F}}^2 < 10^{-6}$.

```
def nmf_mur(A, n_components, max_iter=10000, tol=1e-6):
    # Khởi tạo U và V ngẫu nhiên
    m, n = A.shape
    U = np.random.rand(m, n_components)
    V = np.random.rand(n_components, n)

    for k in range(max_iter):
        # Cập nhật U
        UVV = np.dot(U, np.dot(V, V.T)) # Tính toán UVV^T
        UVV[UVV == 0] = 1e-10  # Tránh chia cho 0
        U *= np.dot(A, V.T) / UVV

        # Cập nhật V
        UUV = np.dot(np.dot(U.T, U), V)  # Tính toán (U^T)UV
        UUV[UUV == 0] = 1e-10  # Tránh chia cho 0
        V *= np.dot(U.T, A) / UUV

        # Kiểm tra điều kiện dừng
        A_approx = np.dot(U, V)
        loss = np.linalg.norm(A - A_approx, 'fro')**2

        if loss < tol:
            print(f"Điều kiện dừng đạt được tại vòng lặp {k+1}.")
            break

    return U, V, A_approx
```

3.3 Đánh giá mô hình NMF

Để đánh giá mô hình NMF xây dựng ta dùng chỉ số RMSE (Root Mean Square Error) là một thước đo được sử dụng để đánh giá độ chính xác của một mô hình dự đoán. RMSE tính toán sự khác biệt giữa ma trận dự đoán và ma trận thực tế với công thức tính:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{A}_{ij} - \mathbf{A}_{approx,ij})^2}$$

Trong đó:

- $\mathbf{A}$: Ma trận thực tế chứa các giá trị đánh giá.
- $\mathbf{A}_{approx}$: Ma trận dự đoán được tạo ra từ mô hình.

- N: Số lượng phần tử khác 0 trong ma trận thực tế $\mathbf{A}$.
- m, n: Số hàng, số cột của ma trận $\mathbf{A}$.

Giá trị RMSE nhỏ hơn cho thấy mô hình có độ chính xác cao hơn.

Ví dụ 3.1. *Cho ma trận dự đoán* $\mathbf{A}_{approx} = \begin{bmatrix} 0.98 & 1.98 & 4.04 \\ 0.99 & 2.87 & 2.13 \\ 1.99 & 5.01 & 3.02 \end{bmatrix}$ *và ma trận thực tế* $\mathbf{A} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 3 & 0 \\ 2 & 0 & 0 \end{bmatrix}$*, thì RMSE của chúng là:*

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{4}\left[(1-0.98)^2 + (4-4.04)^2 + (3-2.87)^2 + (2-1.99)^2\right]} \\ &= \sqrt{0.00475} \approx 0.0689 \end{aligned}$$

Tính RMSE giữa ma trận thực tế $\mathbf{A}$ và ma trận dự đoán $\mathbf{A}_{approx}$ bằng cách tìm các phần tử không bằng 0 trong $\mathbf{A}$ và các phần tử tương ứng trong $\mathbf{A}_{approx}$. Sử dụng hàm *mean_squared_error* trong thư viện *sklearn.metrics* và lấy căn bậc 2 của hàm để tính RMSE giữa hai mảng (mảng 1 chứa các phần tử không bằng 0 trong $\mathbf{A}$ và mảng 2 chứa các phần tử trong $\mathbf{A}_{approx}$ tương ứng với vị trí mà các phần tử không bằng 0 trong $\mathbf{A}$).

```
from sklearn.metrics import mean_squared_error

def calculate_rmse(A, A_approx):
    A = A[A.nonzero()].flatten()
    A_approx = A_approx[A.nonzero()].flatten()
    return np.sqrt(mean_squared_error(A, A_approx))
```

Để tìm giá trị tối ưu r (số lượng thành phần) cho mô hình NMF ta sử dụng **phương pháp kiểm tra chéo (Cross-Validation)** [14]:

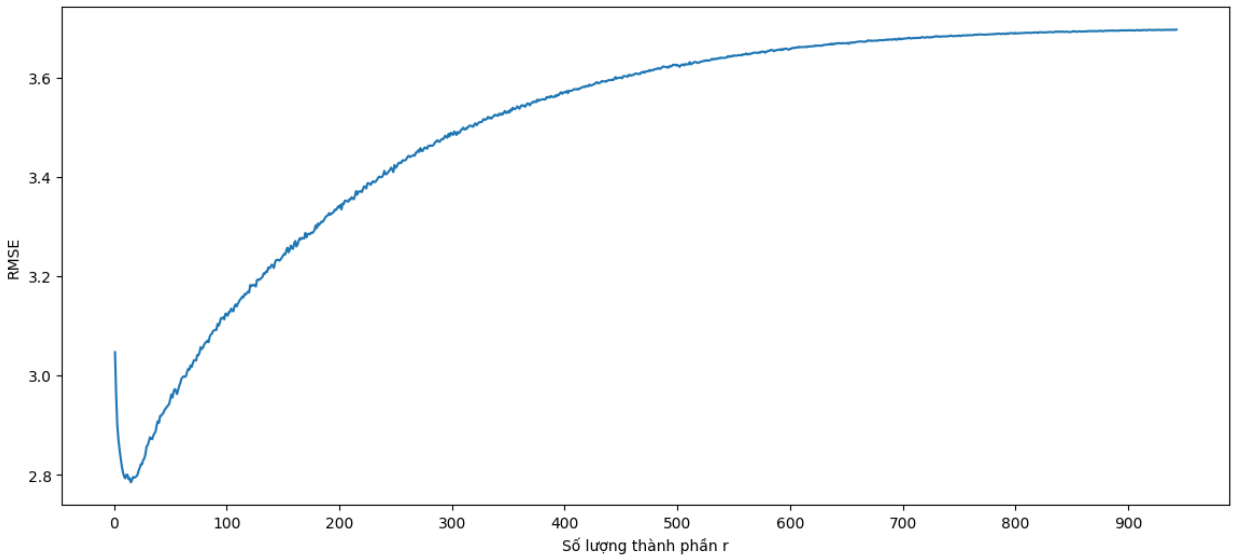
- **Chia dữ liệu:** sử dụng hàm *train_test_split* để chia ma trận dữ liệu $\mathbf{A}$ thành tập huấn luyện $\mathbf{A}_{train}$ (67%) và tập kiểm tra $\mathbf{A}_{test}$ (33%).

