PCA VÀ LDA

PHÂN TÍCH THÔNG KÊ DỮ LIỆU NHIỀU BIẾN

Hướng dẫn: Nguyễn Mạnh Hùng

MỤC LỤC

THÔNG TIN

Cá Nhân

MSSV	HỌ TÊN	EMAIL	SĐT
19127517	Hồ Thiên Phước	htphuoc19@clc.fitus.edu.vn	-

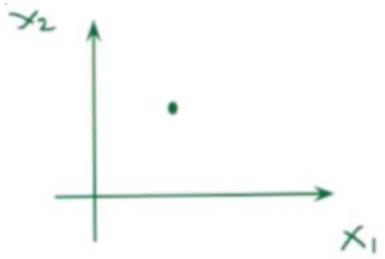
PCA

Tính Variance là khai thác, làm cho khoảng biến thiên lớn nhất có thể. Variance thể hiện độ lệch so với giá trị trung bình.
Khoảng biến thiên thể hiện tính khả tách, khoảng biến thiên càng lớn thì tính khả tách càng cao.
Covariance nói lên sự tương quan của 2 biến.
Ràng buộc:

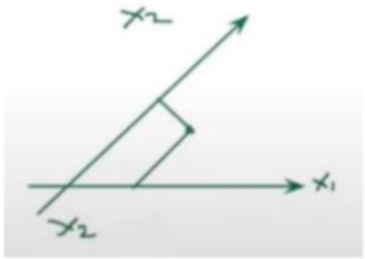
1) e_i (vector cơ sở, vector của hệ tọa độ mới sau khi dùng phép biến đổi tuyến tính chiếu xuống) phải là vector đơn vị, nếu không là vector đơn vị thì ta tiến hành chuẩn hóa theo chiều dài. Sau đây là ràng buộc cho cực trị:

$$e_i' e_i = 1 \text{ hoặc } e_i' e_j = 0, \forall i \neq j$$

2)



Góc vuông (trực giao đôi một), thể hiện độc lập tuyến tính



Góc méo, thể hiện sự phụ thuộc tuyến tính

$$Cov(Y_i, Y_j) = 0, \forall i \neq j$$

NHẬN XÉT: Ràng buộc tại Phát biểu bài toán:

1)Không gian mới có số chiều nhỏ hơn.
2)Các thành phần giữ lại sau khi giảm chiều phải có tính không tương quan (độc lập tuyến tính). Các vector tương ứng với các trục trong không gian mới đòi hỏi trực giao đôi một. Vì thế chúng ta tính Covariance.
3)Khoảng biến thiên mong muốn là lớn nhất có thể.

Các hệ vector riêng. ei'ei=ei'iei (vì ei=iei) = i (vì ei là vector đơn vị)
Giảm chiều là sau khi tìm được các hệ vector riêng và giá trị riêng sắp xếp giảm dần, ta giữ lại k giá trị riêng ứng với k vector riêng lớn nhất, còn lại sẽ bị bỏ đi.

1. **Tính**

$$X' = [X_1, X_2, ..., X_p]$$

Ma trận hiệp phương sai Cov là

$$Y_1 = a_1' X = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Y_2 = a_2' X = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p' X = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$
$$Var(Y_i) = a_i' \sum a_i, \ i = 1,2,...,p$$
$$Cov(Y_i, Y_j) = a_i' \sum a_j, i, j = 1,2,...,p$$

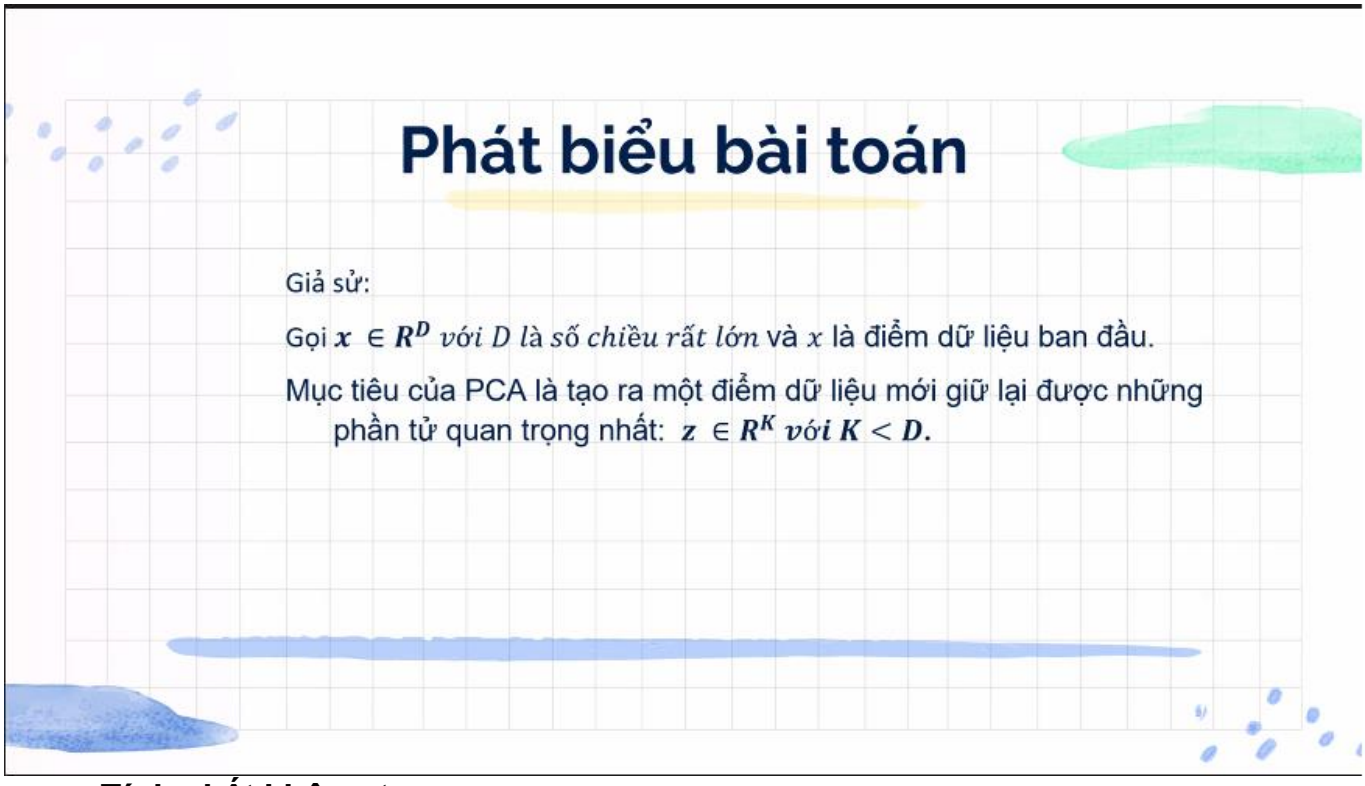
Độ lệch của từng vector (điểm đặc trưng) so với vector trung bình centroid thể hiện ở:

$Var(Y_i) = a_i' \sum a_i = a_i' \lambda_i a_i$ (vì $\sum a_i = \lambda_i a_i$) = λ_i (vì a_i là vector đơn vị)

Khoảng biến thiên cho biết tính khả tách thể hiện ở:

$Cov(Y_i, Y_j) = a_i' \sum a_j = a_i' \lambda_j a_j = 0$ (do điều kiện các vector phải trực giao đôi một)

Phát biểu bài toán



Tính chất không tương quan

Tính chất không tương quan

Cho ma trận

S= $\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ . & . & \cdots & . \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$ có s_{ij} với $i \neq j$ $(i, j \leq n)$ được gọi là hiệp phương sai, thể hiện sự tương quan giữa thành phần thứ i và j của dữ liệu.

Các giá trị Covariance này có thể dương, âm hoặc bằng 0. Khi giá trị nó bằng 0 thì ta có thể nói hai thành phần i, j trong dữ liệu là không tương quan.

Ta sử dụng PCA để giảm số chiều dữ liệu, đồng thời giữ lại các thành phần trong các chiều độc lập (không tương quan với nhau) để tránh việc trùng lắp dữ liệu.

Đặc tính

Đặc tính

1. Giúp giảm số chiều dữ liệu - Giúp hỗ trợ trực quan hóa dữ liệu khi dữ liệu có quá nhiều chiều thông tin.

2. Do dữ liệu ban đầu có số chiều lớn (nhiều biến) thì PCA giúp chúng ta xoay trục tọa độ xây một trục tọa độ mới đảm bảo độ biến thiên của dữ liệu và giữ lại được nhiều thông tin nhất mà không ảnh hưởng tới chất lượng của các mô hình dự báo. (Maximize the variability).

3. Do PCA giúp tạo 1 hệ trục tọa độ mới nên về mặt ý nghĩa toán học, PCA giúp chúng ta xây dựng những biến factor mới là tổ hợp tuyến tính của những biến ban đầu.

4. Trong không gian mới, có thể giúp chúng ta khám phá thêm những thông tin quý giá mới khi mà tại chiều thông tin cũ những thông tin quý giá này bị che mất.

Phương pháp

Phương pháp

Vậy ta có:

$\text{Var}(Y_i) = a_i' \Sigma a_i = \lambda_i$

$\text{Cov}(Y_i, Y_j) = a_i' \lambda_j a_j = 0$

Từ 2 phương trình trên ta tính được các vector λ_i, sau đó sắp xếp chúng theo thứ tự giảm dần như sau:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p$$

Sau đó ta chọn k (với k < p) λ lớn nhất.

Thuật toán

Thuật toán PCA

Bước 1: Tính vector kỳ vọng của toàn bộ dữ liệu:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

Bước 2: Tính vector độ lệch so với giá trị trung bình của toàn bộ dữ liệu:

$$\widehat{X_n} = x_n - \bar{x}$$

Bước 3: Tính ma trận hiệp phương sai:

$$\Sigma = \frac{1}{N} \hat{X} \widehat{X^T}$$

Bước 4: Tìm các giá trị riêng (λ) và vector riêng (a) của ma trận phương sai, sắp xếp chúng theo thứ tự giảm dần:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

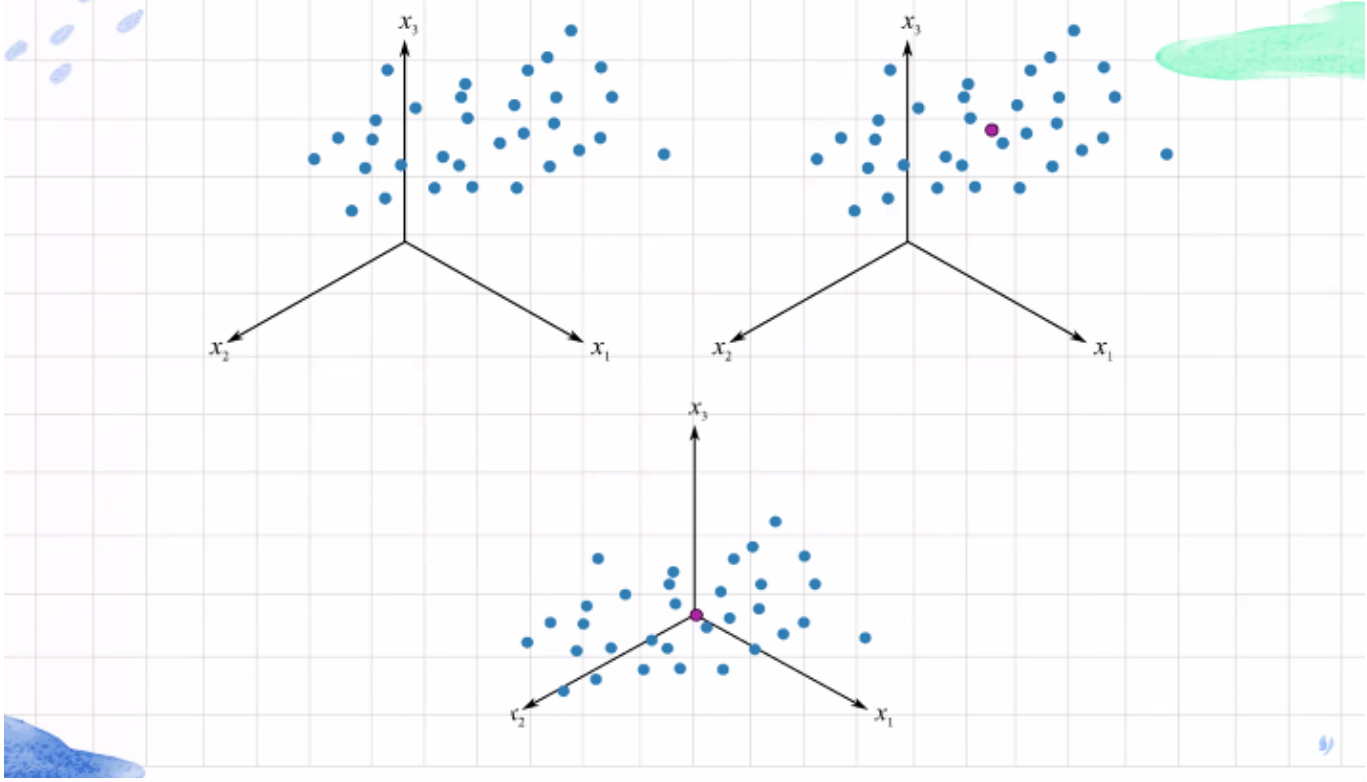
Bước 5: Chọn K vector riêng ứng với K trị riêng lớn nhất để xây dựng ma trận. Chuyển hóa dữ liệu gốc thành dữ liệu trong không gian mới bằng một phép nhân ma trận đơn giản.

