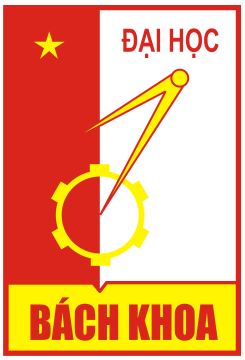
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC
————oOo————



BÁO CÁO PHÂN TÍCH SỐ LIỆU

Đề tài:
Phân tích phân biệt và phân loại

GV hướng dẫn: ThS. Lê Xuân Lý

Mã Học phần: MI4024

Mã Lớp học: 125015

Sinh viên thực hiện: Nhóm 9

Họ và tên	MSSV
Phạm Thúy Vy (Leader)	20173604
Nguyễn Thị Bích Ngọc	20185388
Phan Văn Thanh	20185405
Nguyễn Minh Tuấn	20185420
Hồ Thị Tuyết	20185424
Nguyễn Tiến Vĩ	20185426

HÀ NỘI, 5/2021

Bảng đánh giá công việc

Họ và tên	MSSV	Mức đóng góp
Phạm Thúy Vy (Leader)	20173604	1
Nguyễn Thị Bích Ngọc	20185388	1
Phan Văn Thanh	20185405	2
Nguyễn Minh Tuấn	20185420	2
Hồ Thị Tuyết	20185424	2
Nguyễn Tiến Vĩ	20185426	2

Mục lục

Lời mở đầu

Một trong những nhiệm vụ cơ bản của khoa học để đưa thế giới về trật tự là tiến hành phân loại. Như vậy, cho tập hợp n đối tượng và các quan sát định tính của chúng, ta đòi hỏi tạo chúng thành các nhóm dựa trên tính tương đồng nội tại. Chủ đề phân loại là rất rộng và chúng ta chỉ giới hạn ở những vấn đề mà các nhà thống kê có thể đóng góp. Các nhà phân loại có trong tay các số liệu không thuộc kiểu thống kê: các nhà động vật học ghi lại tiến trình phát triển của loài vật, các nhà ngôn ngữ học có các kiến thức về sự di dời của dân cư ảnh hưởng đến việc phân loại ngôn ngữ nhân loại... Tuy nhiên chúng ta không đề cập trực tiếp đến những ràng buộc xa lạ này trong quá trình phân loại, mặc dù nó cần được đưa vào khi giải thích thực tế. Mục tiêu của chúng ta chỉ đơn thuần là: dựa trên quan sát p dấu hiệu $U_1, U_2, U_3, \ldots$ của một cá thể hoặc một đối tượng, cần phải xác định xem đối tượng hoặc cá thể đó thuộc vào 1 trong k nhóm xác định nào.
Khác với việc phân loại là phân tích phân biệt. Chúng ta biết trước có hai quần thể A và B tồn tại và ta có một mẫu ngẫu nhiên từ đặc trưng nào đó của các cá thể. Ngoài ra ta cũng có một mẫu biết chắc chắn là lầy từ A còn một mẫu khác biết chắc là lấy từ B. Ta muốn đưa ra quy tắc để xếp 1 cá thể, mà ta chưa chắc chắn thuộc loại nào (nhưng biết chắc là thuộc một trong hai loại trên), vào loại A hay loại B. Mong muốn chúng ta là quy tắc này tối ưu theo một nghĩa nào đó chẳng hạn càng ít sai lầm càng tốt hoặc giá (trung bình) phải trả cho các sai lầm là thấp. Sự phân tích phân biệt thường được sử dụng trong trường hợp mất thông tin (chẳng hạn phân loai giới tính mộ cổ); không nhận được đủ thông tin; dự đoán...
Do đó trong bài thuyết trình và báo cáo chúng em trình bày các phần như sau:

- Đặt bài toán
- Quy tắc phân biệt ngẫu nhiên và không ngẫu nhiên
- Bài toán và thực hành Excel
- Phân lớp: Phương pháp khoảng cách và tọa độ
- Phân biệt tuyến tính - LDA
- Mô hình Logistic
- Bài toán thực tế

Phần thuyết trình và báo cáo được nhóm thực hiện và hoàn thành dưới sự hướng dẫn của thầy Lê Xuân Lý. Nhóm 9 chúng em xin chân thành cảm ơn thầy đã giảng dạy và chỉ dẫn chúng em để có thể hoàn thành tốt môn học này cũng như có thêm những kiến thức cần thiết cho bản thân mỗi cá nhân.

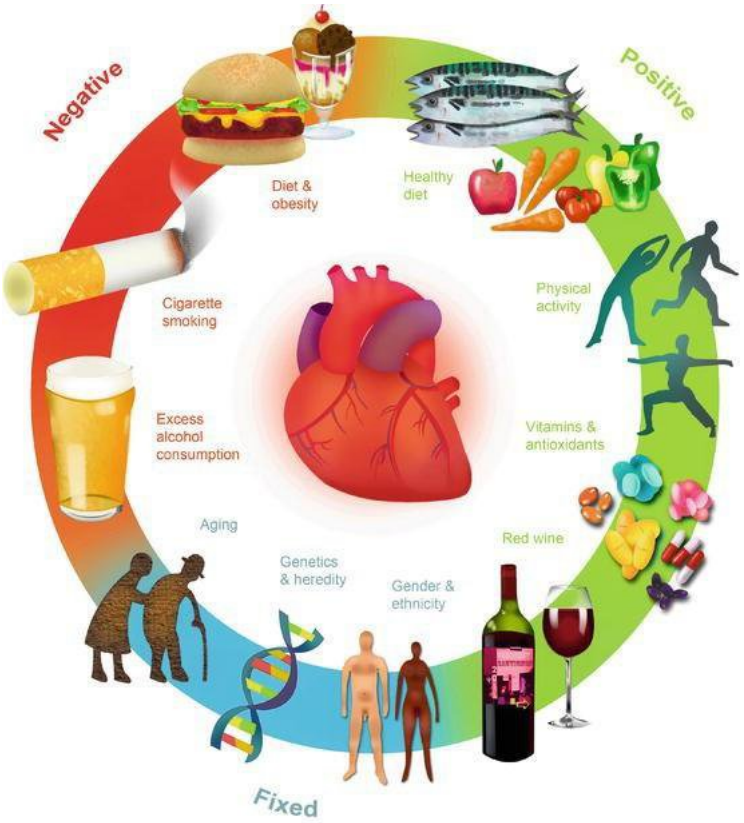
Chương 1

Đặt bài toán

Phân tích phân biệt

- Giả sử có 2 quần thể A và B, một mẫu ngẫu nhiên của các cá thể.
- Bài toán đặt ra: Đưa ra một quy tắc để sắp xếp 2 mẫu ngẫu nhiên chắc chắn thuộc 2 quần thể A và B với mong muốn là quy tắc này tối ưu theo một nghĩa nào đó

Ví dụ thực tế: Dựa trên chỉ số sinh học có thể phân biệt ra nhóm những người có khả năng mắc bệnh hay không?