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.33)
A_train = MaTranA(X_train, y_train, A_shape)
A_test = MaTranA(X_test, y_test, A_shape)
```

- **Thử nghiệm với nhiều giá trị** r**:** tính RMSE cho mỗi giá trị r (r là số nguyên dương thỏa mãn $r \leq 943$) bằng cách tìm ma trận đánh giá dự

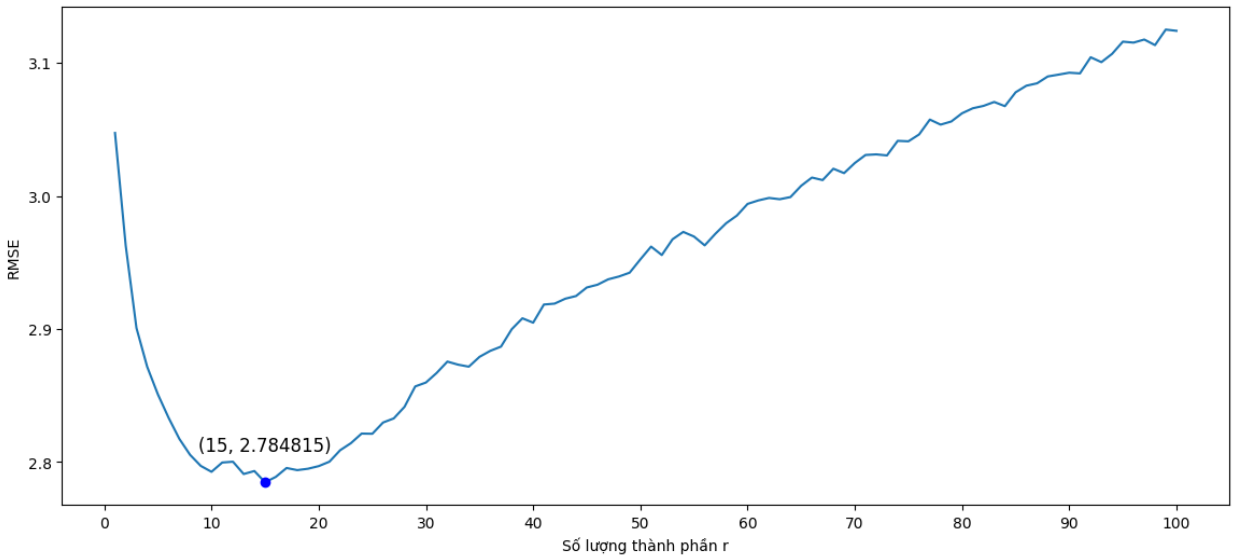
đoán $\mathbf{A}_{train,approx}$ trên tập huấn luyện và đánh giá trên tập kiểm tra (tính RMSE giữa $\mathbf{A}_{test}$ và $\mathbf{A}_{train,approx}$) (Hình 3.1).

```
for r in range(1, 944):
    U_train, V_train, A_train_approx = nmf_mur(A_train, r)

    rmse = calculate_rmse(A_test, A_train_approx)
    print(f"r = {r}, RMSE = {rmse}")
```



Hình 3.1: Giá trị RMSE của mô hình NMF

- **Chọn giá trị tối ưu** r**:** chọn giá trị r mà cho kết quả RMSE thấp nhất trên tập kiểm tra (Hình 3.2).



Hình 3.2: Giá trị tối ưu r của mô hình NMF

Như vậy, từ kết quả thu được ở Hình 3.2, mô hình NMF đạt tối ưu tại $r = 15$ và thu được các ma trận $\mathbf{U}, \mathbf{V}$ như sau:

Bảng 3.5: Ma trận $\mathbf{U}, \mathbf{V}$ tối ưu ($r = 15$)

```
Kích thước của ma trận U: (943, 15)
    [[1.754802985 0.86774089 1.808734665 ... 0.611725783 0.182258332]
     [     0           0           0      ...      0      0.607650069]
     [0.196395847       0           0      ...      0      0.319057179]
     [0.015436883 0.105574984      0      ...      0      0.223160513]
     [     0      1.270511421 2.025879862 ...      0           0     ]
                        ...                ...           ...
     [     0           0           0      ...      0           0     ]
     [     0      0.359971718 0.718974092 ... 0.14948077 0.20920653]
     [0.254158606       0      0.034675304 ...      0           0     ]
     [     0           0           0      ...      0           0     ]
     [1.145434148 1.156763836      0      ...      0           0     ]

Kích thước của ma trận V: (15, 1682)
    [[0.021545838 0.116815732 0.563002465 ...      0      0.012976871]
     [     0           0           0      ...      0           0     ]
     [0.882182974 0.194469016      0      ...      0           0     ]
     [1.549924460       0      0.009736486 ...      0           0     ]
     [1.273244896       0           0      ... 0.002602468      0     ]
                        ...                ...           ...
     [8.914109143       0      0.482282039 ... 0.002636738      0     ]
     [     0           0           0      ...      0           0     ]
     [     0      2.217719468 0.101021831 ... 0.021727777      0     ]
     [     0           0           0      ...      0      0.025255193]
     [     0           0           0      ...      0           0     ]
```

3.4 Ứng dụng của NMF - Gợi ý phim

Đọc dữ liệu trong file *u.user* để lấy thông tin về người dùng:

```
users_file = os.path.join(data_dir, "u.user")
user_column_names = ['user_id', 'age', 'gender', 'occupation',
                                                    'zip_code']
users = pd.read_csv(users_file, sep='|', names=user_column_names)
```

và in ra thông tin về người dùng theo *user_ID*:

```
user_id = 406
user_info = users[users['user_id'] == user_id].iloc[0]
print(f"Tuổi: {user_info['age']}, Giới tính: {user_info['gender']},
        Nghề nghiệp: {user_info['occupation']},
                                Mã bưu điện: {user_info['zip_code']}")
```

Bảng 3.6: Thông tin về người dùng có $user\_ID = 406$

```
Tuổi: 52, Giới tính: M, Nghề nghiệp: educator, Mã bưu điện: 93109
```

Đọc dữ liệu trong file *u.item* để lấy thông tin về bộ phim:

```
movies_file = os.path.join(data_dir, "u.item")
movie_column_names = ['item_id', 'title', 'release_date',
    'video_release_date', 'IMDb_URL', 'unknown', 'Action', 'Adventure',
    'Animation', 'Children', 'Comedy', 'Crime', Documentary', 'Drama',
    'Fantasy', 'Film-Noir', 'Horror', 'Musical', 'Mystery', 'Romance',
                                'Sci-Fi', 'Thriller', 'War', 'Western']
movies = pd.read_csv(movies_file, sep='|', names=movie_column_names,
                                                    encoding='latin-1')
```

Tìm các bộ phim mà người dùng đã xem theo $user\_ID$ và gộp chúng với các thể loại:

```
user_ratings = ratings[ratings['user_id'] == user_id]
user_movies = user_ratings.merge(movies, on='item_id')
```

để thu được danh sách các bộ phim mà người dùng $user\_ID$ đã đánh giá:

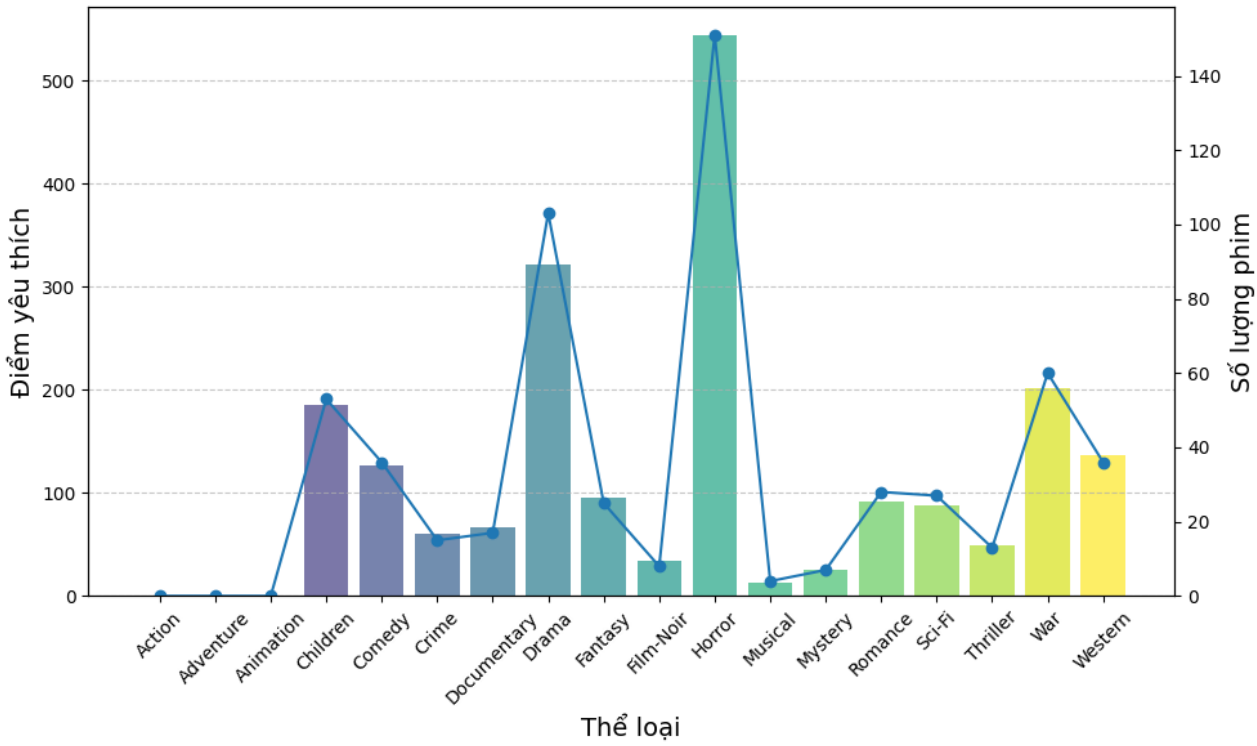
```
print(f"STT|ID phim|Tên phim|Thể loại|Điểm đánh giá")
for index, row in user_movies.iterrows():
    # Lấy danh sách tên các thể loại phim
    genres = movies.columns[6:]
    genre_list = ", ".join([genres[i] for i in range(len(genres)) if
                                                    row.iloc[i + 6] == 1])
    print(f"{index + 1}|{row['item_id']}|{row['title']}|{genre_list}|
                                                    {row['rating']}")
```

Bảng 3.7: Danh sách các bộ phim mà người dùng có $user\_ID = 406$ đã đánh giá

STT	ID phim	Tên phim	Thể loại	Điểm đánh giá
1	48	Hoop Dreams (1994)	Film-Noir	5
2	479	Vertigo (1958)	Thriller	4
3	664	Paris, Texas (1984)	Horror	2
4	823	Mulholland Falls (1996)	Fantasy, Mystery	3
5	962	Ruby in Paradise (1993)	Horror	4
6	239	Sneakers (1992)	Fantasy, Horror, Western	3
7	156	Reservoir Dogs (1992)	Fantasy	5
...	...	...	...	...
340	393	Mrs. Doubtfire (1993)	Drama	4
341	654	Chinatown (1974)	Mystery, Thriller	4
342	971	Mediterraneo (1991)	Drama	3

Để tìm thể loại phim mà người dùng $user\_ID$ yêu thích ta cần phải tính tổng các điểm đánh giá cho từng thể loại phim dựa trên các phim mà người dùng đã xem (không tính thể loại phim *unknown*):

```
def SoThich_TheLoai(user_movies, movies):
    genre_scores = {genre: 0 for genre in movies.columns[6:]}
    genre_counts = {genre: 0 for genre in movies.columns[6:]}

    for index, row in user_movies.iterrows():
        genres = movies.columns[6:]
        for i in range(len(genres)):
            if row.iloc[i + 6] == 1:
                genre_scores[genres[i]] += row['rating']
                genre_counts[genres[i]] += 1

    return genre_scores, genre_counts
```



Hình 3.3: Thể loại phim yêu thích của người dùng có $user\_ID = 406$

Nhận xét 3.1. *(Hình 3.3)*
Phân tích các thể loại yêu thích của người dùng có $user\_ID = 406$*:*

- ***Horror**: 544 điểm từ 151 bộ phim cho thấy người dùng rất thích thể loại này. Số điểm trung bình là khoảng 3.6, cho thấy họ đánh giá rất cao những bộ phim kinh dị đã xem.*
- ***Drama**: là thể loại được ưa chuộng với 321 điểm từ 103 bộ phim, điểm*

trung bình là khoảng 3.1. Đây cũng là thể loại mà người dùng có sự quan tâm đáng kể.

- ***War*** *(201 điểm từ 60 phim) và* ***Children*** *(185 điểm từ 53 phim) cũng cho thấy sự quan tâm của người dùng với điểm trung bình lần lượt là 3.35 và 3.49.*
- ***Action***, ***Adventure***, *và* ***Animation*** *không có điểm nào, có thể do người dùng chưa xem phim thuộc thể loại này hoặc không có sự quan tâm.*

Tóm lại, người dùng này có sở thích rõ ràng đối với thể loại phim ***kinh dị*** *và* ***tâm lý****, thể hiện qua số điểm cao và nhiều phim đã xem. Họ cũng có sự yêu thích nhất định với phim* ***chiến tranh*** *và* ***trẻ em****. Việc không có điểm đánh giá cho các thể loại như* ***hành động***, ***phiêu lưu*** *và* ***hoạt hình*** *có thể cho thấy họ không quan tâm tới những thể loại này hoặc chưa có cơ hội xem.*

Nhìn chung, sở thích của người dùng này thiên về những câu chuyện sâu sắc, cảm xúc hoặc hồi hộp hơn là những hành động nhanh và hài hước.

Sử dụng mô hình NMF để dự đoán điểm đánh giá cho các bộ phim mà người dùng $user\_ID$ chưa xem: các giá trị trong ma trận $\mathbf{A}_{approx} = \mathbf{UV}$ là các điểm đánh giá dự đoán cho các bộ phim mà người dùng đó chưa đánh giá.