$F = X.A$ với $A = [a_1 | a_2 | \ldots | a_k]$ là tập hợp các vector riêng được chọn làm thành phần chính

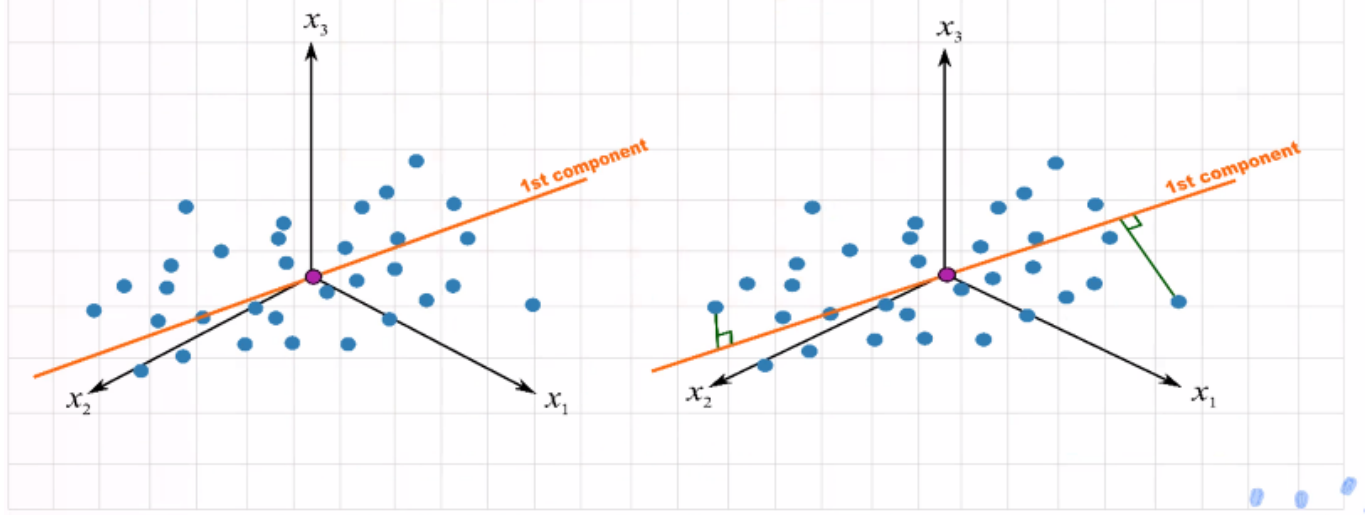
Độ đo

Ý nghĩa hình học PCA

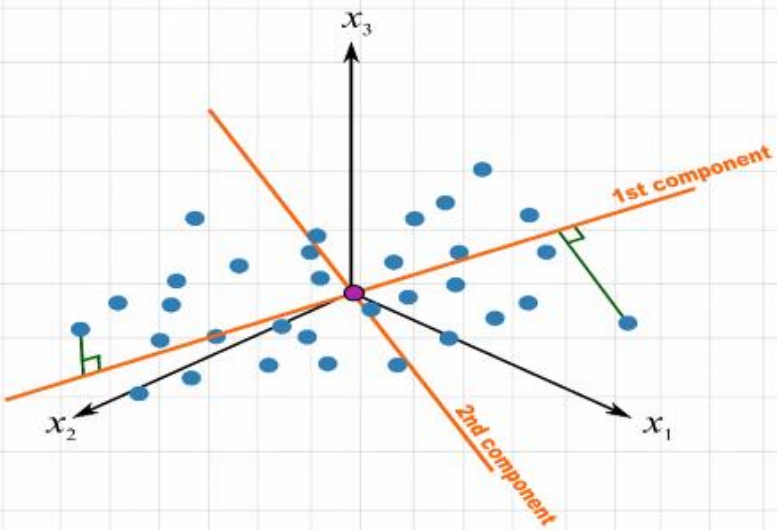
Bước 1: Dời hệ trục tọa độ.



Bước 2: Vẽ đường thành phần chính (principal component) đầu tiên.

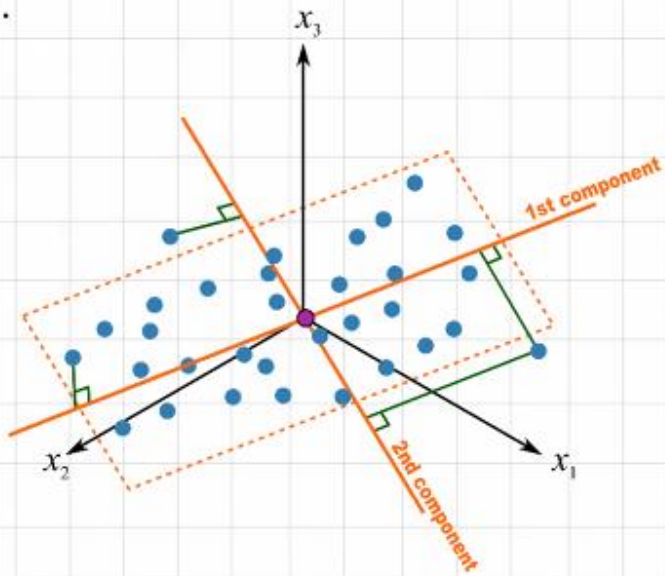


Bước 3: Vẽ các đường thành phần còn lại.



Bước 4: Giảm số chiều của hệ trục tọa độ.

Bước 5: Vẽ lại hệ trục tọa độ mới dựa theo số lượng đường thành phần chính.

