**BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA CÔNG NGHỆ THÔNG TIN**

A picture containing chart

Description automatically generated

BÁO CÁO   
BÀI PROJECT - PCA

Giảng viên hướng dẫn: Trần Hà Sơn

Họ và tên: Trần Đình Nhật Trí – Trương Anh Tuấn

Mã số sinh viên: 21120576 - 21120589

1. Thông tin cá nhân:

|  |  |
| --- | --- |
| Họ và tên | MSSV |
| Trần Đình Nhật Trí | 21120576 |
| Trương Anh Tuấn | 21120589 |

1. Định nghĩa PCA:

* PCA (Principal Component Analysis) là phương pháp phân tích thành phần chính, mục đích là để giảm chiều dữ liệu của một tập dữ liệu lớn. Phương pháp này dùng chủ yếu để đơn giản hóa một tập dữ liệu lớn nhưng vẫn giữ lại những thành phần quan trọng nhất của tập dữ liệu đó (thành phần chính).

Lợi ích của PCA:

+ Giảm chiều dữ liệu nhưng vẫn giữ được đặc trưng cơ bản của dữ liệu.

+ Tiết kiếm thời gian và chi phí để phân tích dữ liệu

+ Đưa ra cái nhìn trực quan, tổng quát

+ Xác định các biến tương quan.

Ví dụ ta có một hệ tọa độ thể hiện một tập dữ liệu như sau:

A picture containing line, handwriting, diagram, text

Description automatically generated

*Hình 1: Hệ tọa độ thể hiện tập dữ liệu với số thành phần chính là 2*

* Theo lý thuyết của PCA, các hướng có phương sai lớn nhất là quan trọng nhất, bởi vì ở các hướng đó, các điểm dữ liệu biến động nhiều nhất. Vì thế trên ảnh ta thấy hai hướng 1st Component và 2nd Component. Trục 1st Component là hướng chính đầu tiên mà các điểm dữ liệu đó biến thiên nhiều nhất, tiếp đến là trục 2ndComponent. Xét một cách nhìn khác, PCA cũng là một bài toàn chuyển hệ tọa độ: A picture containing line, handwriting, diagram, text

  Description automatically generated

*Hình 2. Thể hiện việc chuyển hệ tọa độ của các thành phần chính*

* Lý do chọn các chiều dữ liệu mà theo chiều đó các điểm mẫu có độ biến thiên lớn nhất là vì nếu độ biến quá ít, thì các các điểm mẫu sẽ co cụm lại với nhau thì ta sẽ không phân loại được. Sau đây là một hình ảnh minh họa khác của PCA:

A picture containing text, diagram, line, plot

Description automatically generated  
*Hình 3. Minh họa khả năng phân hoạch của PCA theo số chiều được chọn.*

1. Liên hệ với SVD:

* SVD và PCA có liên hệ với nhau nhưng không giống nhau hoàn toàn.
* SVD là một phương pháp phân rã ma trận thành ba ma trận khác:

A = UΣVT . Cho A là ma trận trực chuẩn m x m, thì trong đó, U là ma trận trực chuẩn m x n, Σ là ma trận chéo hóa m x n (không vuông), V là ma trận trực chuẩn n x n.

* Mặt khác, PCA là một phương pháp giúp giảm chiều dữ liệu nhưng vẫn đảm bảo những đặc trưng chính của tập dữ liệu đó. PCA cũng được xem là một ứng dụng đơn giản của SVD. Áp dụng SVD lên ma trận dữ liệu, các thành phần chính (principle components) sẽ được lấy từ các vector riêng (singular vectors).

1. Ý tưởng thực hiện:
2. Image compressing:

Ý tưởng của việc sử dụng PCA (Principal Component Analysis) trong việc nén ảnh (image compressing) là giảm số chiều của dữ liệu ảnh bằng cách giữ lại chỉ các thành phần chính quan trọng nhất.

Quá trình nén ảnh sử dụng PCA thông thường có các bước sau:

1. Chuyển đổi ảnh màu sang ảnh đen trắng (grayscale) nếu cần thiết, để giảm số lượng kênh màu từ 3 xuống 1.
2. Áp dụng PCA lên tập dữ liệu các khối ảnh. PCA sẽ tìm các thành phần chính của các khối ảnh và giảm số chiều của tập dữ liệu.
3. Chọn số thành phần chính cần giữ lại để đảm bảo tỷ lệ nén mong muốn.
4. Sử dụng các thành phần chính đã chọn để xây dựng lại ảnh từ các vectơ dữ liệu thu được sau PCA.
5. Khi xây dựng lại ảnh, thông tin không quan trọng đã bị mất đi trong quá trình giảm số chiều, điều này dẫn đến việc nén ảnh.
6. Dự đoán bệnh ung thư vú:

Để sử dụng PCA cho dự đoán bệnh ung thư vú, ta làm như sau:

* 1. Chuẩn hóa dữ liệu: ta sử dụng StandardScaler để chuẩn hóa các dữ liệu đầu vào thành một phân phối chuẩn với trung bình bằng 0 và độ lệch chuẩn bằng 1 để đảm bảo rằng mỗi đặc trưng có cùng phạm vi giá trị, giúp mô hình hoạt động tốt hơn.
  2. Tiến hành PCA: huấn luyện PCA trên tập dữ liệu, sau đó áp dụng nó để giảm số chiều của dữ liệu ban đầu. PCA sẽ tạo ra các thành phần chính mới, được sắp xếp theo độ quan trọng giảm dần tùy vào số chiều mình mong muốn (luôn bé hơn số chiều ban đầu).

---- HẾT ----