```
U, V, A_approx = nmf_mur(A, 15)

def recommend_movies(user_id, A_approx, A, n=10):
    # Lấy điểm đánh giá của người dùng từ ma trận A
    user_ratings = A[user_id - 1]
    # Tìm các phim mà người dùng chưa đánh giá
    unrated_movies = np.where(user_ratings == 0)[0]
    # Lấy điểm dự đoán cho các phim chưa đánh giá từ ma trận A_approx
    predicted_scores = A_approx[user_id - 1, unrated_movies]
    # Giới hạn điểm dự đoán (1 - 5)
    predicted_scores = np.clip(predicted_scores, 1, 5)
    # Sắp xếp và chọn ra 10 phim có điểm cao nhất
    recommended_indices = np.argsort(predicted_scores)[::-1][:n]
    return unrated_movies[recommended_indices],
                                    predicted_scores[recommended_indices]
```

và in ra danh sách 10 bộ phim mà người dùng có khả năng thích dựa trên các điểm đánh giá đã được dự đoán:

```
recommended_movies, predicted_scores = recommend_movies(user_id,
                                                        A_approx, A)
print(f"ID phim|Tên phim|Thể loại|Điểm dự đoán")
for movie_index, score in zip(recommended_movies, predicted_scores):
    movie_info = movies[movies['item_id'] == movie_index + 1]
    # Lấy tên của bộ phim
```