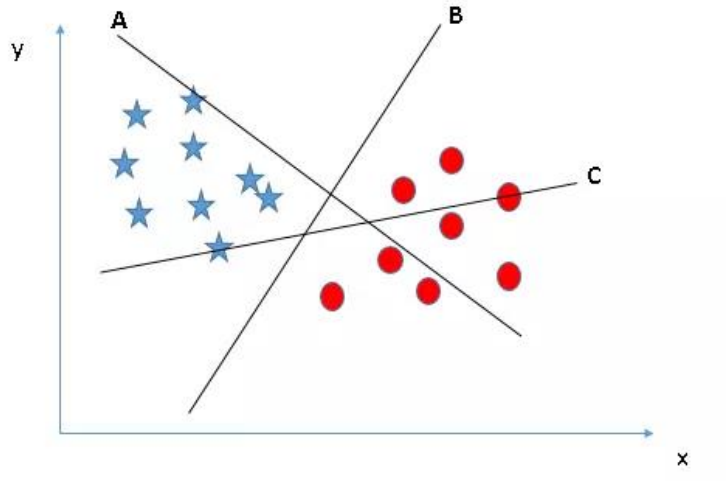


LDA

MỘT SỐ KHÁI NIỆM CƠ BẢN

Từ bức ảnh ban đầu được cấu thành bởi nhiều pixel, sau khi được chia ô lưới trong không gian ảnh, ta duyệt qua ảnh bằng cách nhóm 2,3,4 hoặc 5 pixel lại thành 1 nhóm và chọn ra pixel đặc trưng đại diện cho nhóm đó. Các pixel đặc trưng được đưa vào không gian đặc trưng.

Những dữ liệu đã được gán nhãn là những toạ độ pixel do con người sử dụng tool LabelImg click chuột vào từng ảnh trong dataset để xác định vị trí của đối tượng trong ảnh và các pixel được gom vào cùng 1 class do con người tự đặt tên class.



Không gian đặc trưng

Mỗi đường hình chuông biểu diễn 1 class trong không gian đặc trưng. Độ rộng của mỗi đường hình chuông thể hiện độ lệch chuẩn của dữ liệu. Dữ liệu tập trung thì độ lệch chuẩn nhỏ, dữ liệu "lệch ra xa" thì độ lệch chuẩn lớn.

PHÁT BIỂU BÀI TOÁN

Input: X- là một mảng các toạ độ pixel trong 'dataset đã gán nhãn'

Output: wT (phép chiếu, là một phép biến đổi tuyến tính)

Cơ chế hoạt động:

- B1: Tính ma trận within sw
- B2: Tính ma trận between sB
- B3 Tìm vector phép chiếu tốt nhất (tính trị riêng, tính vector riêng, tính vector cơ sở theo vector riêng) sw-1.sB.w=w
- B4: Giảm số chiều y=wTx

Thách thức:

LDA/Fisher với nhiều quần thể.…

ĐỊNH NGHĨA

N: số điểm dữ liệu 2 class

w^T: phép chiếu - phép biến đổi tuyến tính

N_1: Điểm đầu tiên của class 1

$N_2 = N - N_1$: Điểm cuối cùng class 2

C_1: Các điểm thuộc class 1

C_2: Các điểm thuộc class 2

Mỗi điểm sau khi chiếu:

$$y_n = \mathbf{w}^T\mathbf{x}_n, 1 \leq n \leq N$$

Chain rule cho đạo hàm nhiều biến nếu ma trận A đối xứng:

$$\nabla_{\mathbf{w}}\mathbf{w}\mathbf{A}\mathbf{w} = 2\mathbf{A}\mathbf{w}$$

Ta có đẳng thức:

$$(\mathbf{a}^T\mathbf{b})^2 = (\mathbf{a}^T\mathbf{b})(\mathbf{a}^T\mathbf{b}) = \mathbf{a}^T\mathbf{b}\mathbf{b}^T\mathbf{a}$$

Vector kỳ vọng của mỗi class:

$$\mathbf{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{x}_n, \quad k = 1, 2$$

Khoảng cách 2 chuông sau khi chiếu:

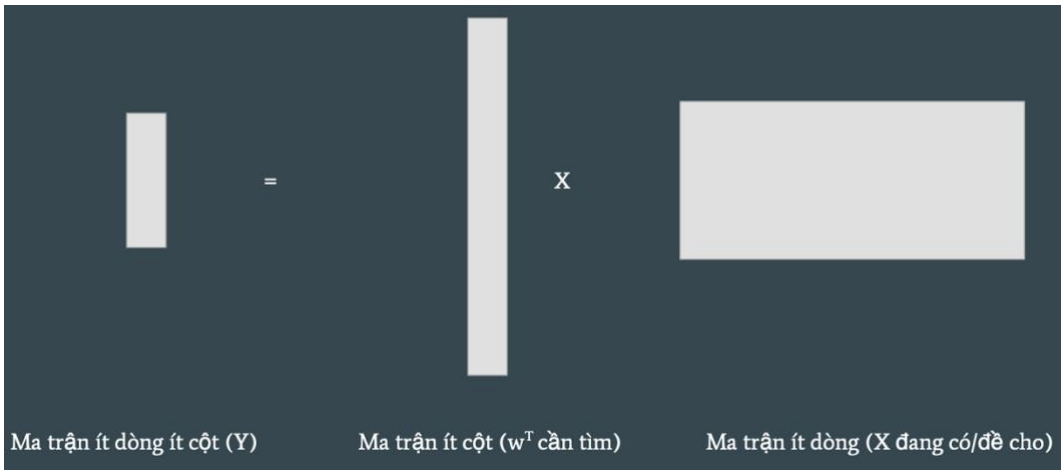
$$m_1 - m_2 = \frac{1}{N_1}\sum_{i\in\mathcal{C}_1} y_i - \frac{1}{N_2}\sum_{j\in\mathcal{C}_2} y_j = \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)$$

Mức tập trung của mỗi class:

$$s_k^2 = \sum_{n\in\mathcal{C}_k}(y_n - m_k)^2, \quad k = 1, 2$$

Vấn đề là:

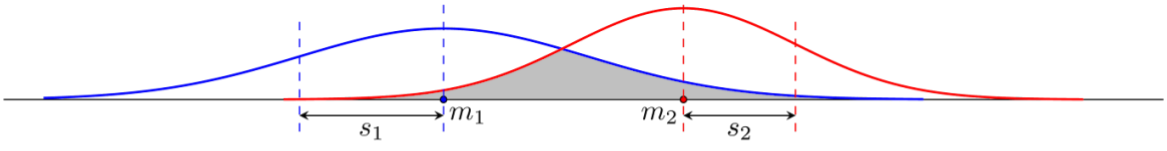
1) Số chiều lớn (các pixel trong dataset sau gán nhãn nhiều) nên ta cần phải giảm chiều dữ liệu.