Hình 1.1: Các nhân tố ảnh hưởng tới việc mắc bệnh tim

Phân loại

Xét bài toán trong đó dựa trên quan sát p dấu hiệu $U_1, U_2, ..., U_p$ của một cá thể (hoặc một đối tượng), cần phải xác định xem cá thể (hoặc đối tượng) đó thuộc vào một trong k nhóm xác định nào.
Ví dụ thực tế: Dựa vào quan sát các dấu hiệu về chiều dài, rộng của đài hoa và cánh hoa Iris từ đó xếp chúng vào 1 trong 3 nhóm : Setosa, Versicolor hay Virginica.

Hình 1.2: Iris flower

Chương 2

Quy tắc phân biệt không ngẫu nhiên và ngẫu nhiên.

2.1 Quy tắc phân biệt

Ký hiệu U là số đo của p dấu hiệu của một đối tượng, S là không gian mẫu (tập hợp tất cả các giá trị có thể có của $U, S \subset$[1])
Trên cơ sở số đo U, ta phải xác định đối tượng quan sát thuộc vào nhóm 1,2,...,hay k.
Nội dung: Chia không gian S thành k miền rời nhau $W_1, W_2, ..., W_k$ và quy tắc phân biệt của ta là: Coi đối tượng thuộc nhóm i nếu $U \in W_i, (i = 1 \div k)$. Điều đó tương đương với việc xác định quy tắc phân biệt $\delta(U) \in \{1, 2, ..., k\}$ với $\delta(U) = \sum_{i=1}^{k} i.I_{W_i}(U)$, trong đó I_A ký hiệu là hàm chỉ tiêu của A.

Xác định k hàm không âm $\lambda_1(U), ..., \lambda_k(U)$ sao cho $\sum_{i=1}^{k} \lambda_i(U) = 1$.
Khi đó quy tắc phân biệt ngẫu nhiên là: coi đối tượng thuộc nhóm i với xác suất $\lambda_i(U)$. Ký hiệu $\lambda = (\lambda_1(U), ..., \lambda_k(U))$ là quy tắc ngẫu nhiên.

Rõ ràng quy tắc không ngẫu nhiên là trường hợp đặc biệt của quy tắc ngẫu nhiên khi $\lambda_i(U)$ chỉ nhận giá trị 0 hoặc 1 (chẳng hạn $\lambda_i(U) = I_{W_i}(U)$).

2.2 Hàm tổn thất

- Giả sử $P_1(u), ..., P_k(u)$ là mật độ xác suất của U (khi U có mật độ), hoặc là xác suất để U nhận giá trị u (khi U rời rạc) khi cá thể thuộc vào nhóm $1, 2, ..., k$.

- Ký hiệu r_{ij} là tổn thất gây ra khi xếp cá thể thuộc nhóm i vào nhóm j. Trước hết ta xét quy tắc phân biệt không ngẫu nhiên xác định bởi quy tắc phân biệt $\delta(U) \in \{1, 2, ..., k\}$ xác định trong phần trước.

- Nếu đối tượng thuộc nhóm i và quy tắc quyết định là $\delta(U)$ thì tổn thất

trung bình sẽ là:

$$L_i^{\delta} = Er_{i\delta(U)} = \sum_{j=1}^{k} r_{ij} P_i\{\delta(U) = j\} = \sum_{j=1}^{k} r_{ij} P_i(W_j)$$
$$L_i^{\delta} = \sum_{j=1}^{k} r_{ij} \int_{W_j} P_i(u)du = r_{i1} \int_{W_1} P_i(u)du + \cdots + r_{ik} \int_{W_k} P_i(u)du \tag{2.1}$$

(Nếu U rời rạc ta thay tích phân bởi tổng).

Nếu quy tắc phân biệt ngẫu nhiên xác định bởi các xác suất $\lambda_1(U), ..., \lambda_k(U)$ thì tổn thất khi ta áp dụng phương pháp đó là:
$\widetilde{L}_i = r_{i1}\lambda_1(U) + \cdots + r_{ik}\lambda_k(U)$
Tổn thất trung bình sẽ là:

$$L_i^{\lambda} = EL_i(U) = \int_S [r_{i1}\lambda_1(U) + \cdots + r_{ik}\lambda_k(U)]P_i(U)du \tag{2.2}$$

Như vậy véctơ tổn thất:

$$(L_1^{\delta}, ..., L_k^{\delta}) \tag{2.3}$$

đặc trưng cho chất lượng của quy tắc quyết định sự lựa chọn. Cho hai quy tắc lựa chọn δ_1, δ_2 với véctơ tổn thất tương ứng là:

$$(L_1^{\delta_1}, ..., L_k^{\delta_1}), (L_1^{\delta_2}, ..., L_k^{\delta_2})$$

Quy tắc δ_1 được gọi là tốt hơn δ_2 nếu:

$$\boxed{L_i^{\delta_1} \leq L_i^{\delta_2}, \forall i = 1, ..., k} \tag{2.4}$$

và với ít nhất một $i : L_i^{\delta_1} < L_i^{\delta_2}$.
Nếu đẳng thức (4) xảy ra với $\forall i = 1 \div k$ thì hai quy tắc δ_1, δ_2 là tương đương theo định nghĩa trên về tính tốt ($\leq$) của các quy tắc quyết định. Ta thấy lớp các quy tắc quyết định là lớp được sắp từng phần.

Quy tắc phân biệt chấp nhận được: Quy tắc phân biệt δ gọi là quy tắc chấp nhận được nếu không tồn tại quy tắc nào khác tốt hơn nó (theo công thức 2.4).

