Bài tập Lý thuyết môn Khai phá Dữ liệu

Họ và tên: Huỳnh Nguyễn Thế Dân

MSSV: 21110256

Lớp: 21TTH1

Questions

Các em phân biệt sự khác nhau và điểm mạnh/điểm yếu của 03 phương pháp: LSA, PCA, và SVD trong việc giảm số chiều dữ liệu.

Các phương pháp Latent Semantic Analysis (LSA), Principal Component Analysis (PCA), và Singular Value Decomposition (SVD) đều được sử dụng để giảm số chiều dữ liệu trong các bài toán xử lý ngôn ngữ tự nhiên và khai phá dữ liệu. Dưới đây là sự khác nhau và điểm mạnh/điểm yếu của mỗi phương pháp:

Latent Semantic Analysis (LSA):

- **Sự khác nhau**:
  - **LSA** là một phương pháp phân tích ma trận được sử dụng để xác định các mối quan hệ ngữ nghĩa giữa các từ và văn bản trong một tập dữ liệu lớn. Phương pháp này dựa trên giả định rằng các từ có liên quan đến nhau thường xuất hiện cùng nhau trong các văn bản và ngược lại, các văn bản có chứa các từ tương tự cũng có nội dung tương tự.
- **Điểm mạnh**:
  - **Tìm kiếm semantic**: LSA cho phép tìm ra các từ có ý nghĩa tương tự trong không gian vector, ngay cả khi chúng không xuất hiện cùng nhau trong cùng một văn bản.
  - **Giảm số chiều dữ liệu**: LSA giúp giảm số chiều của ma trận term-document một cách hiệu quả, giúp giảm bớt vấn đề về chiều dữ liệu và cải thiện hiệu suất tính toán.
  - **Xử lý dữ liệu thưa**: LSA xử lý được dữ liệu thưa (sparse data) bằng cách chuyển đổi nó thành một không gian vector thưa.
- **Điểm yếu**:

    - **Không biểu diễn được mối quan hệ cụ thể giữa các từ**: Mặc dù LSA có thể xác định các từ có ý nghĩa tương tự, nhưng nó không cung cấp thông tin cụ thể về mối quan hệ semantic giữa các từ.
    - **Phụ thuộc vào biểu diễn ma trận term-document**: LSA phụ thuộc vào biểu diễn ma trận term-document, điều này có nghĩa là nó có thể không hiệu quả khi xử lý các loại dữ liệu khác ngoài văn bản.
    - **Khả năng xử lý dữ liệu lớn**: Trong một số trường hợp, tính toán LSA có thể trở nên phức tạp đối với các tập dữ liệu lớn.

Principal Component Analysis (PCA):

- **Sự khác nhau**:
    - **PCA** là một phương pháp thống kê được sử dụng để giảm số chiều của dữ liệu bằng cách tìm ra các thành phần chính (principal components) của dữ liệu. Các thành phần chính là các vector eigenvector ứng với các eigenvalue lớn nhất của ma trận hiệp phương sai (covariance matrix) của dữ liệu.
- **Điểm mạnh**:
    - **Giảm số chiều dữ liệu**: PCA giúp giảm số chiều của dữ liệu một cách hiệu quả bằng cách chọn ra các thành phần chính giữ lại phần lớn thông tin quan trọng của dữ liệu.
    - **Giữ lại cấu trúc quan trọng của dữ liệu**: Các thành phần chính được chọn sao cho chúng biểu diễn được phần lớn sự biến thiên trong dữ liệu, giữ lại cấu trúc quan trọng của dữ liệu gốc.
    - **Khả năng khám phá các mối quan hệ tuyến tính**: PCA là một công cụ mạnh mẽ để khám phá các mối quan hệ tuyến tính giữa các biến trong dữ liệu.
- **Điểm yếu**:
    - **Không phân biệt được các mối quan hệ phi tuyến tính**: PCA chỉ tập trung vào các mối quan hệ tuyến tính giữa các biến, do đó nó không phân biệt được các mối quan hệ phi tuyến tính trong dữ liệu.
    - **Khả năng ảnh hưởng bởi các outliers**: Các outliers có thể ảnh hưởng đến kết quả của PCA bằng cách thay đổi ma trận hiệp phương sai và do đó ảnh hưởng đến các thành phần chính được chọn.
    - **Chỉ phù hợp cho dữ liệu có phân phối Gaussian**: PCA giả định rằng dữ liệu có phân phối Gaussian, do đó nó không phù hợp cho các loại dữ liệu không tuân theo phân phối này.

Singular Value Decomposition (SVD):

- **Sự khác nhau**:
    - **SVD** là một phương pháp phân rã một ma trận thành ba ma trận con: một ma trận unitary, một ma trận đường chéo có các singular values trên đường chéo, và một ma trận unitary khác. SVD có thể được áp dụng cho các ma trận không vuông và có thể giải quyết một loạt các vấn đề trong khoa học dữ liệu, xử lý ảnh, xử lý tín hiệu, và các lĩnh vực khác.
- **Điểm mạnh**:

    - **Giảm số chiều dữ liệu**: SVD có thể giảm số chiều của dữ liệu một cách hiệu quả bằng cách chọn ra các singular values và các singular vectors quan trọng nhất.
    - **Biểu diễn dữ liệu gốc một cách hiệu quả**: SVD tạo ra một biểu diễn hiệu quả của dữ liệu gốc bằng cách sử dụng các singular vectors và singular values.
    - **Khả năng xử lý dữ liệu thưa**: SVD có khả năng xử lý dữ liệu thưa (sparse data), tức là dữ liệu mà nhiều phần tử có giá trị 0.
- **Điểm yếu**:
    - **Tính toán phức tạp**: Tính toán SVD có thể trở nên phức tạp đối với các ma trận lớn, đặc biệt là khi cần tính toán tất cả các singular values và singular vectors.
    - **Khó hiểu và khó diễn giải**: Kết quả của SVD có thể khó hiểu và khó diễn giải, đặc biệt đối với người không chuyên về toán học hoặc thống kê.
    - **Không hiệu quả với dữ liệu lớn và thưa**: Mặc dù SVD có khả năng xử lý dữ liệu thưa, nhưng đối với các tập dữ liệu lớn và thưa, tính toán có thể trở nên không hiệu quả.

Hết.