```
        title = movie_info['title'].values[0]
        # Lấy danh sách tên các thể loại phim
        genres = movie_info.columns[6:]
        genre_list = [genre for genre in genres if
                                        movie_info[genre].values[0] == 1]
        print(f"{movie_index + 1}|{title}|{', '.join(genre_list)}|
                                                        {score:.2f}")
```

Bảng 3.8: Danh sách các bộ phim gợi ý cho người dùng có $user\_ID = 406$

ID phim	Tên phim	Thể loại	Điểm dự đoán
603	Rear Window (1954)	Mystery, Thriller	5.00
200	Shining, The (1980)	Horror	4.08
484	Maltese Falcon, The (1941)	Film-Noir, Mystery	4.07
423	E.T. the Extra-Terrestrial (1982)	Children, Drama, Fantasy, Sci-Fi	3.88
192	Raging Bull (1980)	Drama	3.12
83	Much Ado About Nothing (1993)	Comedy, Romance	3.11
178	12 Angry Men (1957)	Drama	3.09
659	Arsenic and Old Lace (1944)	Comedy, Mystery, Thriller	2.92
475	Trainspotting (1996)	Drama	2.90
288	Scream (1996)	Horror, Thriller	2.75

Nhận xét 3.2. *(Bảng 3.8 và Hình 3.3)*
Phân tích danh sách các bộ phim gợi ý cho người dùng có $user\_ID = 406$*:*

- ***Rear Window (1954)****: có điểm dự đoán rất cao (5.00), phù hợp với sở thích người dùng về thể loại hồi hộp và bí ẩn.*
- ***The Shining (1980)****: là một bộ phim kinh dị nổi tiếng với điểm dự đoán 4.08, phù hợp với sở thích hàng đầu của người dùng.*
- ***The Maltese Falcon (1941)****: cũng phù hợp với sở thích người dùng về thể loại bí ẩn.*
- ***E.T. the Extra-Terrestrial (1982)****: mặc dù là một bộ phim kinh điển nhưng điểm dự đoán 3.88 có thể không hoàn toàn phản ánh sở thích của người dùng vì họ không có nhiều hứng thú với thể loại khoa học viễn tưởng, giả tưởng và trẻ em.*
- ***Raging Bull (1980)****: là một tác phẩm tâm lý nổi bật, phim này hoàn toàn phù hợp với sở thích của người dùng. Điểm dự đoán 3.12 cho thấy đây là một bộ phim có giá trị nghệ thuật cao và sẽ thu hút sự quan tâm của người xem.*
- ***Much Ado About Nothing (1993)****: mặc dù đây là một bộ phim hài, thể loại không phải là sở thích chính của người dùng, nhưng điểm dự đoán 3.11 cho thấy nó vẫn có thể mang lại sự giải trí cho khán giả.*

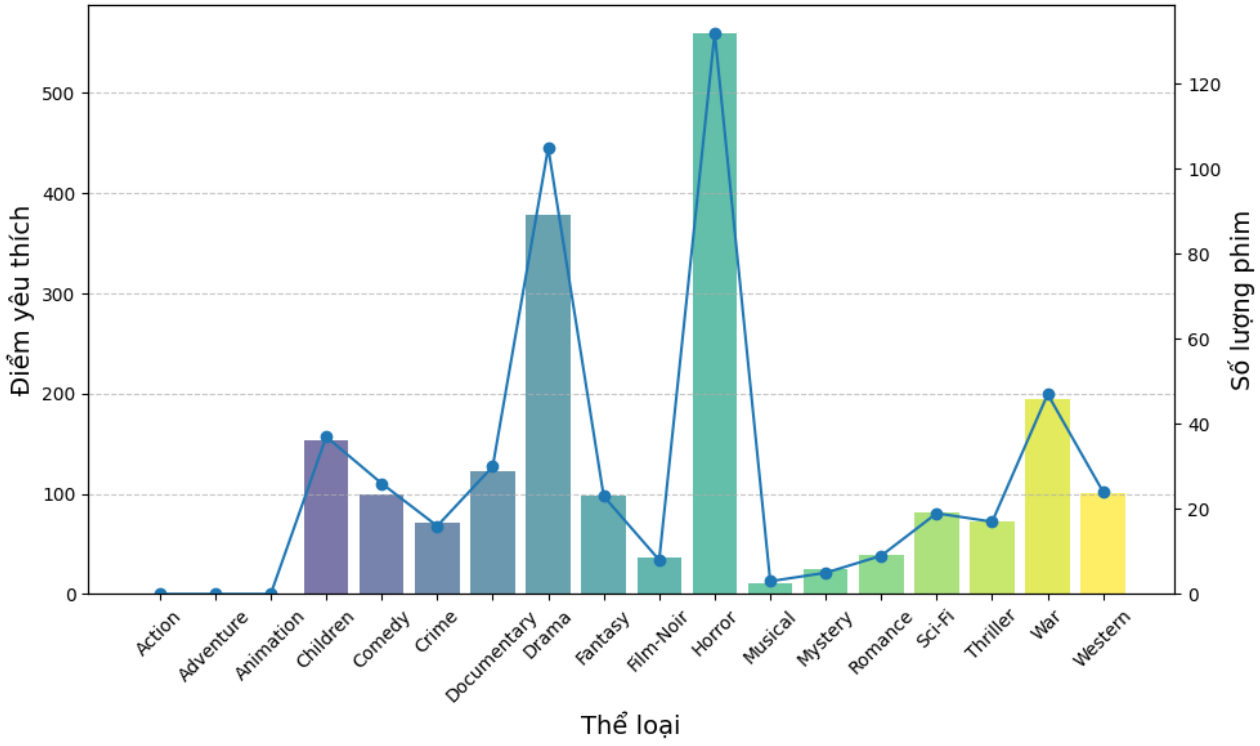
- ***12 Angry Men (1957)**: là một tác phẩm kinh điển trong thể loại tâm lý, rất phù hợp với sở thích của người dùng.*
- ***Arsenic and Old Lace (1944)**: sự kết hợp giữa hài hước và bí ẩn có thể thu hút người dùng, nhưng điểm dự đoán thấp (2.92) cho thấy đây không phải là lựa chọn hàng đầu.*
- ***Trainspotting (1996)**: là một bộ phim tâm lý mạnh mẽ, mặc dù điểm dự đoán không cao, nhưng nó vẫn đáng để người dùng xem.*
- ***Scream (1996)**: là một bộ phim kinh dị và hồi hộp, hoàn toàn phù hợp với sở thích của người dùng. Tuy nhiên, điểm dự đoán thấp (2.75) có thể cho thấy nó không phải là một tác phẩm xuất sắc nhất trong thể loại này.*

*Tóm lại, các bộ phim gợi ý phần lớn đều phù hợp với sở thích của người dùng này, đặc biệt là các phim thuộc thể loại kinh dị và tâm lý. Những bộ phim như **The Shining (1980)**, **Rear Window (1954)**, và **The Maltese Falcon (1941)** chắc chắn sẽ thu hút sự quan tâm của họ. Các lựa chọn khác như **E.T. the Extra-Terrestrial (1982)** và **Raging Bull (1980)** cũng có thể mang lại trải nghiệm thú vị. Mặc dù một số bộ phim có yếu tố hài hước như **Much Ado About Nothing (1993)** và **Arsenic and Old Lace (1944)**, không phải là sở thích chính của người dùng, nhưng chúng vẫn có thể mang đến sự giải trí cho người xem.*