2.3 Quy tắc phân biệt Bayes

Giả sử quan sát một đối tượng sẽ thuộc nhóm i với xác suất tiên nghiệm π_i. Trong trường hợp đó, giá trị trung bình của tổn thất sẽ là:

$$L^{\delta} = \pi_1 L_1^{\delta} + \pi_2 L_2^{\delta} + \cdots + \pi_k L_k^{\delta} \tag{2.5}$$

Đối với quy tắc phân biệt không ngẫu nhiên $d(U)$ ta có:

$$\begin{aligned} L^{\delta} &= \sum_{j=1}^{k} \int_{W_j} [\pi_1 r_{1j} P_1(u) + \cdots + \pi_k r_{kj} P_k(u)] du \\ &= -\int_{W_1} S_1 du - \int_{W_2} S_2 du - \cdots - \int_{W_k} S_k du \end{aligned} \tag{2.6}$$

trong đó:

$$S_j(u) = -(\pi_1 r_{1j} P_1(u) + \cdots + \pi_k r_{kj} P_k(u)) \tag{2.7}$$

được gọi là thông tin phân biệt thứ j
Với quy tắc phân biệt ngẫu nhiên $(\lambda_1(U), ..., \lambda_k(U))$ kỳ vọng toán của tổn thất có dạng:

$$L^{\lambda} = \int_S - \left[\sum_{i=1}^{k} \lambda_i(u) S_i(u) \right] du \tag{2.8}$$

Định lý 2.1: Giả sử $W_1^*, .., W_k^*$ là các miền rời nhau và $\bigcup_{i=1}^{k} W_i^* = S$ sao cho:

$$u \in W_i^* \Leftrightarrow S_i(u) = \max_{1 \le j \le k} S_j(u) \tag{2.9}$$

Khi đó kỳ vọng toán của tổn thất của quy tắc phân biệt không ngẫu nhiên tính theo (2.6),(2.7) là nhỏ nhất khi W_i được thay bởi $W_i^*, i = 1 \div k$.

Định lý 2.2: Đối với quy tắc phân biệt ngẫu nhiên ta đặt:

- $\lambda_i^*(u) = 1, \lambda_j^*(u) = 0, \forall i \neq j$ nếu $S_i(u) > S_j(u), \forall i \neq j$.

- Nếu $S_{i_1}(u) = \cdots = S_{i_r}(u) > S_{i_{r+1}}(u) \geq \cdots \geq S_{i_k}(u)$, ta đặt:
  $\lambda_{i_{r+1}}^* = \cdots = \lambda_{i_k}^* = 0$
  $\lambda_{i_1}^*, ..., \lambda_{i_r}^*$ có thể chọn tùy ý sao cho $\sum_{j=1}^{r} \lambda_{ij}^*(u) = 1$.
  Khi đó: $L^{\lambda^*} \leq L^{\lambda}$

Các quy tắc phân biệt σ^*, λ^* xác định trong định lý (2.1),(2.2) được gọi là các quy tắc quyết định Bayes đối với phân bố tiên nghiệm $\pi = (\pi_1, ..., \pi_k)$.

Chú ý: Trong trường hợp đặc biệt khi $r_{ii} = 0, r_{ij} = 1, \forall i \neq j$, khi đó thông tin phân biệt (2.7) sẽ có dạng:

$$S_i(u) = -\sum_{j=1}^{k} \pi_j P_j + \pi_i P_i = \pi_i P_i + C(u)$$

với $C(u) = -\sum_{j=1}^{k} \pi_j P_j$.
Do đó ta sẽ xếp đối tượng vào nhóm thứ i nếu $\pi_i P_i$ lớn nhất.

2.4 Quy tắc phân biệt khi U có phân bố chuẩn $N_p(\mu, A)$

- Giả sử nếu cá thể thuộc nhóm thứ i thì dấu hiệu $U = (U_1, ..., U_p)$ của nó có phân bố chuẩn $N_p(\mu_i, A_i), i = 1 \div k$. Khi đó:

$$P_i(u) = (2\pi)^{p/2}|A_i|^{-1/2} exp\left\{-\frac{1}{2}(u-\mu_i)^T A_i^{-1}(u-\mu_i)\right\}$$

- Do đó khi $r_{ii} = 0, r_{ij} = 1$ với $i \neq j$ ta có:

$$\widetilde{S}_i(U) = ln(\pi_i P_i(U)) = -\frac{1}{2}ln|A_i| - \frac{1}{2}(U-\mu_i)^T A_i^{-1}(U-\mu_i) + ln\pi_i - \frac{p}{2}ln2\pi \tag{2.10}$$

- Nếu ma trận $A_i \equiv A$ thì mọi $\widetilde{S}_i(U)$ đều chứa một số hạng giống nhau: $-\frac{1}{2}ln|A| - \frac{p}{2}ln2\pi - \frac{1}{2}U^T A^{-1}U$.

- Do đó ta có thể thay thông tin phân biệt $\widetilde{S}_i(U)$:

$$\overline{S}_i(u) = \mu_i^T A^{-1}U - \frac{1}{2}\mu_i^T A^{-1}\mu_i + ln\pi_i, i = 1 \div k \tag{2.11}$$

  đây là hàm tuyến tính đối với U. Hàm trên gọi là **hàm phân biệt tuyến tính**.
  Trong trường hợp này ta sẽ xếp cá thể vào nhóm thứ i nếu $\overline{S}_i(U)$ lớn nhất.

- Ta xét trường hợp khi k=2, là ta cần phải liệt cá thể có dấu hiệu U vào nhóm 1 hoặc 2 Đặt:

$$\begin{aligned}
&\overline{L}(u) = \overline{S}_1(u) - \overline{S}_2(u) \\
&= (\mu_1^T - \mu_2^T)A^{-1}U - [\frac{1}{2}(\mu_1^T A^{-1}\mu_1 - \mu_2^T A^{-1}\mu_2) + ln\pi_2 - ln\pi_1] \\
&= (\mu_1^T - \mu_2^T)A^{-1}U - C
\end{aligned}$$

- khi đó ta có thể thay $\overline{L}(u)$, bởi :

$$L(u) = (\mu_1^T - \mu_2^T)A^{-1}U \tag{2.12}$$

- Ta sẽ liệt kê cá thể vào nhóm 1 khi và chỉ khi:

$$L(u) \geq \mathrm{C}$$

Chương 3

Bài toán và thực hành Excel

3.1 Bài toán 1

Đề bài: Xét việc phân biệt trạng thái thần kinh của một người dựa trên các số đo về 3 dấu hiệu của tâm thần U_1, U_2, U_3. Sau đây là số liệu thống kê dựa trên việc đo 3 dấu hiệu trên 256 người dưới dạng các trung bình mẫu và ma trận hiệp phương sai mẫu. Giả sử quan sát 1 đối tượng có $U = (0.8201, 1.6, 0.68)^T$, hãy xếp đối tượng đó vào một trong các nhóm.

