MA TRẬN KHÔNG ÂM

Tran Van Long

AI Academy Vietnam

20/7/2024

Nội dung

Ma trận không âm [1]

Ma trận không âm

Ma trận $A = (a_{ij})_{m\times n}$ gọi là ma trận không âm nếu $a_{ij} \geq 0$.

Ma trận dương

Ma trận $A = (a_{ij})_{m\times n}$ gọi là ma trận dương nếu $a_{ij} > 0$.

Ký hiệu

$$|x| = (|x_1|, |x_2|, \dots, |x_n|)$$

[1] (Tài liệu 1)

Ma trận dương

Tính chất

A là ma trận vuông cấp n và dương. Khi đó,

- A có giá trị riêng dương lớn nhất ρ (gọi là nghiệm Perron),
- A có véc-tơ riêng dương (véc-tơ riêng Perron) ứng với giá trị riêng dương lớn nhất,
- Nghiệm Perron thỏa mãn

$$\rho = \max\{f(x) = \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} : \quad x \geq 0, x \neq 0\}$$

VINBIGDATA VINGROUP AI Academy Vietnam

Ví dụ

Cho ma trận $A = \begin{bmatrix} 7 & 2 & 3 \\ 1 & 8 & 3 \\ 1 & 2 & 9 \end{bmatrix}$

Các giá trị riêng của A là $12, 6, 6$ nên nghiệm Perron là $\rho = 12$ với véc-tơ Perron $p = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

Ma trận chuyển vị A^T cũng là ma trận dương có nghiệm Perron $\rho = 12$ và véc-tơ Perron $q = (1, 2, 3)$.

Ma trận không âm

Ma trận A không âm gọi là ma trận **rút gọn được** (reducible) nếu tồn tại ma trận giao hoán P sao cho

$$P^T A P = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix}$$

với X, Z là các ma trận vuông.
Trái lại ta nói ma trận A **bất khả quy** (irreducible).

Tính chất

Ma trận A không âm là bất khả quy thì $(I + A)^{n-1} > 0$.

Ma trận không âm

Cho A vuông cấp n, không âm, bất khả quy. Khi đó

Tính chất

- A có giá trị riêng dương lớn nhất ρ (bội 1).
- A có véc-tơ riêng dương ứng với ρ.
- Công thức Collatz-Wielandt:

$$\rho = \max\{f(x) = \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} : \quad x \geq 0, x \neq 0\}$$

VINBIGDATA VINGROUP AI Academy Vietnam

Phân tích ma trận không âm

Phân tích ma trận không âm (non-negative matrix factorization-NMF): Biểu diễn ma trận A cỡ $m \times n$ không âm thành tích hai ma trân WH với $W \geq 0, H \geq 0$ cỡ $m \times r$, $n \times r$.

NMF cực tiểu hóa độ lệch giữa X và WH.

NMF áp dụng các lĩnh vực: topic modeling, text mining, audio source separation, microarray data analysis.

Bài toán NMF là NP-hard, tối ưu không lồi dẫn đến bài toán hội tụ đến điểm dừng (không chắc là tối ưu toàn cục).

Bài toán NMF lồi theo từng biến W, và H.

Bài toán NMF không có nghiệm duy nhất.

Nếu ma trận X thỏa mãn tích chất r-tách (r-separable): Các cột của ma trận X là tổ hợp không âm của r cột của X.

Phân tích ma trận không âm: lời giải chính xác

Giả sử ta có phân tích

$$X = WH,$$

các cột thứ j của ma trận X được viết dưới dạng

$$X_j = WH_j = \sum_{i=1}^{r} W_i H_{ij}$$

Ta ký hiệu nón sinh vởi các cột của ma trận W là

$$Cone(W) = \{x = \sum_{i=1}^{n} \alpha_i W_i : \quad \alpha_i \geq 0\}$$

Bài toán NMF có lời giải chính xác iff

$$Cone(X) \subset Cone(W).$$

Phân tích ma trận không âm: lời giải chính xác

Giả sử ma trận X thỏa mãn tính chất r-tách.
Ta xét tập nón lồi sinh bởi các cột của ma trận X là $Cone(X)$.
Đường sinh cực biên của một nón lồi $Cone(X)$ là tập $R_x = \{\alpha x : \ \alpha \geq 0\}$ với $x \in Cone(X)$ không thể biểu diễn là một tổ hợp lồi của hai điểm phân biệt trong $Cone(X)$ khác với x.
Ta xác định r đường sinh cực biên của nón lồi $Cone(X)$ sinh bởi các cột $W_1, W_2, \ldots, W_r$. Xét cột thứ j của ma trận X có biểu diễn tổ hợp không âm dạng

