Báo cáo Principle Components Analysis (PCA)

Họ và tên: Trần Anh Duy

MSSV: 21120235

I. Động lực tìm hiểu PCA:

Trong các bài toán học máy thì dữ liệu có kích thước rất lớn. Máy tính có thể hiểu và thực thi các thuật toán trên dữ liệu này, tuy nhiên đối với con người để "nhìn" dữ liệu nhiều chiều thật sự là rất khó. Vì vậy bài toán giảm chiều dữ liệu ra đời giúp đưa ra cái nhìn mới cho con người về dữ liệu nhiều chiều. Ngoài để trực quan dữ liệu, các phương pháp giảm chiều dữ liệu còn giúp đưa dữ liệu về một không gian mới giúp khai phá các thuộc tính ẩn mà trong chiều dữ liệu ban đầu không thể hiện rõ, hoặc đơn giản là giảm kích thước dữ liệu để tăng tốc độ thực thi cho máy tính.

II. Giới thiệu PCA:

- Phân tích thành phần chính (PCA) là một kỹ thuật quan trọng trong lĩnh vực xử lý dữ liệu và thống kê.
- PCA là phương pháp biến đổi giúp giảm số lượng lớn các biến có tương quan với nhau thành tập ít các biến sao cho các biến mới tạo ra là tổ hợp tuyến tính của những biến cũ không có tương quan lần nhau.
- Ví dụ, chúng ta có 100 biến ban đầu có tương quan tuyến tính với nhau, khi đó chúng ta sử dụng phương pháp PCA xoay chiều không gian cũ thành chiều không gian mới mà ở đó chỉ còn 5 biến không có tương quan tuyến tính mà vẫn dữ được nhiều nhất lượng thông tin từ nhóm biến ban đầu.

III. Thuật toán PCA:

Ý tưởng: Về mặt ý tưởng, thuật toán PCA tìm một hệ không gian mới và tối đa hóa phương sai dữ liệu của không gian mới đó. Sau đó lựa chọn ra n chiều có phương sai lớn nhất.

1. Đầu vào:

- Ma trận dữ liệu:Ma trận này chứa các điểm dữ liệu được biểu diễn trong không gian nhiều chiều. Mỗi hàng trong ma trận đại diện cho một điểm dữ liệu và mỗi cột đại diện cho một chiều. Kích thước của ma trận là n x m.
- Số lượng thành phần chính: Số lượng thành phần chính (PCs) mong muốn là số chiều của không gian dữ liệu được giảm sau khi áp dụng PCA.

2. Đầu ra:

- Ma trận thành phần chính: Ma trận này chứa các thành phần chính, được gọi là các vectơ riêng của ma trận hiệp phương sai của dữ liệu. Mỗi hàng trong ma trận đại diện cho một thành phần chính và mỗi cột đại diện cho phương sai của dữ liệu được dự đoán bởi thành phần chính đó. Kích thước của ma trận là p x k, với k là số lượng thành phần chính được chọn.
- Ma trận điểm dữ liệu được chiếu: Ma trận này chứa các điểm dữ liệu được biểu diễn trong không gian thành phần chính mới. Mỗi hàng trong ma trận đại diện cho một điểm dữ liệu và mỗi cột đại diện cho tọa độ của điểm dữ liệu đó trên thành phần chính tương ứng. Kích thước của ma trận là n x k.

3. Các bước thực hiện PCA:

Bước 1: Chuẩn bị dữ liệu

- Ta phải chuẩn bị dữ liệu cần được giảm chiều với n mẫu và m đặt trưng

Bước 2: Chuẩn hóa dữ liệu

- PCA chỉ hoạt động hiệu quả khi các đặt trưng trong dữ liệu có cùng tỉ lệ.
- Ở đây ta sẽ xử dụng phương pháp **StandardScaler** để chuẩn hoá dữ liệu đầu vào.

Bước 3: Trừ mỗi điểm dữ liệu cho vector kì vọng:

$$X_k = X_k - X_{mean}$$

Bước 4: Tính ma trận hiệp phương sai

$$Cov(X) = \frac{1}{n-1} \cdot X^T \cdot X$$

Bước 5: Tính các trị riêng và vector riêng của ma trận hiệp phương sai để xác định các thành phần chính:

- Tiếp theo, chúng ta tính toán các vector riêng và giá trị riêng tương ứng của ma trận hiệp phương sai.

Bước 6: Lấy k trị riêng có giá trị lớn nhất, tạo ma trận U với các hàng là các vector riêng ứng với k trị riêng đã chọn.

Bước 7: Ánh xạ không gian ban đầu sang không gian k chiều:

- Để thực hiện việc này, ta nhân ma trận dữ liệu gốc với ma trận chứa các vector riêng của ma trận hiệp phương sai (các thành phần chính). Kết quả thu được chính là dữ liệu đã được tái tổ chức theo các thành phần chính.

IV. Trình bày background toán:

1. Ý nghĩa của việc sử dụng trị riêng:

- **Trị riêng** đại diện cho mức độ **biến đổi** của dữ liệu theo từng **hướng** trong không gian dữ liệu.
- Việc phân tích **trị riêng** giúp ta **hiểu rõ hơn về cấu trúc** của dữ liệu và **xác định** những **hướng** chứa nhiều **thông tin** nhất.
- Bằng cách chọn **k** trị riêng lớn nhất, ta có thể **chiếu dữ liệu** xuống **k** chiều mới, giữ lại **phần thông tin quan trọng nhất**.

2. Ý nghĩa của việc xây dựng phép biến đổi tuyến tính sang không gian mới:

- Giảm số lượng biến trong khi vẫn giữ lại thông tin quan trọng của dữ liệu.
- Việc chiếu dữ liệu xuống không gian mới có thể phơi bày những mối liên hệ giữa các biến mà trước đây bị che khuất bởi nhiễu và độ phức tạp cao của dữ liệu.
- Việc sắp xếp trị riêng theo thứ tự giảm dần giúp ta xác định các hướng quan trọng nhất trong không gian dữ liệu.

3. Có cách nào chọn một số k phù hợp cho PCA không?

- Phương pháp phổ biến nhất là chọn k sao cho tổng tỷ lệ phương sai của các trị riêng được chọn đạt đến một mức mong muốn (thường là 95%)

V. Vấn đề của PCA:

Mặc dù là một kỹ thuật giảm chiều dữ liệu hiệu quả và được sử dụng rộng rãi, Phân tích thành phần chính (PCA) cũng tồn tại một số hạn chế cần được lưu ý:

- PCA không trực tiếp cung cấp cách giải thích cho các thành phần chính được tạo ra. Việc hiểu ý nghĩa của các thành phần này có thể khó khăn, đặc biệt khi dữ liệu có nhiều chiều và các thành phần chính là sự kết hợp tuyến tính của nhiều biến gốc.
- PCA có thể nhạy cảm với nhiễu trong dữ liệu. Nếu dữ liệu bị nhiễu bởi các giá trị ngoại lệ hoặc điểm dữ liệu bất thường, kết quả PCA có thể bị ảnh hưởng đáng kể.
- Quá trình giảm chiều dữ liệu bằng PCA luôn dẫn đến việc mất một phần thông tin trong dữ liệu gốc. Mức độ mất mát thông tin phụ thuộc vào số lượng thành phần chính được chọn.
- PCA không phù hợp với tất cả các loại dữ liệu. Ví dụ, PCA không hiệu quả với dữ liệu thưa thớt hoặc dữ liệu có nhiều giá trị thiếu.