Các nhóm	Cỡ mẫu	$\overline{u_1}$	$\overline{u_2}$	$\overline{u_3}$
1.Tâm thần bất an	114	2.9298	1.667	0.7281
2.Bị điên	33	3.0303	1.2424	0.5455
3.Bệnh thái nhân cách	32	3.8125	1.8438	0.8125
4.Bệnh hoang tưởng	17	4.7059	1.5882	1.1176
5.Thay đổi cá tính	5	1.4000	0.2000	0.0000
6.Trạng thái bình thường	55	0.6000	0.1455	0.2182

Bảng 3.1: Thống kê 3 dấu hiệu trạng thái thần kình trung bình mẫu

Ma trận phương sai mẫu $S = (s_{ij})$				Nghịch đảo của ma trận phương sai mẫu $S^{-1} = (s^{ij})$			
N^o	1	2	3	N^o	1	2	3
1	2.3008	0.2516	0.4742	1	0.5432	-0.2002	-0.4208
2	0.2516	0.6075	0.0358	2	-0.2002	1.7258	0.0558
3	0.4742	0.0358	0.5951	3	-0.4208	0.0558	2.0123

Bảng 3.2: Ma trận và nghịch đảo ma trận hiệp phương sai mẫu

Giải.

Giả sử phân bố xác suất $P_{i(u)}$ của các dấu hiệu U có phân bố chuẩn $N_3(u_i, A)$. Khi đó $\tilde{\mu}_i = \overline{\mu}^{(i)}$; S là ước lượng không chệch của μ_i và A, còn hàm thông tin phân biệt

$$\widehat{S}_i(u) = \overline{u}^{(i)^T} S^{-1} U - \tfrac{1}{2}\overline{u}^{(i)^T} S^{-1} \overline{u}^{(i)} + \ln \pi_i, i = \overline{1,6}$$

Đặt $\overline{u}^T = (\overline{u_1}, \overline{u_2}, \overline{u_3})$, ta có

$$\overline{u}^T S^{-1} = \begin{bmatrix} 2.9298 & 1.667 & 0.7281 \\ 3.0303 & 1.2424 & 0.5455 \\ 3.8125 & 1.8438 & 0.8125 \\ 4.7059 & 1.5882 & 1.1176 \\ 1.4000 & 0.2000 & 0.0000 \\ 0.6000 & 0.1455 & 0.2182 \end{bmatrix} \begin{bmatrix} 0.5432 & -0.2002 & -0.4208 \\ -0.2002 & 1.7258 & 0.0558 \\ -0.4208 & 0.0558 & 2.0123 \end{bmatrix}$$

$$l^T = \overline{u}^T S^{-1} = \begin{bmatrix} 0.9513 & 2.331 & 0.3253 \\ 1.1678 & 1.5678 & -0.1081 \\ 1.3599 & 2.4641 & 0.1336 \\ 1.7680 & 1.8611 & 0.3571 \\ 0.7204 & 0.0649 & -0.5780 \\ 0.2050 & 0.1431 & 0.1947 \end{bmatrix}$$

$$\tfrac{1}{2}\overline{u}^{(1)^T} S^{-1} \overline{u}^{(1)} = l^{(1)^T} \overline{u}^{(1)} = \tfrac{1}{2} \begin{bmatrix} 0.9513 & 2.331 & 0.3253 \end{bmatrix} \begin{bmatrix} 2.9298 \\ 1.667 \\ 0.7281 \end{bmatrix} =$$

3.4549

Tương tự ta có:

$$\tfrac{1}{2}\overline{u}^{(i)^T} S^{-1} \overline{u}^{(i)} = \begin{bmatrix} 3.4549 \\ 2.7139 \\ 4.9182 \\ 5.8375 \\ 0.5107 \\ 0.0931 \end{bmatrix}$$

$$\ln \pi_i = \tfrac{n_i}{N} = \ln \begin{bmatrix} 0.4453 \\ 0.1289 \\ 0.1250 \\ 0.0664 \\ 0.0195 \\ 0.2148 \end{bmatrix} = \begin{bmatrix} -0.8090 \\ -2.0487 \\ -2.0794 \\ -2.7120 \\ -3.9357 \\ -1.5378 \end{bmatrix}$$

Lập được mô hình

$$\widehat{S}(U) = \begin{bmatrix} 0.9513 & 2.331 & 0.3253 \\ 1.1678 & 1.5678 & -0.1081 \\ 1.3599 & 2.4641 & 0.1336 \\ 1.7680 & 1.8611 & 0.3571 \\ 0.7204 & 0.0649 & -0.5780 \\ 0.2050 & 0.1431 & 0.1947 \end{bmatrix} U - \begin{bmatrix} 3.4549 \\ 2.7139 \\ 4.9182 \\ 5.8375 \\ 0.5107 \\ 0.0931 \end{bmatrix} + \begin{bmatrix} -0.8090 \\ -2.0487 \\ -2.0794 \\ -2.7120 \\ -3.9357 \\ -1.5378 \end{bmatrix}$$

$\Rightarrow$ Ta sẽ liệt kê cá thể vào nhóm thứ i nếu $\widehat{S}_i(U)$ là lớn nhất.

Với $U = (0.8201, 1.6, 0.68)^T$

Khi đó áp dụng mô hình trên $\widehat{S}(U) = \begin{bmatrix} 0.4671 \\ -1.3699 \\ -1.8490 \\ -3.8790 \\ -4.1449 \\ -1.1016 \end{bmatrix}$

Như vậy$\widehat{S_1} = \underset{1 \le i \le 6}{max}\ \widehat{S_i}$,cần xếp cá thể vào nhóm I: tâm thần phân liệt.