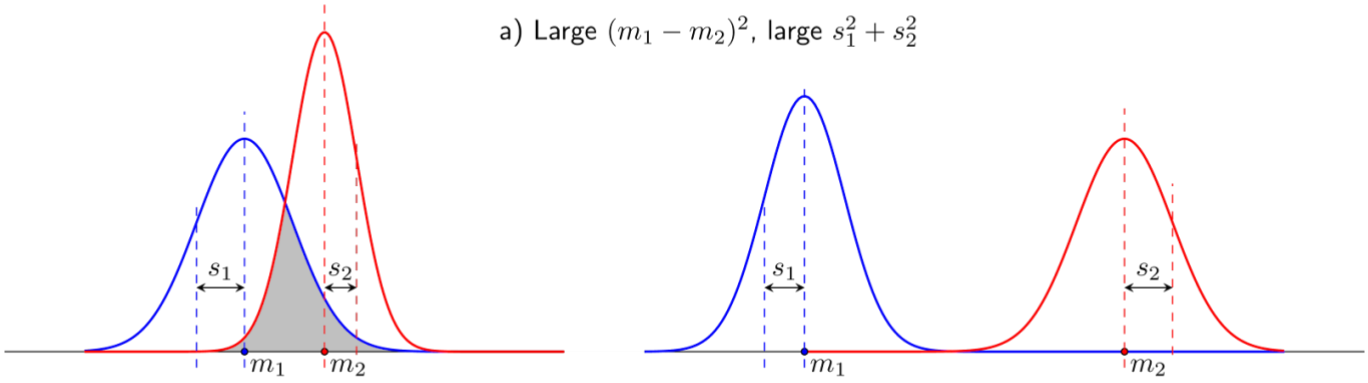
Giảm chiều

2) Trong không gian đặc trưng, vì không có "chỗ đứng" hợp lí nên các vector đặc trưng sẽ đứng hỗn loạn gây lẫn lộn giữa class này với class khác.

LDA/Fisher ra đời để khắc phục tình trạng này. LDA/Fisher là thuật toán có giám sát **đi tìm phép chiếu** từ không gian đặc trưng này sang không gian đặc trưng khác có số chiều nhỏ hơn mà vẫn giữ được tính khả tách của dữ liệu. Vì mong muốn khoảng cách 2 class tách ra xa để đỡ nhầm lẫn class, nên tính khả tách được mô tả là within-class- độ lệch chuẩn trong 1 class nhỏ (đồng nghĩa tính tập trung của các cá thể trong 1 class phải gần nhau) và between-class- khoảng cách 2 kỳ vọng phải đủ lớn.



a) Large $(m_1 - m_2)^2$, large $s_1^2 + s_2^2$



b) Small $(m_1 - m_2)^2$, small $s_1^2 + s_2^2$

c) Large $(m_1 - m_2)^2$, small $s_1^2 + s_2^2$

Hình c) đạt được 2 mong muốn trực quan là hình chuông phải 'gầy' và 2 hình chuông 'xa nhau'

Để chuyển đổi 2 mong muốn này thành biểu thức toán học, mong muốn của Fisher là đồng thời cực đại bình phương khoảng cách của 2 lớp và cực tiểu tổng 2 phương sai của 2 lớp. Đây là 2 mong muốn tỉ lệ với nhau. Để fisher tối ưu thì fisher phải đạt cực đại. Vậy khi chuyển 2 mong muốn này vào phương trình toán học, ta sẽ để mong muốn đạt cực đại vào tử số của phân số và mong muốn cực tiểu sẽ để ở mẫu số của phân số, và khi ta cực đại phân số này, fisher sẽ tối ưu, mẫu số sẽ tự khắc cực tiểu và tử số sẽ tự khắc cực đại.

$$\frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

Vì ta đang tìm phép chiếu w^T, nên ta cần làm xuất hiện w^T trong phân số này. Và sau đó cho đạo hàm của phân số = 0 để tìm nghiệm sao cho phân số này tối ưu nhất.

Tử số - Between:

$$\left(\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1-\mathrm{m}_2)\right)^2=\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w}=\mathrm{w}^{\mathrm{T}}s_B w$$

Mẫu số - Within:

$$\mathrm{s}_1^2+\mathrm{s}_2^2=\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}\left(\mathrm{w}^{\mathrm{T}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})\right)^2=\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}w=w^T s_w w$$

$$\Rightarrow\ \mathrm{w}^*=\underset{\mathrm{w}}{\operatorname{argmax}}\frac{\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w}}{\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w}}=\underset{\mathrm{w}}{\operatorname{argmax}}\frac{\mathrm{w}^{\mathrm{T}}\mathrm{s}_{\mathrm{B}}\mathrm{w}}{\mathrm{w}^{\mathrm{T}}\mathrm{s}_{\mathrm{w}}w}$$

$(\mathrm{w}^*)'=0$

$$\Rightarrow\ \frac{2(\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w}.\left(\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w}\right)-2(\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w})(\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w})}{(\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w})^2}=0$$

$$\Rightarrow\ (\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w}=\frac{\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}\mathrm{w}}{\mathrm{w}^{\mathrm{T}}\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w}}.\ \sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\mathrm{w}$$

$$\Rightarrow\ \left(\sum_{\mathrm{k}=1}^{2}\sum_{\mathrm{n}\in\mathrm{C}_\mathrm{k}}(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})(\mathrm{x}_\mathrm{n}-\mathrm{m}_\mathrm{k})^{\mathrm{T}}\right)^{-1}((\mathrm{m}_1-\mathrm{m}_2)(\mathrm{m}_1-\mathrm{m}_2)^{\mathrm{T}}).\mathrm{w}=\mathrm{w}^*.\mathrm{w}$$

$$\Rightarrow\ s_w^{-1}s_B.\mathrm{w}=\mathrm{w}^*.\mathrm{w}$$

Vậy w là vector riêng của $s_w^{-1}s_B$ ứng với trị riêng w*. Để w* tối ưu thì w* phải đạt cực đại. Suy ra w* phải là trị riêng lớn nhất.