Bảng 3.9: Danh sách các bộ phim mà người dùng có $user\_ID = 747$ đã đánh giá

Tuổi: 19, Giới tính: M, Nghề nghiệp: other, Mã bưu điện: 93612

STT	ID phim	Tên phim	Thể loại	Điểm đánh giá
1	228	Star Trek: The Wrath of Khan (1982)	Children, Comedy, Western	4
2	25	Birdcage, The (1996)	Drama	3
3	108	Kids in the Hall: Brain Candy (1996)	Drama	4
4	208	Young Frankenstein (1974)	Drama, Romance	5
5	23	Taxi Driver (1976)	Horror	5
6	223	Sling Blade (1996)	Horror	5
7	432	Fantasia (1940)	Crime, Documentary, Sci-Fi	5
8	48	Hoop Dreams (1994)	Film-Noir	5
9	1050	Ghost and Mrs. Muir, The (1947)	Horror, War	3
...	...	...	...	...
288	865	Ice Storm, The (1997)	Horror	5
289	1	Toy Story (1995)	Crime, Documentary, Drama	5
290	58	Quiz Show (1994)	Horror	3
291	514	Annie Hall (1977)	Drama, War	4
292	124	Lone Star (1996)	Horror, Thriller	5
293	510	Magnificent Seven, The (1954)	Children, Horror	5



Hình 3.4: Thể loại phim yêu thích của người dùng có $user\_ID = 747$

Bảng 3.10: Danh sách các bộ phim gợi ý cho người dùng có $user\_ID = 747$

ID phim	Tên phim	Thể loại	Điểm dự đoán
197	Graduate, The (1967)	Drama, Romance	4.96
191	Amadeus (1984)	Drama, Mystery	4.86
484	Maltese Falcon, The (1941)	Film-Noir, Mystery	4.51
435	Butch Cassidy and the Sundance Kid (1969)	Action, Comedy, Western	4.35
657	Manchurian Candidate, The (1962)	Film-Noir, Thriller	4.29
523	Cool Hand Luke (1967)	Comedy, Drama	3.87
527	Gandhi (1982)	Drama	3.87
143	Sound of Music, The (1965)	Musical	3.73
89	Blade Runner (1982)	Film-Noir, Sci-Fi	3.58
186	Blues Brothers, The (1980)	Action, Comedy, Musical	3.43

Nhận xét 3.3. *(Hình 3.4 và Bảng 3.10)*

*Người dùng $user\_ID = 747$ có sở thích rõ ràng đối với thể loại phim **kinh dị** và **tâm lý**, thể hiện qua số điểm cao và nhiều phim đã xem. Họ cũng có sự yêu thích nhất định với thể loại **trẻ em**, **tài liệu** và **chiến tranh**. Việc không có điểm đánh giá cho các thể loại như **hành động** và **phiêu lưu** có thể cho thấy họ không quan tâm tới những thể loại này hoặc chưa có cơ hội xem.*

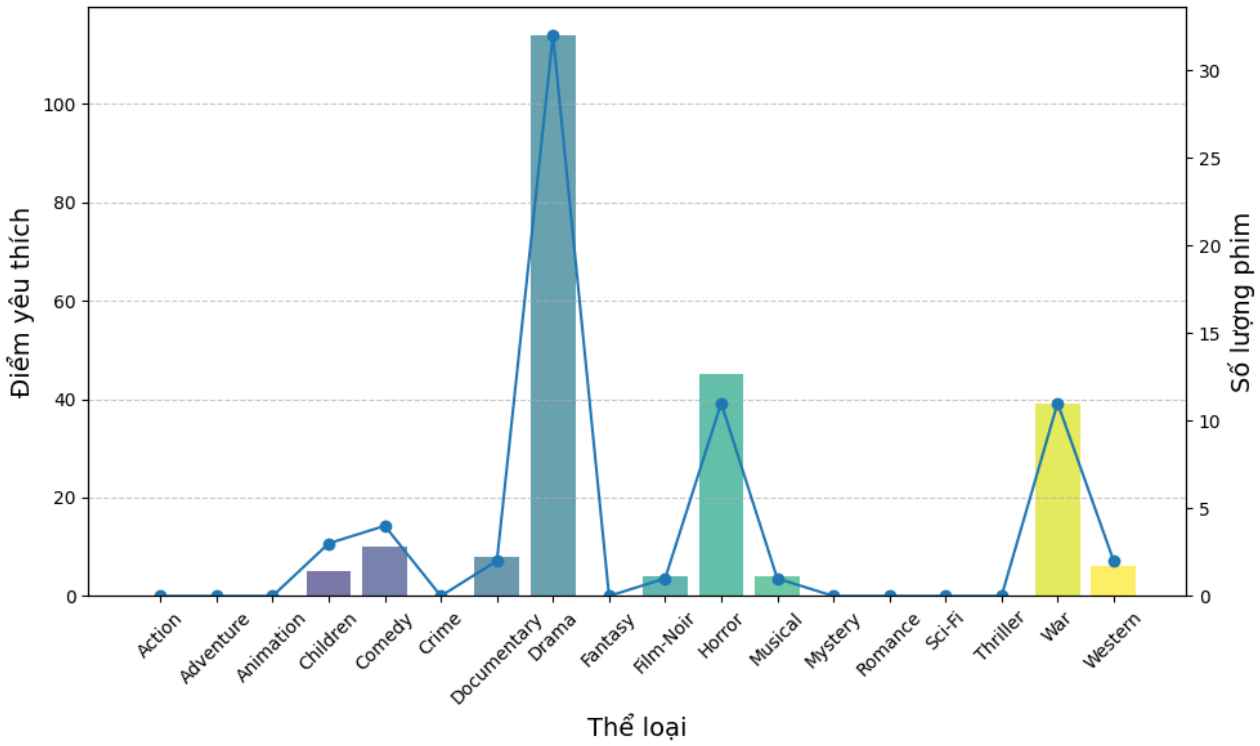
Nhìn chung, người dùng này có xu hướng thích những câu chuyện sâu sắc và hồi hộp hơn là những thể loại giải trí nhẹ nhàng hay hành động.

Các bộ phim gợi ý chủ yếu thuộc thể loại tâm lý, kinh dị và một số thể loại khác như hài, hành động và nhạc kịch, rất phù hợp với sở thích của người này. Hơn nữa, tất cả các bộ phim trong danh sách gợi ý đều có điểm dự đoán cao, cho thấy chúng là những lựa chọn tốt cho người xem. Có thể thấy rằng các bộ phim như ***The Graduate (1967)***, ***Amadeus (1984)*** *và* ***The Maltese Falcon (1941)*** *rất phù hợp với sở thích của người dùng này vì chúng thuộc các thể loại mà họ yêu thích.*

Bảng 3.11: Danh sách các bộ phim mà người dùng có $user\_ID = 196$ đã đánh giá

Tuổi: 49, Giới tính: M, Nghề nghiệp: writer, Mã bưu điện: 55105

STT	ID phim	Tên phim	Thể loại	Điểm đánh giá
1	242	Kolya (1996)	Drama	3
