ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

Trực quan hóa dữ liệu

ĐỀ TÀI

PCA Althorigm

Lớp: 21_21 **Sinh viên thực hiện:**	
Tên	**MSSV**
Lương Thành Đạt	21120428

MỤC LỤC

1 Động lực

Giảm Chiều Dữ Liệu: Trong nhiều lĩnh vực, dữ liệu thường có số chiều cao, đặc biệt là khi làm việc với các tập dữ liệu từ cảm biến, hình ảnh, hoặc văn bản. PCA giúp giảm số chiều này, từ đó làm cho việc xử lý dữ liệu hiệu quả hơn, giảm độ phức tạp tính toán và tránh hiện tượng overfitting.

Trích Xuất Đặc Trưng: PCA có thể được sử dụng để trích xuất các đặc trưng quan trọng từ dữ liệu, giúp tăng cường khả năng hiểu và diễn giải dữ liệu. Điều này đặc biệt hữu ích khi phải làm việc với các tập dữ liệu lớn hoặc có nhiều biến.

Phân Tích Thống Kê: Trong phân tích thống kê, PCA được sử dụng để khám phá cấu trúc ẩn trong dữ liệu và xác định các biến quan trọng. Điều này có thể giúp trong việc hiểu và diễn giải các mối quan hệ phức tạp trong dữ liệu.

Trực Quan Hóa Dữ Liệu: PCA cho phép biến đổi dữ liệu vào một không gian có số chiều thấp hơn, giúp trực quan hóa dữ liệu một cách dễ dàng và hiệu quả hơn. Việc này giúp người nghiên cứu hoặc nhà phân tích dữ liệu hiểu rõ hơn về cấu trúc và mối quan hệ trong dữ liệu.

2 Thuật toán PCA

Khái niệm

Principal Component Analysis (PCA), là một phương pháp thống kê được sử dụng để giảm chiều dữ liệu trong khi vẫn giữ lại phần lớn thông tin quan trọng. PCA giúp tìm ra các hướng trong không gian dữ liệu mà biến thiên của dữ liệu là lớn nhất.

Vấn đề

Cho một tập dữ liệu có không gian đặc trưng có số chiều cao, mục tiêu là tìm ra một biểu diễn có số chiều thấp hơn nhưng vẫn giữ lại phần lớn thông tin quan trọng.

Đầu vào

- Tập dữ liệu X được biểu diễn dưới dạng ma trận $m \times n$ trong đó m là số mẫu và n là số chiều của các đặc trưng.
- Số chiều mong muốn k cho biểu diễn giảm chiều.

Đầu ra

- Các thành phần chính: Các vector riêng đại diện cho các hướng của phương sai lớn nhất trong dữ liệu.
- Dữ liệu biến đổi: Tập dữ liệu Y thu được bằng cách chiếu dữ liệu gốc lên các thành phần chính đã chọn, tạo ra một biểu diễn giảm chiều.

Ràng buộc

- Số chiều k phải nhỏ hơn n (số chiều của các đặc trưng).
- Thông thường, k được xác định dựa trên mức độ giữ lại phương sai mong muốn hoặc ràng buộc tính toán.

Mục tiêu

- Tối đa hóa phương sai được giữ lại.
- Tối thiểu hóa sự mất mát thông tin.

Các bước

Bước 1: Tính toán ma trận hiệp phương sai

Covariance Matrix: Ma trận hiệp phương sai Σ của dữ liệu X có thể được tính bằng công thức sau:

$$\Sigma = \frac{1}{n}(X - \bar{X})(X - \bar{X})^T$$

Trong đó:

- X là ma trận dữ liệu với mỗi hàng là một mẫu và mỗi cột là một biến.
- $\bar{X}$ là vector trung bình của các cột của X
- n là số lượng mẫu.

Bước 2: Phân tích giá trị riêng và vector riêng

Eigenvalue Decomposition: Tính toán các giá trị riêng λ và các vector riêng v của ma trận hiệp phương sai Σ. Các giá trị riêng và các vector riêng này thỏa mãn phương trình sau:

$$\Sigma v = \lambda v$$

Trong đó, λ là giá trị riêng và v là vector riêng tương ứng.

Bước 3: Lựa chọn thành phần chính

Select Principal Components: Chọn các thành phần chính bằng cách sắp xếp các giá trị riêng theo thứ tự giảm dần và chọn ra các vector riêng tương ứng với các giá trị riêng lớn nhất.

Bước 4: Chiếu dữ liệu lên các thành phần chính

Projection: Chiếu dữ liệu gốc X lên các thành phần chính đã chọn để thu được dữ liệu mới Y có số chiều thấp hơn. Quá trình này có thể được thực hiện bằng cách nhân ma trận dữ liệu X với ma trận các thành phần chính được chọn.

$$Y = X.V_k$$

Trong đó: V_k là ma trận chứa các vector riêng tương ứng với k giá trị riêng lớn nhất đã chọn.

3 Mức độ hoàn thành

Study PCA	*45%*
Implementation PCA	*45%*
Overall comprehensive of submitted source code	*10%*
Total	*100%*