$$X_j = \sum_{i=1}^{r} H_{ij} W_i = WH_j \Rightarrow X = WH.$$

VINBIGDATA VINGROUP AI Academy Vietnam

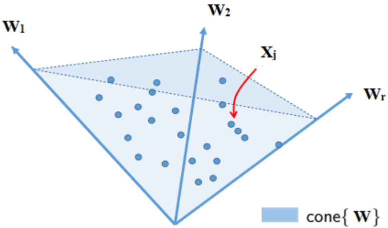
Phân tích ma trận không âm

Cho ma trận X cỡ $m \times n$ không âm và $r < \min\{m, n\}$. Tìm hai ma trận không âm $W \in \mathbb{R}_+^{m\times r}$ và $H \in \mathbb{R}_+^{r\times n}$ sao cho có xấp xỉ

$$X \approx WH.$$

Các cột của ma trận X xấp xỉ bởi tổ hợp tuyến tính các cột của W với

$$X_j \approx H_{1j}W_1 + H_{2j}W_2 + \cdots + H_{rj}W_r$$



Phân tích ma trận không âm

Cho ma trận X tìm $W \geq 0, H \geq 0$ sao cho

$$\min_{W,H} f(X - WH)$$

với $f(\cdot)$ là hàm tổn thất.
Chuẩn của ma trận

- $||A||_2 = \sigma_{\max}(A)$
- $||A||_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{trace(A^T A)} = \sqrt{\sum_i \sigma_i^2(A)}$ (chuẩn Frobenius)
- $||A||_* = trace(\sqrt{A^T A}) = \sum_i \sigma_i(A)$ (chuẩn nuclear)

Phương pháp tối ưu theo tọa độ [2]

Bài toán

$$\min f(x_1, x_2, \ldots, x_n)$$

Phương pháp tối ưu theo tọa độ

- chọn ngẫu nhiên $x^{(0)}$
- Repeat
  - for i from 1 to n
  - $x_i^{(k+1)} = \arg\min_y f(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, \mathbf{y}, x_{i+1}^{(k)}, \ldots, x_n^{(k)})$
- stop condition

[2] (Tài liệu 2)

Phương pháp tối ưu theo tọa độ

Bài toán

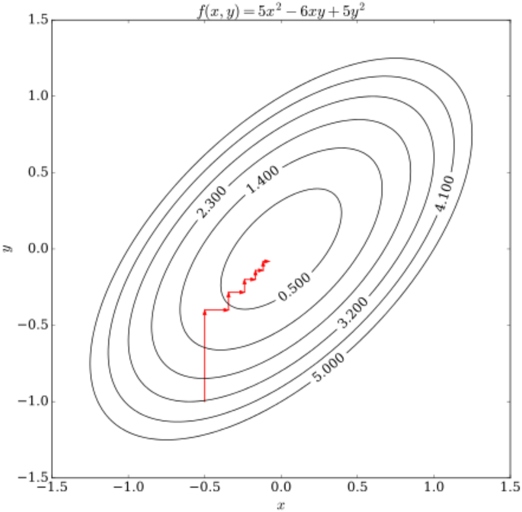
$$\min f(x) = \min ||Ax - b||^2/2$$

$$\frac{\partial f}{\partial x_i} = A_i^T(Ax - b) = A_i^T(A_i x_i + A_{-i}x_{-i} - b) = 0 \Rightarrow x_i = \frac{A_i^T(A_{-i}x_{-i} - b)}{A_i^T A_i}$$

Công thức lặp

$$x_i^{(k+1)} = \frac{A_i^T(A_{1:i-1}x_{1:i-1}^{(k+1)} + A_{i+1:n}x_{i+1:n}^{(k)} - b)}{A_i^T A_i}$$

Phương pháp tối ưu theo tọa độ



Phân tích ma trận không âm [3]

Xét bài toán

$$F(W,H) = ||X - WH||_F^2 \tag{1}$$

Điểm dừng (U, V) là cực tiểu địa phương hàm $F(W, H)$ thỏa mãn các tính chất sau:

Tính chất

- $\left\langle UV, UV - X \right\rangle = 0$
- $||X - UV||^2 = ||X||_F^2 - ||UV||_F^2$

Nếu X_k là ma trận xấp xỉ tốt nhất hạng k của ma trận X thì ma trận không âm của X_k, ký hiệu là $[X_k]_+$ thỏa mãn

$$||X - [X_k]_+||_F \leq ||X - X_k||_F.$$

[3] (Tài liệu 2)