2	393	Mrs. Doubtfire (1993)	Drama	4
...	...	...	...	...
33	13	Mighty Aphrodite (1995)	Drama	2
34	762	Beautiful Girls (1996)	Horror	3
35	173	Princess Bride, The (1987)	Children, Comedy, Drama, War	2
36	1022	Fast, Cheap & Out of Control (1997)	Film-Noir	4
37	845	That Thing You Do (1996)	Drama	4
38	269	Full Monty, The (1997)	Drama	3
39	110	Operation Dumbo Drop (1995)	Children, Comedy, Drama	1



Hình 3.5: Thể loại phim yêu thích của người dùng có $user\_ID = 196$

Bảng 3.12: Danh sách các bộ phim gợi ý cho người dùng có $user\_ID = 196$

ID phim	Tên phim	Thể loại	Điểm dự đoán
88	Sleepless in Seattle (1993)	Comedy, Romance	1.56
216	When Harry Met Sally... (1989)	Comedy, Romance	1.43
275	Sense and Sensibility (1995)	Drama, Romance	1.33
732	Dave (1993)	Comedy, Romance	1.21
451	Grease (1978)	Comedy, Musical, Romance	1.17
100	Fargo (1996)	Crime, Drama, Thriller	1.11
283	Emma (1996)	Drama, Romance	1.07
781	French Kiss (1995)	Comedy, Romance	1.02
204	Back to the Future (1985)	Comedy, Sci-Fi	1.00
1676	War at Home, The (1996)	Drama	1.00

Nhận xét 3.4. *(Hình 3.5 và Bảng 3.12)*

Người dùng $user\_ID = 196$ *có sở thích rõ ràng đối với thể loại phim* ***tâm lý****, thể hiện qua số điểm cao và nhiều phim đã xem. Họ cũng có sự yêu thích nhất định với thể loại* ***kinh dị*** *và* ***chiến tranh****. Việc không có điểm đánh giá cho các thể loại như* ***hành động****,* ***phiêu lưu****,* ***hoạt hình****, ... có thể cho thấy họ không quan tâm tới những thể loại này hoặc chưa có cơ hội xem.*

Nhìn chung, người dùng này có sự yêu thích mạnh mẽ với những câu chuyện tâm lý sâu sắc và có chiều sâu hơn là những thể loại giải trí nhẹ nhàng hay hành động.

Tuy nhiên, hầu hết các bộ phim gợi ý trong danh sách đều thuộc thể loại hài và lãng mạn, không phải là sở thích chính của người dùng này. Hơn nữa, tất cả các bộ phim gợi ý đều có điểm dự đoán thấp (dưới 2) cho thấy rằng chúng có thể không phù hợp với sở thích của người dùng.

3.5 So sánh NMF với SVD

Phân tích giá trị kỳ dị (SVD - Singular Value Decomposition [16]) là một công cụ mạnh mẽ trong đại số tuyến tính, thường được sử dụng trong các bài toán giảm chiều dữ liệu và hệ thống gợi ý. Với một ma trận $\mathbf{A}$ kích thước $m \times n$ (trong đó m là số người dùng và n là số bộ phim), SVD phân rã ma trận này thành ba ma trận:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$$

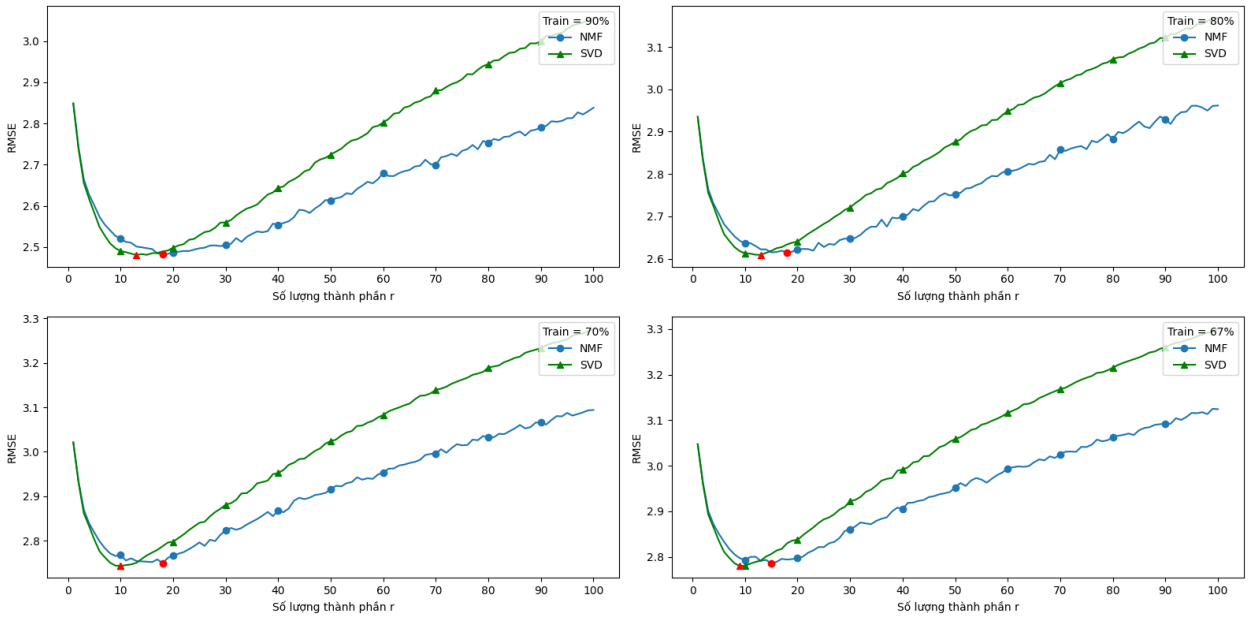
- $\mathbf{U}$: có kích thước $m \times r$ chứa các vector riêng của $\mathbf{A}\mathbf{A}^\top$, thể hiện mối quan hệ giữa người dùng và các yếu tố tiềm ẩn.
- $\mathbf{S}$: là ma trận chéo có kích thước $r \times r$, chứa các giá trị riêng lớn nhất của $\mathbf{A}$, thể hiện độ mạnh của từng yếu tố tiềm ẩn.

- $\mathbf{V}^\top$: là ma trận chuyển vị của $\mathbf{V}$ có kích thước $r \times n$, chứa các vector riêng của $\mathbf{A}^\top\mathbf{A}$, thể hiện sự tương đồng giữa các bộ phim và các yếu tố tiềm ẩn.

Mô hình SVD được xây dựng bằng cách sử dụng thư viện *scikit-learn* với hàm *TruncatedSVD* cho phép thực hiện phân tích SVD trên ma trận lớn mà không cần phải tính toán hoàn toàn các giá trị kỳ dị, giúp tiết kiệm bộ nhớ và thời gian tính toán.