3.2 Thực hành Excel

3.2.1 Bài toán 1

Các câu lệnh sử dụng: MMULT(), TRANSPOSE(),LN()
Ta thực hiện lập mô hình dự đoán dựa trên công thức

$$\widehat{S_i}(u) = \overline{u}^{(i)^T} S^{-1} U - \frac{1}{2} \overline{u}^{(i)^T} S^{-1} \overline{u}^{(i)} + \ln \pi_i, i = \overline{1,6}$$

bằng các lệnh Excel đã liệt kê ở trên.
Tìm $\overline{u}^{(i)^T} S^{-1}$ bằng lệnh MMULT(C11:E16,M3:O5), trong đó C11:E16,M3:O5 là vị trí ma trận $\overline{u}^{(i)^T}$ và S^{-1} trong file Excel

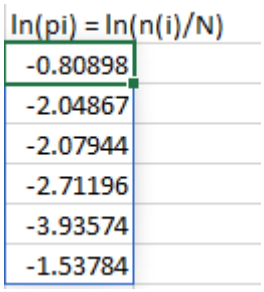
L^T = u^T*S^-1		
0.951349	2.330991	0.325314
1.167784	1.567907	-0.10811
1.359921	2.464105	0.133578
1.768001	1.861156	0.357325
0.72044	0.06488	-0.57796
0.204972	0.143159	0.194723

Hình 3.1: $\overline{u}^{(i)^T} S^{-1}$

$\frac{1}{2} \overline{u}^{(i)^T} S^{-1} \overline{u}^{(i)}$ bằng MMULT(Gi:Ii,TRANSPOSE(Ci:Ei))/2, $i = \overline{11,16}$, trong đó Gi:Ii, Ci:Ei ứng với $\overline{u}^{(i)^T} S^{-1}$ và $\overline{u}^{(i)}$ trong file Excel

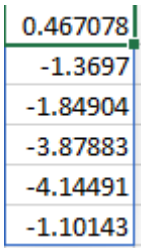
(1/2)*u(i)^T*S^-1*u(i)
3.454943
2.713863
4.918274
5.837636
0.510796
0.093151

Hình 3.2: $\frac{1}{2} \overline{u}^{(i)^T} S^{-1} \overline{u}^{(i)}$



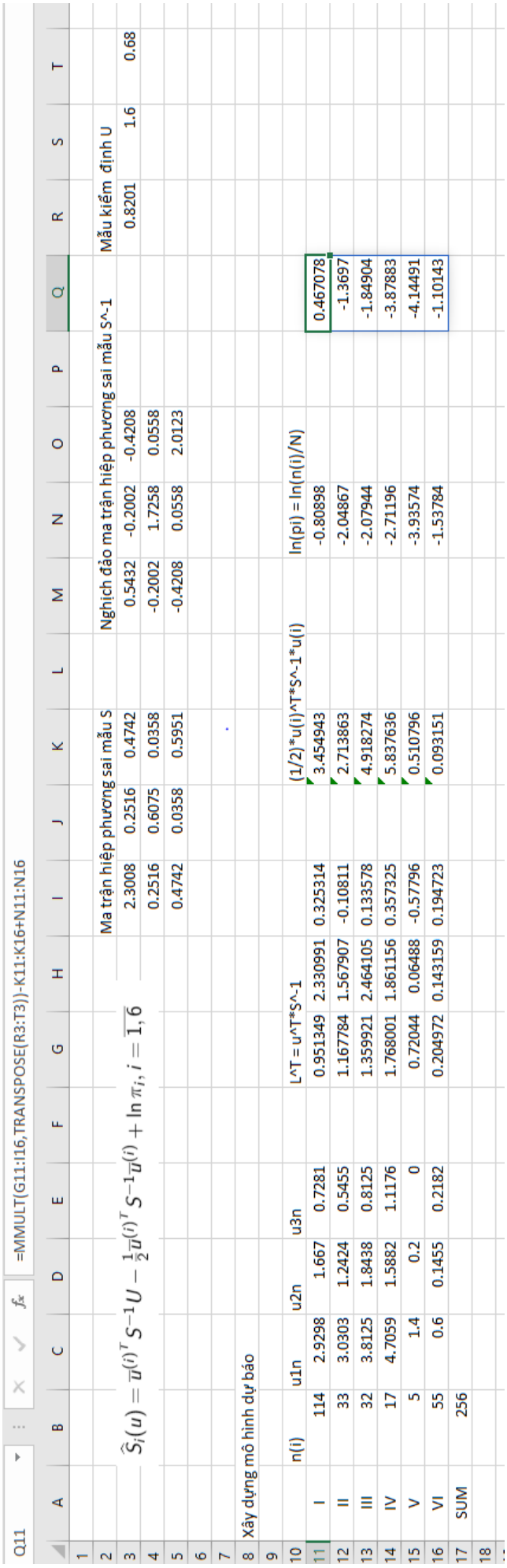
Hình 3.3: $\ln \pi_i$

$\ln \pi_i$ bằng LN(B11:B16/B17), trong đó B11:B16 lần lượt là cỡ mẫu n_i của các nhóm, B17 là tổng cỡ mẫu trong file Excel
Sau khi lập được mô hình ta tiến hành thử với mẫu kiểm định bằng lệnh ,MMULT(G11:I16,TRANSPOSE(R3:T3))-K11:K16+N11:N16, với R3:T3 là vị trí mẫu kiểm định và G11:I16, K11:K16, N11:N16 lần lượt là vị trí của các giá trị của mô hình trong file Excel ta được kết quả



Hình 3.4: Giá trị kiểm định

Ta thấy phần tử thứ nhất là lớn nhất vì vậy sẽ xếp mẫu vừa kiểm định vào nhóm I: Tâm thần bất an.
Dưới đây là toàn bộ kết quả thực hiện trên Excel



11

Hình 3.5: Kết quả thực hành bài toán 1

3.2.2 Bài toán 2: Iris dataset

Iris dataset là bộ dữ liệu gồm 150 các thể về các chỉ số chiều dài, rộng của đài hoa và cách hoa Iris để phân biệt chúng vào 3 nhóm Sestosa , Versicolor, Virginica.
Với bộ giữ liệu này chúng em sẽ sử dụng $\frac{7}{10}$ ứng với 105 cá thể hoa để xây dụng mô hình và$\frac{3}{10}$ ứng với 45 để kiểm định.
Các câu lệnh sử dụng :
SUM(),COUNT(),COUNTIF(),MMULT(),TRANSPOSE(),MINVERSE(),IF()
Bước đầu tiên chúng ta sẽ tìm trung bình mẫu cho các dấu hiệu nhận biết của từng nhóm và trung bình mẫu toàn bộ dữ liệu xây dựng mô hình bằng lệnh SUM()/COUNT()

	I	II	III	IV	ni
Nhóm 1	5.017647	3.461765	1.485294	0.264705882	34
Nhóm 2	5.876471	2.797059	4.241176	1.332352941	34
Nhóm 3	6.605405	2.959459	5.581081	2.010810811	37
mẫu	5.855238	3.069524	3.820952	1.225714286	105

Hình 3.6: Trung bình mẫu

Tiếp theo chúng ta sẽ tính sai số của dữ liễu so với trung bình mẫu từ đó tính ma trận hiệp phương sai mẫu bằng lệnh MMULT()/COUNT() và nghịch đảo của ma trận hiệp phương sai mẫu bằng MINVERSE()