Thuật toán Lee-Seung

Ta viết

$$F(W,H) = ||X - WH||^2 = \sum_{i=1}^{n} ||X_i - WH_i||^2$$

Ta cực tiểu hóa các cột của H tách rời nhau. Dẫn đến bài toán tìm cực tiểu

$$\min_{h\geq 0} F(h) \text{ với } F(h) = ||x - Wh||^2$$

Giả sử nghiệm xấp xỉ là $\bar{h}$, xét hàm trội của $F(h)$ tại $\bar{h}$

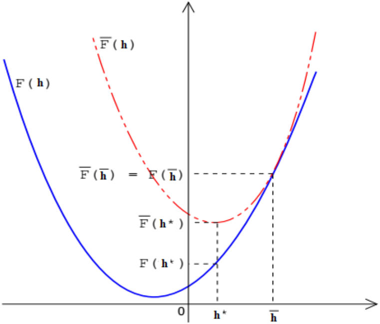
$$\bar{F}(h) = ||x - Wh||^2 + (h - \bar{h})^T H_{\bar{h}}(h - \bar{h}),$$

với $H_{\bar{h}} = D_a - W^T W$ với $a = \dfrac{[W^T W \bar{h}]}{[\bar{h}]}$ là phép chia Hadamard.

Thuật toán Lee-Seung

Giá trị cực tiểu của hàm $\bar{F}(h)$ tại

$$h^* = \bar{h} \circ \frac{[W^T x]}{[W^T W \bar{h}]}$$



Thuật toán Lee-Seung

- Cố định W, cập nhật ma trận H.
- Cố định H, cập nhật ma trận W.

Thuật toán Lee-Seung

- Initialize $W^0, H^0, k = 0$
- REPEAT
  - $W^{k+1} = W^k \circ \dfrac{[X(H^k)^T]}{[W^k H^k (H^k)^T]}$
  - $H^{k+1} = H^k \circ \dfrac{[(W^{k+1})^T X]}{[(W^k)^T W^{k+1} H^k]}$
  - $k = k + 1$
- STOPPING CONDITION

Thuật toán ALS (Alternating Least Squares)

Ta giải hai bài toán

- $W = \arg\min_{W \geq 0} ||X - WH||_F^2$
- $H = \arg\min_{H \geq 0} ||X - WH||_F^2$

Ta xét bài toán tìm

$$\min ||Ax - b||^2 \text{ với} x \geq 0.$$

Có nghiệm tối ưu là $x = [A^\dagger b]_+$
Từ đó ta có công thức cập nhật W, H là

- $W \leftarrow [XH^\dagger]_+$
- $H \leftarrow [W^\dagger X]_+$

Ví dụ 1

Cho ma trận không âm X.

$$
\begin{array}{cccccccccc}
\text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} & \textit{Articles} \\
\end{array}
$$

$$
\begin{array}{c}
\left(\begin{array}{ccccccccc}
6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\
1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\
0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\
0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\
1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1
\end{array}\right)
\end{array}
\begin{array}{c}
\text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f}
\end{array}
$$

Ví dụ 1

Phân tích $X = WH$

$$W = \begin{pmatrix} 0.278 & 0. \\ 0. & 0.34 \\ 0.289 & 0. \\ 0.267 & 0. \\ 0.166 & 0.306 \\ 0. & 0.354 \end{pmatrix} \text{Documents x Topics}$$

$$H = \begin{pmatrix} 14. & 14. & 2.314 & 1.279 & 0. & 13.38 & 19. & 12.776 & 16.76 \\ 2. & 0. & 22.686 & 16.721 & 22. & 5.62 & 0. & 3.224 & 1.24 \end{pmatrix}$$

Topics x Terms

Ví dụ 1

$$WH = \begin{bmatrix} 3.892 & 3.892 & 0.643 & 0.356 & 0. & 3.72 & 5.282 & 3.552 & 4.659 \\ 0.68 & 0. & 7.713 & 5.685 & 7.48 & 1.911 & 0. & 1.096 & 0.422 \\ 4.046 & 4.046 & 0.669 & 0.37 & 0. & 3.867 & 5.491 & 3.692 & 4.844 \\ 3.738 & 3.738 & 0.618 & 0.341 & 0. & 3.572 & 5.073 & 3.411 & 4.475 \\ 2.936 & 2.324 & 7.326 & 5.329 & 6.732 & 3.941 & 3.154 & 3.107 & 3.162 \\ 0.708 & 0. & 8.031 & 5.919 & 7.788 & 1.989 & 0. & 1.141 & 0.439 \end{bmatrix}$$