VÍ DỤ

$$X_1=(4,1),(2,4),(2,3),(3,6),(4,8)$$
$$X_2=(9,10),(6,8),(9,5),(8,7),(10,8)$$

Bước 1: $s_w=s_1+s_2$

$$S_1=\sum_{x\in C_1}(x-\mu_1)(x-\mu_1)^\top$$

$$\mu_1=\left\{\frac{4+2+2+3+4}{5},\frac{1+4+3+6+4}{5}\right\}$$
$$\mu_1=[3.00\quad 8.60]$$

$$(\chi_1-\mu_1)=\begin{bmatrix}1. & -1 & -1 & 0 & 1\\ -2.6 & 0.4\cdot & -0.6 & 2.4 & 0.4\end{bmatrix}$$

$x_{11}x_{11}^T=$

$$\begin{bmatrix}1\\-2.6\end{bmatrix}\cdot[1 \quad -2.6]=\begin{bmatrix}1 & -2.6\\-2.6 & 6.76\end{bmatrix}$$

Chứng minh tương tự (cmtt):

$$\begin{bmatrix}-1\\0.4\end{bmatrix}[-1 \quad 0.4]=\begin{bmatrix}1 & -0.4\\-0.4 & 0.16\end{bmatrix}$$
$$\begin{bmatrix}-1\\-0.6\end{bmatrix}[-1-0.6]=\begin{bmatrix}1 & 0.6\\0.6 & 0.36\end{bmatrix}$$
$$\begin{bmatrix}0\\2.4\end{bmatrix}[0 \quad 2.4]=\begin{bmatrix}0 & 0\\0 & 5.76\end{bmatrix}$$
$$\begin{bmatrix}1\\0.4\end{bmatrix}[1 \quad 0.4]=\begin{bmatrix}1 & 0.4\\0.4 & 0.16\end{bmatrix}$$

Tính tổng các cột ta thu được

$$S_1=\begin{bmatrix}0.8 & -0.4\\-0.4 & 2.6\end{bmatrix}$$

Cmtt ta thu được s_2 rồi ta tính s_w

$$S_2=\begin{bmatrix}1.84 & -0.04\\-0.04 & 2.64\end{bmatrix}$$
$$S_W=S_1+S_2.$$
$$S_W=\begin{bmatrix}2.64 & -0.44\\-0.44 & 6.28\end{bmatrix}$$

Bước 2: Tìm s_B

$$\begin{aligned}S_B=&(\mu_1-\mu_2)(\mu_1-\mu_2)^\top\\=&\begin{pmatrix}-5.4\\-4\end{pmatrix}(-5.4 \quad -4)=\begin{pmatrix}29.16 & 21.6\\21.6 & 16.00\end{pmatrix}\end{aligned}$$

Bước 3: Tìm w^T từ $s_w^{-1}s_B.\mathrm{w}=\lambda.\mathrm{w}$

$$S_W^{-1}S_BW=\lambda W \quad (*)$$
$$\left|S_W^{-1}S_B-\lambda I\right|=0$$
$$\begin{vmatrix}11.89-\lambda & 8.81\\5.08 & 3.76-\lambda\end{vmatrix}=0$$

I=[1 0 0
0 1 0
0 0 1]

Giải và chọn λ lớn nhất

$$\lambda=15.65$$

Đã có s_B, tìm s_w^{-1}

$$\begin{bmatrix}a & b\\c & d\end{bmatrix}^{-1}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}$$
$$S_W=\begin{bmatrix}2.64 & -0.44\\-0.44 & 5.28\end{bmatrix}$$
$$S_W=\frac{1}{13.74}\begin{bmatrix}5.28 & 0.44\\0.44 & 2.64\end{bmatrix}=\begin{bmatrix}0.384 & 0.032\\0.032 & 0.192\end{bmatrix}$$
$$S_W^{-1}=\frac{1}{13.74}\begin{bmatrix}5.28 & 0.44\\0.44 & 2.64\end{bmatrix}=\begin{bmatrix}0.384 & 0.032\\0.032 & 0.122\end{bmatrix}$$

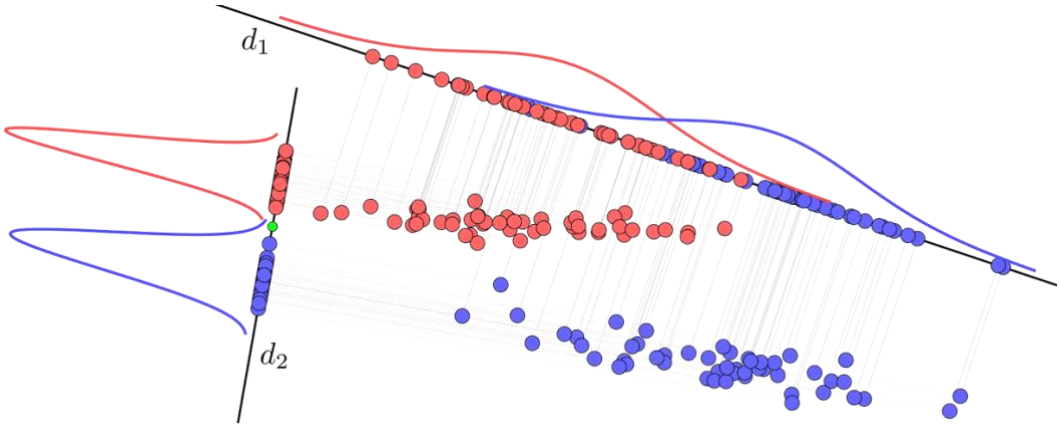
Bước 4: Chiếu

$$y=W^\top x$$

y_1= [0.91 0.39][4 1

2 4
2 3
3 6
4 8]
y_2=[0.91 0.39][9 10
6 8
9 5
8 7
10 8]

SO SÁNH PCA VÀ LDA



d1: PCA, d2: LDA

	LDA	PCA
Mô hình học	Học có giám sát - là thuật toán tiên đoán nhãn cho dữ liệu mới dựa trên tập huấn luyện (các mẫu trong tập này đều đã được gán nhãn)	Học không giám sát - là thuật toán tiên đoán nhãn cho dữ liệu mới dựa trên tập huấn luyện (các mẫu trong tập này đều chưa được gán nhãn)
Xét phương sai	Có	Không
Giảm chiều	Có	Có
Số lượng thông tin	Cần thiết	Nhiều nhất
Kết quả phân loại	Tốt hơn	..