B3: Tính ma trận hiệp phương sai mẫu MMULT()/COUNT()			
(SSE^T*SSE)/(N-1)			
0.729034799	-0.03436	1.311716	0.507508
-0.03435806	0.195985	-0.33993	-0.12854
1.311716117	-0.33993	3.154172	1.279937
0.507508242	-0.12854	1.279937	0.559236
và nghịch đảo của ma trận hiệp phương sai mẫu MINVERSE()			
11.48449496	-8.24859	-9.35256	9.087332
-8.24859361	12.28861	8.012821	-8.02907
-9.35255566	8.012821	12.32944	-17.8895
9.087331951	-8.02907	-17.8895	32.64013

Hình 3.7: Ma trận hiệp phương sai và nghịch đảo ma trận hiệp phương sai

Sau đó thực hiện tương tự như Bài toán 1 trên Excel ta sẽ được mô hình

B4: Lặp lại các bước ở bài toán 1			
L^T			
17.58463	10.9278	-5.61194	-0.12892
16.85819	9.185574	-4.09152	-1.44082
17.5239	10.45759	-5.22465	2.054021
(1/2)*u(i)^T*S^-1*u(i)			
58.8467			
52.74336			
60.83617			
Ln(pi(i))			
-1.1276			
-1.1276			
-1.04304			

Hình 3.8: Mô hình Iris

Thực hiện kiểm định với mẫu kiểm định ta được kết quả và so sánh với phân loại ban đầu của dữ liệu ta thấy có $\frac{6}{45}$ dự đoán bị sai. Tý số $\frac{6}{45}$ cho ta thông tin về chất lượng của mô hình phân biệt được xây dựng.
Chú thích: Vì phần dữ liệu kiểm định khá lớn nên chúng em không thể chụp kết quả vào hết trong 1 ảnh nên thầy có thể xem ở trong file Excel đính kèm ạ.

Chương 4

Phân lớp: Phương pháp khoảng cách và tọa độ

Việc phân lớp dựa trên hiểu biết về bản chất của các mối quan hệ xác định bởi nhiều biến mô tả trạng thái của các đối tượng và sự vật.
Kỹ thuật sẽ được sử dụng trong phần này dựa trên việc tính khoảng cách mô tả sự gần nhau của các đối tượng và ghép dàn các đối tượng thành các nhóm các đối tượng "gần nhau".

4.1 Các độ đo về sự gần nhau của các phần tử

Xét 2 đối tượng có véctơ trạng thái là $x = (x_1, ..., x_k)$ và $y = (y_1, ..., y_k)$. Sau đây là các khoảng cách thường dùng để đo sự "gần nhau" của các đối tượng:
Khoảng cách Euclide:

$$d_1^2(x,y) = \sum_{i=1}^{k} (x_i - y_i)^2 = (x-y)(x-y)^T \tag{4.1}$$

Khoảng cách thống kê:

$$d_2^2(x,y) = (x-y)A(x-y)^T \tag{4.2}$$

trong đó A là ma trận đối xứng xác định dương

Khoảng cách Minkowski:

$$d_3(x,y) = \left(\sum_{i=1}^{k} |x_i - y_i|^m \right)^{1/m}, m = 1,2,3,... \tag{4.3}$$

Khoảng cách Canberra:

$$d_4(x,y) = \sum_{i=1}^{k} \frac{|x_i - y_i|}{x_i + y_i} \tag{4.4}$$

(Chỉ xác định cho các $x_i, y_i > 0$)
Khoảng cách Czkanowski:

$$d_5(x,y) = 1 - 2\sum_{i=1}^{k} min(x_i, y_i) / \sum_{i=1}^{k} (x_i + y_i) \tag{4.5}$$

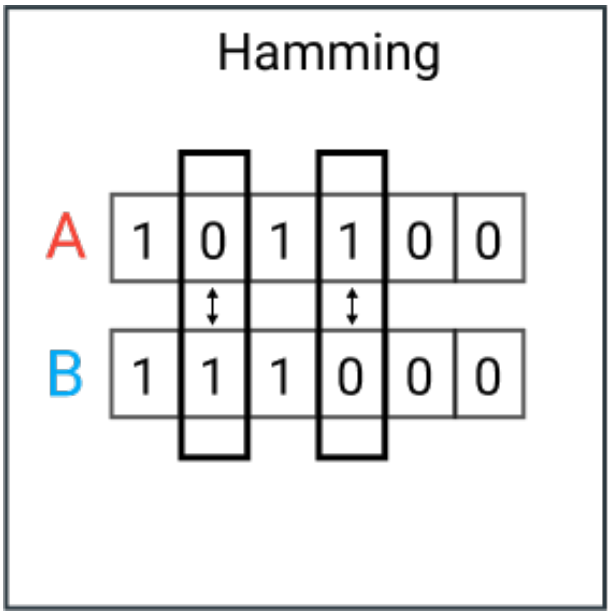
(Chỉ xác định cho các $x_i, y_i > 0$)

Khoảng cách Hamming:

Khoảng cách Hamming là số các giá trị khác nhau giữa hai vector, thường được sử dụng để so sánh các chuỗi dạng nhị phân.

Chú thích 4.1:

Hệ số Czkanowski không phải một khoảng cách thông thường.



Hình 4.1: Khoảng cách Hamming

Chú thích 4.2: Với các biến nhị phân x_i, y_i chỉ nhận 2 giá trị 1 và 0 thì khoảng cách $d_1(x,y) = m =$ số các cặp (x_i, y_i) sao cho $x_i \neq y_i$.

Trong trường hợp này ta đã không tính đến tầm quan trọng của các cặp 1-1 và 0-0. Đối với các biến nhị phân hoặc các biến định tính khác, người ta thường tính đến trọng số của các cặp 1-0; 0-0 và thay vì xét $(x_i - y_i)^2$ ta xét các hệ số sau:

Xét 2 đối tượng có véctơ trạng thái $x = (x_1, ..., x_k)$ và $y = (y_1, ..., y_k), x_i, y_i \in \{0; 1\}$. Bảng sau đây thể hiện số các cặp (x_i, y_i) nhận giá trị 1-1, 1-0, 0-1, 0-0.

	Phần tử thứ nhất 1	 0	Tổng số
Phần tử thứ hai: 1	a	b	$a+b$
0	c	d	$c+d$
Tổng số	$a+c$	$b+d$	$k=a+b+c+d$

$m = c + b =$ số các cặp 1-0, 0-1.