$$E = \begin{bmatrix} 2.10 & -2.89 & 0.35 & -0.35 & 0. & -2.72 & 3.71 & -3.55 & 3.34 \\ 0.32 & 0. & 1.28 & -0.68 & 0.52 & -0.91 & 0. & -0.09 & -0.42 \\ 3.95 & -3.04 & -0.66 & 0.63 & 0. & -3.86 & 3.50 & -2.69 & 2.15 \\ -3.73 & 3.26 & 0.38 & -0.34 & 0. & 5.42 & -4.07 & 3.58 & -4.47 \\ -2.93 & 2.67 & -1.32 & 1.67 & -1.73 & 2.05 & -3.15 & 3.89 & -1.16 \\ 0.29 & 0. & -0.03 & -0.91 & 1.21 & 0.011 & 0. & -1.14 & 0.56 \end{bmatrix}$$

Ví dụ 1

singer	GDP	senate	election	vote	stock	bass	market	band
14.	14.	2.314	1.279	0.	13.38	19.	12.776	16.76
2.	0.	22.686	16.721	22.	5.62	0.	3.224	1.24
1	1	2	2	2	1	1	1	1

tocic 1: singer, GDP, stock, bass, market, band
topic 2: senate, election, vote

Ví dụ 2: Dữ liệu 20 Newsgroups

Thực hiện phân tích ma trận X cỡ 2000×5136 dạng document-term với $k = 20$. Giá trị x_{ij} biểu diễn số lần xuất hiện term (word) j trong văn bản (document) i. (Chú ý rằng ma trận rất thưa)
File dữ liệu: news.csv

Thực hành 1

Cho ma trận không âm X.

$$\begin{array}{cccccccccc}
\text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} & \\
\end{array}$$

$$\begin{array}{c}
\\
\end{array}
\left(\begin{array}{ccccccccc}
6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\
1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\
0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\
0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\
1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1
\end{array}\right)
\begin{array}{c}
Articles \\ \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f}
\end{array}$$

Thực hiện thuật toán Lee-Seung với $k = 3$ tìm W, H để $X \approx WH$.

Thực hành 2

Cho ma trận không âm X.

$$\begin{array}{cc}
\begin{array}{ccccccccc} \text{singer} & \text{GDP} & \text{senate} & \text{election} & \text{vote} & \text{stock} & \text{bass} & \text{market} & \text{band} \end{array} & \mathit{Articles} \\
\begin{pmatrix}
6 & 1 & 1 & 0 & 0 & 1 & 9 & 0 & 8 \\
1 & 0 & 9 & 5 & 8 & 1 & 0 & 1 & 0 \\
8 & 1 & 0 & 1 & 0 & 0 & 9 & 1 & 7 \\
0 & 7 & 1 & 0 & 0 & 9 & 1 & 7 & 0 \\
0 & 5 & 6 & 7 & 5 & 6 & 0 & 7 & 2 \\
1 & 0 & 8 & 5 & 9 & 2 & 0 & 0 & 1
\end{pmatrix} & \begin{array}{c} \text{a} \\ \text{b} \\ \text{c} \\ \text{d} \\ \text{e} \\ \text{f} \end{array}
\end{array}$$

Thực hiện thuật toán ALS với $k = 3$ tìm W, H để $X \approx WH$.

Bài tập 1

The Olivetti faces dataset
Classes: 40
Samples total: 400
Dimensionality: 4096
Features: real, between 0 and 1
Thực hiện thuật toán Lee-Seung với dữ liệu faces.

Bài tập 2

The Olivetti faces dataset
Classes: 40
Samples total: 400
Dimensionality: 4096
Features: real, between 0 and 1
Thực hiện thuật toán ALS với dữ liệu faces.

Bài tập 3

The Olivetti faces dataset
Classes: 40
Samples total: 400
Dimensionality: 4096
Features: real, between 0 and 1
Thực hiện thư viện sklearn.decomposition.NMF với dữ liệu faces.

Tài liệu tham khảo

1. Carl D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, 2110.
2. Charu C. Aggarwal; Linear Algebra and Optimization for Machine Learning, Springer, 2020.
3. David C. Lay, Steven R. Lay, Judi J. McDonald; Linear Algebra and Its Applications, Fifth edition, Pearson, 2016
4. Gilbert Strang; Linear Algebra and Learning from Data, Wellesley- Cambridge Press, 2019.
5. Stephen Boyd, Lieven Vandenberghe; Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares, Cambridge University Press, 2018.
6. Tom Lyche; Numerical Linear Algebra and Matrix Factorizations, Springer, 2020.
7. Hyun-Seok Son; Linear Algebra Coding with Python, 2020.