KHAI THÁC DỮ LIỆU

Nguyễn Xuân Việt Đức - 22280012

Bài tập lý thuyết - Lần 1

1 So sánh giữa PCA, SVD và LSA trong giảm chiều dữ liệu

1.1 Phân tích thành phần chính (Principal Component Analysis - PCA)

PCA là phương pháp giảm chiều dữ liệu không giám sát phổ biến nhất, với các đặc điểm sau:

- **Nguyên lý**: PCA tìm các hướng (thành phần chính) có phương sai lớn nhất trong không gian dữ liệu nhiều chiều và chiếu dữ liệu lên các hướng này.
- **Mục tiêu**: Tối đa hóa phương sai được giữ lại sau khi giảm chiều.
- **Cách thức hoạt động**:
  1. Chuẩn hóa dữ liệu
  2. Tính ma trận hiệp phương sai
  3. Tính các vector riêng và giá trị riêng
  4. Sắp xếp các vector riêng theo giá trị riêng giảm dần
  5. Chọn k vector riêng đầu tiên để tạo không gian con
  6. Chiếu dữ liệu lên không gian con này
- **Công thức toán học**: Nếu X là ma trận dữ liệu đã chuẩn hóa, và W là ma trận các vector riêng đã chọn, thì dữ liệu sau khi giảm chiều là: $Z = XW$
- **Ưu điểm**: Đơn giản, hiệu quả tính toán, giữ lại được nhiều thông tin với ít thành phần.
- **Nhược điểm**: Chỉ nắm bắt mối quan hệ tuyến tính, nhạy cảm với tỷ lệ của các biến.

1.2 Phân tích giá trị suy biến (Singular Value Decomposition - SVD)

SVD là một kỹ thuật đại số tuyến tính có thể được áp dụng cho giảm chiều dữ liệu:

- **Nguyên lý**: SVD phân tách một ma trận thành tích của ba ma trận: $A = U\Sigma V^T$
  - U: Ma trận chứa các vector riêng trái
  - Σ: Ma trận đường chéo chứa các giá trị suy biến (singular values)
  - V^T: Ma trận chuyển vị của ma trận chứa các vector riêng phải
- **Mục tiêu**: Phân rã ma trận thành các thành phần cơ bản và giữ lại các thành phần quan trọng nhất.
- **Cách thức hoạt động**:
  1. Phân tách ma trận dữ liệu A thành $U\Sigma V^T$
  2. Sắp xếp các giá trị suy biến theo thứ tự giảm dần
  3. Chọn k giá trị suy biến lớn nhất và vector tương ứng
  4. Tạo ma trận xấp xỉ $A_k = U_k\Sigma_k V_k^T$
- **Mối liên hệ với PCA**: Thực tế, PCA có thể được thực hiện thông qua SVD. Nếu áp dụng SVD cho ma trận dữ liệu đã chuẩn hóa, các thành phần chính sẽ tương ứng với các cột của ma trận V.
- **Ưu điểm**: Ổn định về số học, có thể áp dụng cho ma trận không vuông, hiệu quả hơn PCA khi ma trận thưa thớt.
- **Nhược điểm**: Chi phí tính toán cao hơn với ma trận lớn.

1.3 Phân tích ngữ nghĩa tiềm ẩn (Latent Semantic Analysis - LSA)

LSA là một kỹ thuật xử lý ngôn ngữ tự nhiên, được áp dụng chủ yếu cho phân tích văn bản:

- **Nguyên lý**: LSA là ứng dụng của SVD trong xử lý ngôn ngữ tự nhiên, áp dụng trên ma trận term-document.
- **Mục tiêu**: Khám phá các cấu trúc ngữ nghĩa tiềm ẩn trong văn bản, vượt qua những hạn chế của việc so khớp từ vựng đơn thuần.
- **Cách thức hoạt động**:
  1. Tạo ma trận term-document A, trong đó A_{ij} thể hiện tần suất xuất hiện của từ i trong văn bản j

2. Áp dụng trọng số TF-IDF (Term Frequency-Inverse Document Frequency)
3. Thực hiện SVD: $A = U\Sigma V^T$
4. Chọn k giá trị suy biến lớn nhất
5. Không gian ngữ nghĩa được biểu diễn bởi ma trận U_k, và văn bản được biểu diễn trong không gian này bởi $\Sigma_k V_k^T$

- **Ưu điểm**: Có khả năng nhận biết đồng nghĩa, bao hàm ngữ nghĩa, giải quyết một phần vấn đề đồng âm khác nghĩa.
- **Nhược điểm**: Không nắm bắt được trật tự từ trong câu, mất thông tin ngữ cảnh, hiệu quả giảm với kho văn bản lớn và đa dạng.

1.4 Bảng so sánh tổng quan

Tiêu chí	**PCA**	**SVD**	**LSA**
Nguyên lý cốt lõi	Tìm hướng có phương sai lớn nhất	Phân tách ma trận thành tích của ba ma trận	Áp dụng SVD cho ma trận term-document
Đầu vào	Ma trận dữ liệu đã chuẩn hóa	Bất kỳ ma trận nào	Ma trận term-document (thường với TF-IDF)
Lĩnh vực ứng dụng	Phân tích dữ liệu chung, phát hiện mẫu	Nén ảnh, lọc nhiễu, khuyến nghị	Xử lý ngôn ngữ tự nhiên, tìm kiếm thông tin
Độ phức tạp tính toán	$O(min\{mn^2, m^2n\})$ với $m \times n$ là kích thước ma trận	$O(mn^2)$ với $m \geq n$	Tương tự như SVD
Xử lý dữ liệu thưa	Không hiệu quả	Hiệu quả hơn PCA	Được thiết kế đặc biệt cho dữ liệu thưa
Cấu trúc toán học	Dựa trên ma trận hiệp phương sai	Phân tích trực tiếp ma trận dữ liệu	Phân tích ma trận tần suất từ-văn bản

1.5 Kết luận

Trong khi cả ba phương pháp đều liên quan đến việc giảm chiều dữ liệu, mỗi phương pháp có những đặc điểm và ứng dụng riêng:

- **PCA** là một kỹ thuật đơn giản và hiệu quả cho dữ liệu nhiều chiều, tập trung vào việc tối đa hóa phương sai.
- **SVD** là một công cụ đại số tuyến tính mạnh mẽ hơn, có thể áp dụng cho nhiều loại ma trận, và là nền tảng toán học cho PCA.

- **LSA** là một ứng dụng cụ thể của SVD trong lĩnh vực xử lý ngôn ngữ tự nhiên, giúp khám phá mối quan hệ ngữ nghĩa tiềm ẩn giữa các từ và văn bản.

Lựa chọn phương pháp nào phụ thuộc vào bản chất của dữ liệu và mục tiêu cụ thể của bài toán giảm chiều.