```
from sklearn.decomposition import TruncatedSVD

def svd(A, n_components):
    model_svd = TruncatedSVD(n_components, random_state=42)
    model_svd.fit(A)
    A_approx = model_svd.inverse_transform(model_svd.transform(A))
    return A_approx
```

Để so sánh hai mô hình NMF và SVD ta tiến hành chia ma trận dữ liệu $\mathbf{A}$ thành tập huấn luyện ($\mathbf{A}_{train}$) và tập kiểm tra ($\mathbf{A}_{test}$) với tỷ lệ của tập $\mathbf{A}_{\mathbf{train}}$ lần lượt là $90\%, 80\%, 70\%$ và 67%. Tính giá trị RMSE cho từng tỷ lệ đó của cả hai mô hình để thu được kết quả như Hình 3.6.



Hình 3.6: Giá trị RMSE của mô hình NMF và SVD

Nhận xét 3.5. *(Hình 3.6)*

- *NMF có RMSE tương đối cao ở các giá trị r nhỏ (1-10) và gần như tương đương với SVD. Điều này cho thấy cả hai mô hình đều gặp khó khăn trong việc khôi phục dữ liệu với số lượng thành phần ít.*

- *Từ $r = 11$ trở đi, RMSE của SVD bắt đầu tăng nhanh hơn so với NMF, cho thấy NMF có xu hướng duy trì độ chính xác tốt hơn khi số lượng thành phần tăng.*
- *Nhìn chung, cả hai mô hình NMF và SVD đều cho giá trị RMSE nhỏ nhất trong khoảng từ 10 đến 20. Điều đó cho thấy rằng cả hai đều có khả năng hoạt động hiệu quả trong khoảng này.*
- *Ngoài ra, NMF được thiết kế đặc biệt để xử lý dữ liệu không âm, giúp mô hình dễ dàng giải thích hơn. Khi dữ liệu có nhiều giá trị không âm, NMF có thể phát hiện các cấu trúc tiềm ẩn tốt hơn, từ đó cải thiện độ chính xác và khả năng giải thích của mô hình.*

Tóm lại, từ những phân tích trên cho thấy NMF có lợi thế rõ ràng hơn SVD trong việc duy trì độ chính xác khi số lượng thành phần r tăng, đặc biệt khi làm việc với dữ liệu không âm. Cả hai mô hình đều có thể đạt được hiệu suất tốt trong khoảng r nhất định, nhưng NMF nổi bật hơn trong khả năng giải thích và phát hiện cấu trúc dữ liệu.

Có thể xem chi tiết code chương trình gợi ý phim cho người dùng theo $user\_ID$ trên bộ dữ liệu MovieLens 100K tại địa chỉ sau: https://github.com/NguyenThiBichNgoc-20185388/NMF.

Kết luận

Qua quá trình nghiên cứu và ứng dụng phương pháp nhân tử hóa ma trận không âm (NMF) để phát triển hệ thống gợi ý phim cho người dùng, em đã hiểu và tiếp thu được một số kiến thức như sau:

- Phương pháp NMF được giới thiệu như một công cụ mạnh mẽ trong việc phân tích dữ liệu không âm. Trong đồ án này, em đã phát biểu bài toán NMF và mô tả cách thức phân rã một ma trận dữ liệu thành hai ma trận không âm để giúp cho việc nhận diện các thành phần tiềm ẩn trong dữ liệu trở nên dễ dàng hơn.
- Tìm hiểu về thuật toán Multiplicative Update Rules (MUR) do Lee và Seung phát triển, một trong những thuật toán phổ biến nhất để giải bài toán NMF. Các bước của thuật toán đã được mô tả rõ ràng, từ khởi tạo ma trận đến việc lặp lại quá trình cập nhật cho đến khi đạt được điều kiện dừng và chứng minh được sự hội tụ của thuật toán này.
- Trình bày chi tiết cách thức triển khai NMF trong việc gợi ý phim cho người dùng dựa trên bộ dữ liệu MovieLens 100K. Em đã xây dựng ma trận đánh giá và áp dụng mô hình NMF để dự đoán điểm đánh giá cho các phim mà người dùng chưa xem, từ đó đưa ra danh sách các bộ phim gợi ý phù hợp.
- Kết quả phân tích cho thấy một số người dùng nhận được danh sách phim gợi ý có điểm dự đoán thấp, cho thấy rằng chúng có thể không phù hợp với sở thích của họ. Điều này có thể là do bộ phim gợi ý không tương thích với thể loại yêu thích của họ hoặc số lượng bộ phim đã đánh giá quá ít nên dẫn đến việc chưa phán ánh rõ thể loại mà họ yêu thích.
- Mặc dù hệ thống gợi ý đã hoạt động tốt với dữ liệu hiện tại, nhưng vẫn còn một số hạn chế, chẳng hạn như việc chỉ dựa vào các thể loại phim có sẵn trong cơ sở dữ liệu mà không xem xét sâu hơn về nội dung phim. Hướng phát triển trong tương lai có thể bao gồm việc cải thiện thuật toán để tăng độ chính xác của các dự đoán, cũng như tích hợp thêm thông tin về nội dung phim nhằm tạo ra trải nghiệm gợi ý tốt hơn cho người dùng.
- Có thể chỉ ra rằng NMF có lợi thế rõ ràng hơn SVD trong việc duy trì độ chính xác khi số lượng thành phần r tăng, đặc biệt khi làm việc với dữ

liệu không âm. Cả hai mô hình đều có thể đạt được hiệu suất tốt trong khoảng r nhất định, nhưng NMF nổi bật hơn trong khả năng giải thích và phát hiện cấu trúc dữ liệu, mở ra nhiều cơ hội cho các ứng dụng và nghiên cứu trong tương lai.

- Ngoài việc gợi ý phim, NMF còn có nhiều ứng dụng tiềm năng khác trong các lĩnh vực như:
  - **Phân tích hình ảnh**: NMF có thể được sử dụng để nhận diện khuôn mặt hoặc phân nhóm hình ảnh dựa trên các đặc trưng tiềm ẩn, giúp cải thiện độ chính xác trong các hệ thống nhận diện thị giác máy tính..
  - **Xử lý tín hiệu**: trong âm thanh và video, NMF giúp tách tín hiệu để cải thiện chất lượng phát lại, ứng dụng trong việc loại bỏ tiếng ồn hoặc tách các nguồn âm thanh khác nhau.
  - **Sinh học**: NMF được ứng dụng trong phân tích gene và protein, giúp nhận diện các mẫu biểu hiện gen trong dữ liệu sinh học phức tạp, từ đó hỗ trợ trong nghiên cứu bệnh lý và phát triển thuốc.
  - **Tài chính**: NMF có thể hỗ trợ trong việc phân tích dữ liệu tài chính, nhận diện các yếu tố ảnh hưởng đến giá cả và dự đoán xu hướng thị trường, giúp các nhà đầu tư đưa ra quyết định thông minh hơn.
  - **Khám phá dữ liệu**: NMF có thể được sử dụng để khai thác và phân tích các tập dữ liệu lớn, giúp nhận diện các mẫu và xu hướng tiềm ẩn, từ đó hỗ trợ ra quyết định trong các lĩnh vực khác nhau như marketing và quản lý chuỗi cung ứng.
  - **Phân tích văn bản**: NMF có thể được ứng dụng để phân nhóm văn bản hoặc trích xuất chủ đề từ các tập hợp tài liệu lớn, giúp cải thiện khả năng tìm kiếm và phân tích nội dung.

Tóm lại, nghiên cứu này đã chứng minh tính khả thi của phương pháp NMF trong việc xây dựng hệ thống gợi ý phim, đồng thời mở ra những hướng đi mới cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực phân tích dữ liệu và học máy. Phương pháp này không chỉ giới hạn trong lĩnh vực phim ảnh mà còn có thể được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ khoa học đến công nghiệp, góp phần vào việc nâng cao hiệu quả phân tích và ra quyết định.

Tài liệu tham khảo

[1] Paul Pauca V, Shahnaz F, Berry Michael W, Plemmons Robert J (2004) **Text mining using non-negative matrix factorization**, *Proceedings of the Fourth SIAM International Conference on Data Mining.*

[2] Kim PM, Tidor B (2003) **Subsystem identification through dimensionality reduction of large-scale gene expression data**, *Genome Res 13: 1706 - 1718.*

[3] Kim H, Park H (2007) **Sparse non-negative matrix factorizations via alternating non-negativity constrained least squares for microarray data analysis**, *Bioinformatics 23: 1495 - 1502.*

[4] Richardson WH (1972) **Bayesian-based iterative method of image restoration**, *J Opt Soc Am 62: 55 - 59.*

[5] Rao N, Shepherd SJ (2004) **Extracting characteristic patterns from genomewide expression data by nonnegative matrix factorization**, *In Computational Systems Bioinformatics Conference.*

[6] Spratling MW (2006) **Learning image components for object recognition**, *J Mach Learn Res 7: 793 - 815.*

[7] Paatero P, Tapper U (1994) **Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values**, *Environmetrics 5: 111 - 126.*

[8] Lee DD, Seung HS (1999) **Learning the parts of objects by non-negative matrix factorization**, *Nature 401: 788 - 791.*

Lee DD, Seung HS (2001) **Algorithms for non-negative matrix factorization**, *Proceedings of Neural Information Processing Systems, vol 13, pp 556 - 562.*

[9] Chu M, Diele F, Plemmons R, Ragni S (2004) **Optimality, computation and interpretations of nonnegative matrix factorizations**, *SIAM J Matrix Anal 4 - 8030.*

[10] Ngoc-Diep Ho (2008) **Non negative matrix factorization algorithms and applications**, *PhD thesis, Université catholique de Louvain.*

[11] Vũ Hữu Tiệp (2019) **Machine Learning cơ bản**, *Nhà xuất bản Đại học Quốc gia Hà Nội.*

[12] Hyunsoo Kim, Haesun Park (2007) **Sparse NMF via Alternating Non-negativity Constrained Least Squares**, *College of Computing Georgia Institute of Technology Atlanta, GA 30332, USA, Nonnegative Matrix Factorization Workshop.*

[13] Sheng Zhang, Weihong Wang, James Ford, Fillia Makedon (2006) **Learning from incomplete ratings using non-negative matrix factorization**, *Proceedings of the 2006 SIAM International Conference on Data Mining, pages 549 - 553.*

[14] Mahesh Mali, Dhirendra Mishra, M. Vijayalaxmi (2022) **MF-RISE: Benchmarking for Multifaceted Recommender System Engine**, *Journal: Research Square - 4/1/2023.*

[15] Shalin S Shah (2021) **A Survey of Latent Factor Models for Recommender Systems and Personalization**, *Data Mining Course Project, Johns Hopkins University.*

[16] Bhagyashree Shelke, Suraj Mandhane, Rishav Agrawal, Prasanna Jain, Umakant Mandawkar (2021) **Movie Recommendation System Using SVD**, *Journal of Engineering Sciences, Vol 12, Issue 06.*