MÔI TRƯỜNG

CÔNG CỤ	NGÔN NGỮ

Visual Studio Code	Python

HƯỚNG DẪN CÀI ĐẶT

Cài đặt pip

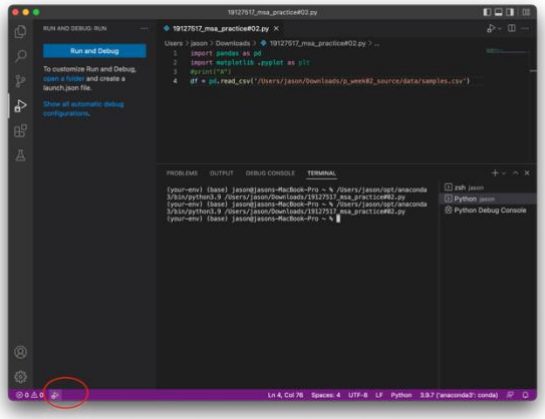
```
python get-pip.py
```

hoặc

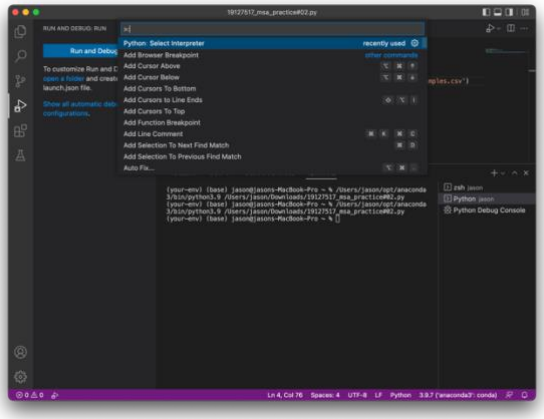
```
python -m pip install --upgrade pip
```

Cài đặt venv

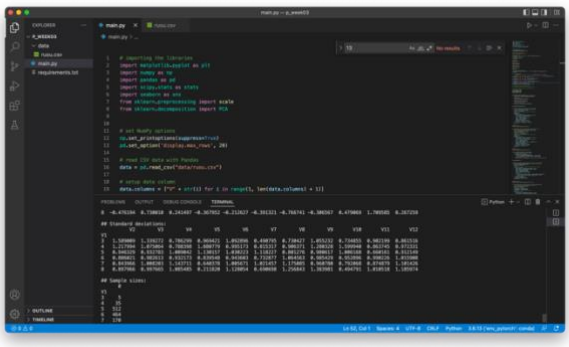
```
pip install virtualenv
virtualenv your-env
source your-env/bin/activate
pip install numpy
pip install pandas
pip install matplotlib
pip install -r requirements.txt
```



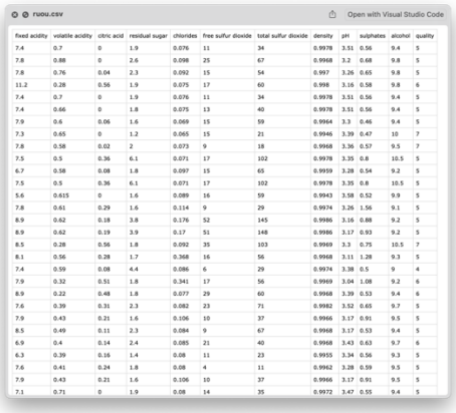
Sau khi tạo env, chọn interpreter



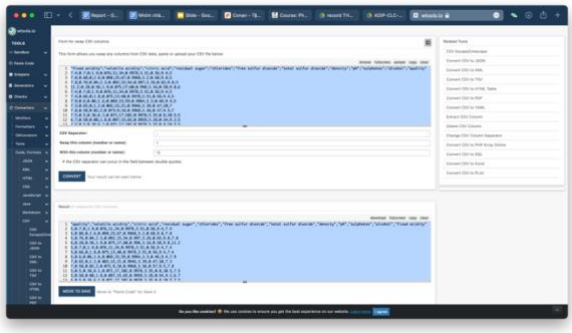
Chọn path your-env/bin/python3.9



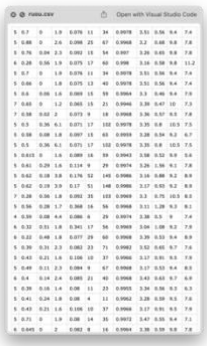
Chạy code có import thư viện và không báo lỗi



File ruou.csv về rượu tìm được ở môn Toán ứng dụng và thống kê

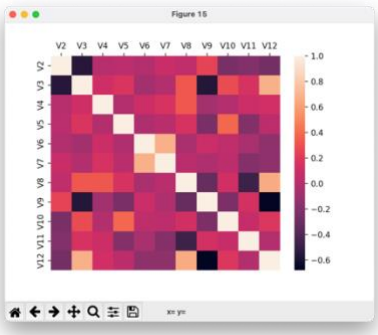


Swap cột đầu và cột cuối tại https://wtools.io/swap-csv-columns

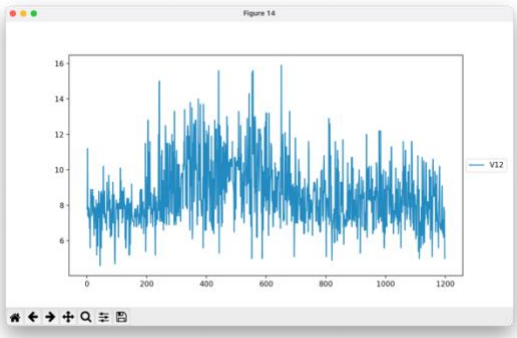


Sau swap, ta xóa dòng đầu tiên rồi sửa code cho khớp với số cột của data.

