

Bài tập 2 - Tìm compact SVD của một ma trận

Viết chương trình cho phép nhập vào một ma trận, sau đó tính compact SVD của ma trận đó. Sử dụng ma trận dưới đây để kiểm tra.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$

Thuật toán:

- ▶ Đầu vào: Ma trận A .
- ▶ Đầu ra: Ma trận U_r , V_r^T và Σ_r

Các bước làm bài:

1. Tìm SVD của ma trận A như trên (dùng scipy).
2. Tìm số giá trị λ khác 0 của ma trận A , đặt là r .
3. Giữ lại r hàng và r cột đầu tiên của ma trận Σ để tạo thành Σ_r .
4. Giữ lại r cột đầu tiên của ma trận U để tạo ra U_r .
5. Giữ lại r hàng đầu tiên của ma trận V^T để tạo ra V_r^T .
6. Xuất kết quả.

Lưu ý: Khi nhân $U_r \Sigma_r V_r^T$ với nhau, ta vẫn được ma trận A như ban đầu. Phép giảm chiều compact SVD không làm thay đổi ma trận A . Vì vậy, sau khi tìm được compact SVD của A , hãy thử nhân lại $U_r \Sigma_r V_r^T$ để kiểm tra kết quả.

Truncated SVD

Trong ma trận Σ , các giá trị trên đường chéo chính là không âm và giảm dần

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0.$$

Thông thường, chỉ có một số các trị riêng λ_i mang giá trị lớn, các giá trị còn lại thường nhỏ và gần 0. Khi đó, với phương pháp truncated SVD, ta có thể xấp xỉ ma trận A dưới dạng

$$A \approx A_k = U_k \Sigma_k V_k^T \text{ với } k < r \quad (5)$$

trong đó

- ▶ U_k và V_k lần lượt là ma trận được tạo bởi k cột đầu tiên của U và V .
- ▶ Σ_k là ma trận được tạo bởi k hàng đầu tiên và k cột đầu tiên của Σ .

Việc xóa bớt những giá trị λ nhỏ và gần bằng 0 khiến cho việc lưu trữ thông tin của ma trận A không thể y như ban đầu. Khi chúng ta nhân ba ma trận U_k, Σ_k, V_k^T lại với nhau, ta chỉ nhận được một ma trận gần đúng với ma trận A .

\implies Số giá trị λ được giữ lại ảnh hưởng đến lượng thông tin được giữ lại của ma trận A sau khi thực hiện phép giảm chiều truncated SVD.

Đây chính là điểm khác biệt giữa compact SVD và truncated SVD.

k và lượng thông tin trong ma trận A

Công thức tính phần trăm lượng thông tin giữ lại sau khi giữ lại k giá trị λ :

$$I = \left(1 - \frac{\|A - A_k\|_F^2}{\|A\|_F^2}\right) * 100 = \left(\frac{\sum_{i=1}^k \lambda_i^2}{\sum_{j=1}^r \lambda_j^2}\right) * 100 \quad (6)$$

Công thức tính số k cần giữ lại nếu muốn giữ lại ít nhất $I\%$ lượng thông tin trong A :

Chọn k nhỏ nhất sao cho

$$\frac{\sum_{i=1}^k \lambda_i^2}{\sum_{j=1}^r \lambda_j^2} \geq I\% \quad (7)$$