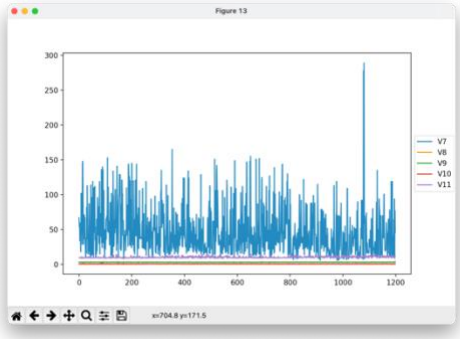
THỰC NGHIỆM



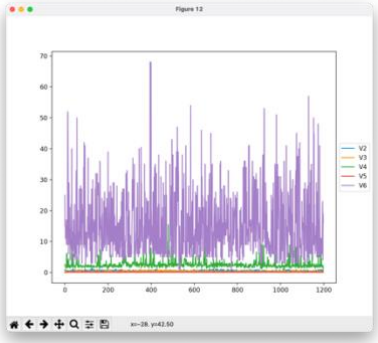
Biểu đồ



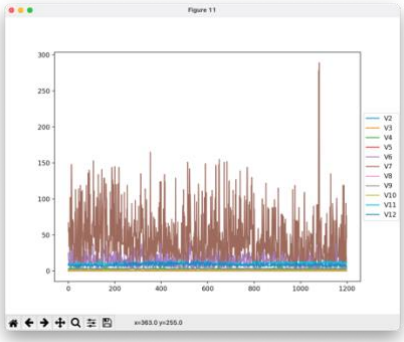
fixed acidity



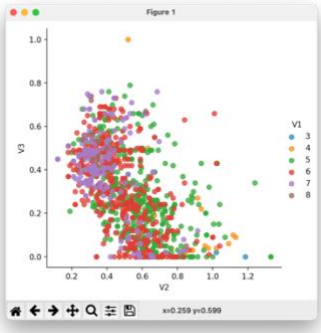
total sulfur dioxide, density, pH, sulphates, alcohol



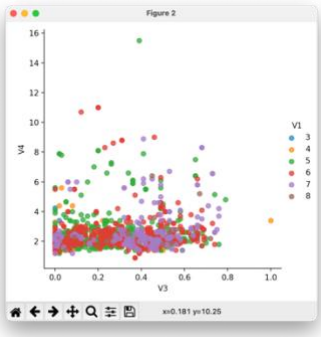
volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide



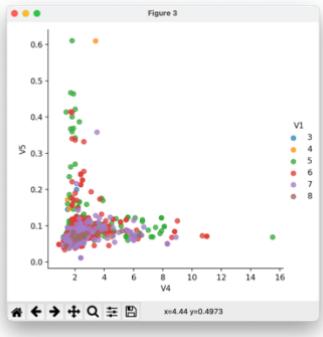
volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, fixed acidity



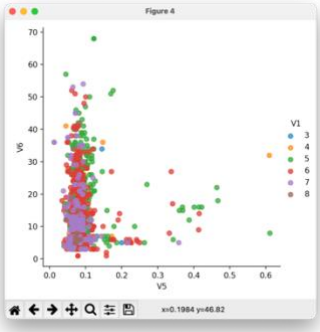
volatile acidity



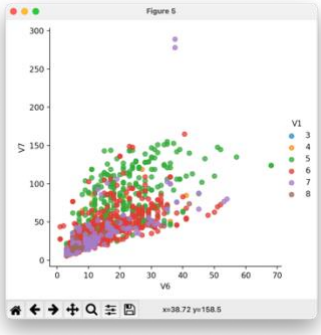
citric acid



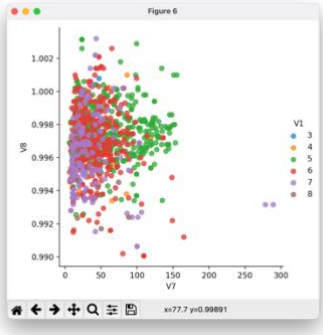
residual sugar



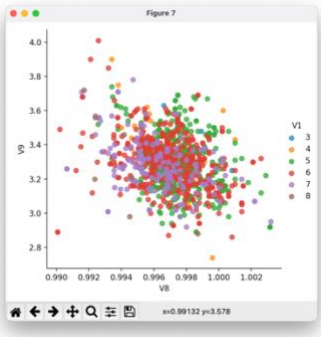
chlorides



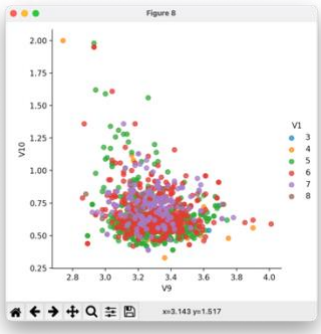
free sulfur dioxide



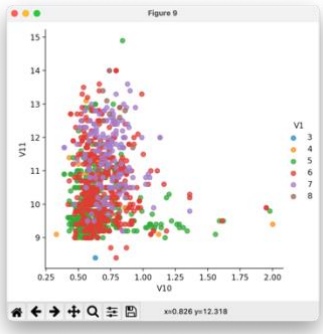
total sulfur dioxide



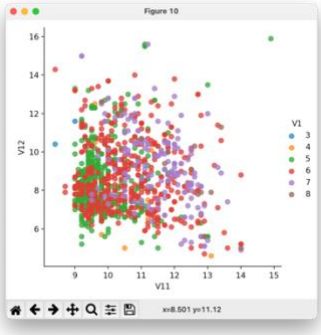
density



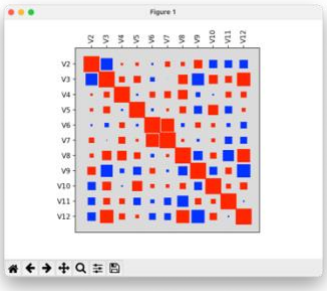
pH



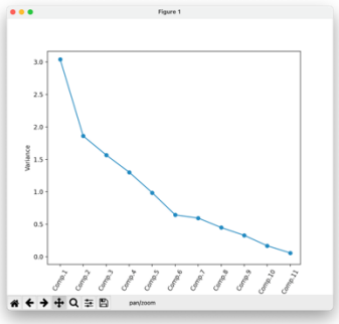
sulphates



alcohol



volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, fixed acidity



quality, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

TỔNG KẾT

Qua bài thực hành này, em được tiếp cận thêm một công cụ Visual Studio Code và file .csv mà đó giờ em không biết xài.

Xung quanh còn rất nhiều đồ án khác (cũng có bài thi cuối kỳ) nên thời gian dành cho đồ án này không đủ để phát triển thêm. Trong phạm vi đồ án, em cũng đã thử các tổng hợp nhiều kiến thức nhất có thể từ bài giảng và tham khảo thêm nhiều tư liệu khác.

Để hoàn thành được bài thực hành này, chúng em xin cảm ơn các giảng viên đã hỗ trợ nhiệt tình, tận tâm hết lòng vì chúng em trong môn học này, kính chúc thầy cô sức khỏe và niềm vui trong công việc giảng dạy tại HCMUS. Lời nói cuối cùng, mong sao ta sẽ được gặp lại.

THAM KHẢO

[1] https://drive.google.com/file/d/1PGRs1yGwrDk4dOX507KImwKs7XBmENvm/view?usp=sharing