Khi đó thay vì các khoảng cách $d_1 - d_5$ người ta xét các hệ số đo sự gần nhau giữa 2 phần tử như sau:

	Hệ số	Ý nghĩa
1	$\frac{a+d}{k}$	cặp 1-1 và 0-0 có trọng số như nhau
2	$\frac{2(a+d)}{2(a+d)+b+c}$	cặp 1-1,0-0 có trọng số gấp đôi
3	a/k	Tỷ lệ cặp 1-1 trên tổng số các cặp
4	$\frac{a+d}{a+d+2(b+c)}$	cặp 1-0,0-1 có trọng số gấp đôi
5	$\frac{a}{a+b+c}$	tỷ lệ các cặp 1-1 trên tổng số không có cặp 0-0
6	$\frac{2a}{a+b+c}$	trọng số gấp đôi cho các cặp 1-1 không tính đến cặp 0-0
7	$\frac{a}{a+2(b+c)}$	trọng số gấp đôi cho cặp 1-0,0-1 không tính đến cặp 0-0
8	$\frac{a}{b+c}$	tỷ lệ các cặp 1-1 trên tổng số các cặp 1-0,0-1

Chú thích 4.3: Ta có thể xây dựng các hệ số đo sự gần nhau hoặc tương tự với nhau từ các khoảng cách $d(x,y)$, chẳng hạn hệ số:

$$e(x,y) = \frac{1}{1+d(x,y)} \; ; 0 < e \leq 1$$

Các giá trị tương tự càng lớn thì các phần tử càng gần nhau hoặc tương tự với nhau.

Chú thích 4.4: Nếu các thành phần của véctơ trạng thái x, y có một số biến định lượng, một số biến định tính, ta có thể sử dụng tổng các hệ số tương tự cho các thành phần định lượng và các thành phần định tính. Nếu không ta có thể đưa các biến định tính về biến định lượng để tính.

Ví dụ 1:

Đề bài:Tính giá trị của các hệ số tương tự cho các cặp của các phần tử có véctơ trạng thái cho trong bảng sau:

Phần tử	Chiều cao (inch)	Trọng lượng (pound)	Màu mắt	Màu tóc	Thuận tay	Giới tính
1	68	140	xanh	vàng	phải	nữ
2	73	185	nâu	nâu	phải	nam
3	67	165	xanh nước biển	vàng	phải	nam
4	64	120	nâu	nâu	phải	nữ
5	76	210	nâu	nâu	trái	nam

1 inch = 2,54 cm ; 1 pound = 453,584 g.

Giải.

$$X_1 = \begin{cases} 1 \text{ nếu chiều cao} \geq 72inch \\ 0 \text{ nếu chiều cao} < 72inch \end{cases} \quad X_2 = \begin{cases} 1 \text{ nếu trọng lượng} \geq 150pound \\ 0 \text{ nếu trọng lượng} < 150pound \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ nếu màu mắt nâu} \\ 0 \text{ nếu ngược lại} \end{cases} \quad X_4 = \begin{cases} 1 \text{ nếu tóc màu vàng} \\ 0 \text{ nếu ngược lại} \end{cases}$$

$$X_5 = \begin{cases} 1 \text{ nếu thuận tay phải} \\ 0 \text{ nếu ngược lại} \end{cases} \quad X_6 = \begin{cases} 1 \text{ nếu là nam} \\ 0 \text{ nếu là nữ} \end{cases}$$

Khi đó véctơ trạng thái của 5 cá thể trên là:

Cá thể	X_1	X_2	X_3	X_4	X_5	X_6
1	0	0	0	1	1	1
2	1	1	1	0	1	0
3	0	1	0	1	1	0
4	0	0	1	0	1	1
5	1	1	1	0	0	0

Khi đó hệ số tương tự $e_{ij}, i, j = 1 \div 5$ giữa các cặp cá thể được cho trong bảng dưới đây nếu ta sử dụng hệ số $e = \frac{a+d}{k} = \frac{a+d}{6}$, trong đó $a+d$ là số các cặp (x_i, y_i) là 1-1, 0-0.
Ta có bảng

Cá thể	1	2	3	4	5
1	1				
2	1/6	1			
3	4/6	3/6	1		
4	4/6	3/6	2/6	1	
5	0	5/6	2/6	2/6	1

Ta có $e_{25} = 5/6$ là lớn nhất. Vậy hai phần tử 1 và 2 gần nhau nhất. $e_{15} = 0$ là bé nhất nên phân tử 1 và 5 ít gần nhau nhất. Từ đó nếu phân làm 2 lớp thì $\{2;5\}, \{1,3,4\}$ sẽ là hai lớp có phần tử tương tự nhau.

Các độ đo khác trong thực tế:

Độ tương đồng cosin (Cosin Similarity:)

$$d(x,y) = cos(\theta) = \frac{x \cdot y}{||x||\,||y||} \tag{4.6}$$

Khoảng cách Manhattan:

$$d(x,y) = \sum_{i=1}^{k} |x_i - y_i| \tag{4.7}$$

Khoảng cách Chebyshev:

$$d(x,y) = max_i(|x_i - y_i|) \tag{4.8}$$

Khoảng cách Haversine:

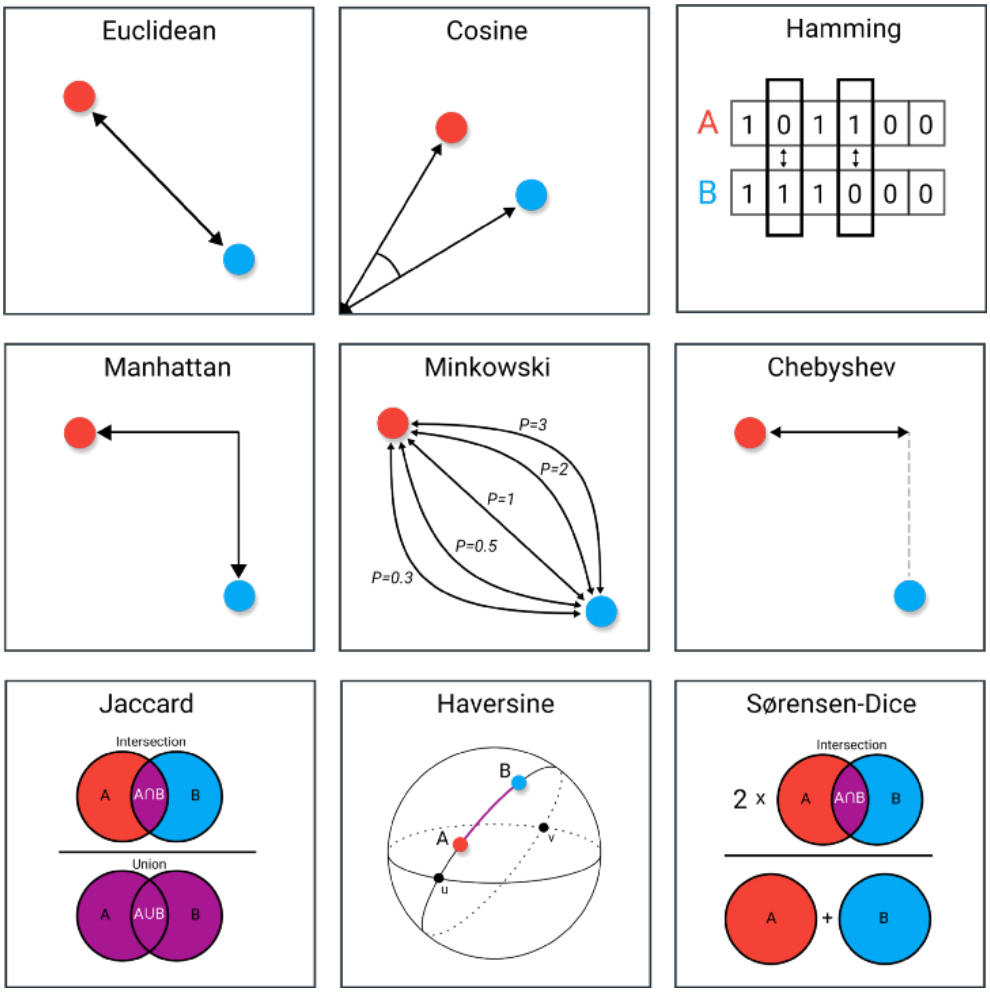
$$d = 2\arcsin\sqrt{\sin^2\frac{\varphi_2 - \varphi_1}{2} + \cos\varphi_1 \cos\varphi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2}} \tag{4.9}$$

Chỉ mục Jaccard:

$$d(x,y) = 1 - \frac{x \cap y}{x \cup y} \tag{4.10}$$

Chỉ mục Sørensen-Dice:

$$d(x,y) = \frac{2|x \cup y|}{|x| + |y|} \tag{4.11}$$



Hình 4.2: Hình ảnh minh họa các khoảng cách

4.2 Phương pháp phân lớp theo thứ bậc

Ta khó có thể kiểm tra tất cả các cách một tổng thể gồm n phần tử thành tất cả các lớp có thể khi n lớn kể cả với máy tính tốt. Vì vậy ta cố gắng phân lớp một cách hợp lý.

Đầu tiên ta chia tất cả các phần tử thành 2 lớp sao cho mỗi phần tử của lớp này cách xa các phần tử của lớp kia. Sau đó mỗi lớp con lại phân thành 2 lớp con theo quy tắc trên cho đến khi mỗi lớp con chỉ còn 1 phần tử. Phương pháp phân lớp như vậy gọi là **phương pháp phân chia lớp theo thứ bậc**.

Một kỹ thuật khác được gọi là **phương pháp gộp theo thứ bậc** được thực hiện như sau: Các cá thể gần nhau hoặc tương tự với nhau nhất được ghép với nhau thành một nhóm. Sau đó các nhóm ban đầu đó lại được ghép với nhau thành các nhóm lớn hơn tương ứng với các khoảng cách bé nhất giữa các nhóm. Tiếp tục quá trình cho đến khi chỉ còn một nhóm duy nhất.

Ta sẽ tập trung nghiên cứu phương pháp gộp theo thứ bậc và cụ thể hơn là nghiên cứu phương pháp kết nối. Phương pháp này rất thích hợp cho việc phân lớp các đối tượng.

Chúng ta sẽ nghiên cứu một số phương pháp sau để xác định khoảng cách giữa các lớp:

1. phương pháp kết nối đơn: dựa trên khoảng cách ngắn nhất hoặc tương tự nhau nhất)
2. phương pháp kết nối đầy đủ: dựa trên khoảng cách lớn nhất hoặc ít tương tự nhất
3. phương pháp kết nối trung bình: dựa trên khoảng cách trung bình

Các bước phân lớp theo thứ bậc kết nối một tập gồm N phần tử:

1. Bắt đầu với N cụm, mỗi cụm chứa 1 phần tử và lập ma trận các khoảng cách cấp N là $D = \{d_{ik}\}$.
2. Tìm một ma trận khoảng cách của các cặp(các cụm) gần nhất. Giả sử khoảng cách giữa 2 cụm gần nhất U, V là d_{UV}.
3. Gộp cụm U với V. Ký hiệu cụm mới là (UV). Lập các phần tử của ma trận khoảng cách mới bằng cách:
   (a) loại các hàng và các cột tương ứng với cumg U,V
   (b) thêm vào một hàng và một cột gồm các khoảng cách từ cụm (UV) đến các cụm còn lại.
4. Lặp lại bước 2-3 $(N-1)$ lần, tất cả các phần tử sẽ tạo thành một cụm duy nhất sau khi kết thúc thuật toán. Ghi lại sự nhận dạng của các cụm đã được kết hợp và mức độ(khoảng cách hoặc sự tương tự) mà ở đó việc kết hợp các cụm đã được thực hiện.

Ví dụ 2 (Phân cụm theo kết nối đơn)
Đề bài:Xét ma trận khoảng cách của 5 cá thể:

$$D = [d_{ik}] = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1| & 0 & & & & \\ 2| & 9 & 0 & & & \\ 3| & 3 & 7 & 0 & & \\ 4| & 6 & 5 & 9 & 0 & \\ 5| & 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

Hãy phân cụm 5 cá thể trên theo kết nối đơn (khoảng cách các cụm gần nhất).

Giải.

Bước 1:

- Ta có $min[d_{ik}] = d_{35} = 2$. Vậy kết hợp 3 và 5 thành một cụm (35).
- Ta tính các khoảng cách từ cụm (35) đến các phần tử còn lại 1,2,4:
  $d_{(35),1} = min(d_{3;1}, d_{5;1}) = 3$;
  $d_{(35),2} = min(d_{3;2}, d_{5;2}) = 7$;
  $d_{(35),4} = min(d_{3;4}, d_{5;4}) = 8$;
- Xóa đi các dòng và các cột 3 và 5 tương ứng với các phần tử thứ 3 và 5 và thay bởi một hàng và một cột các khoảng cách $d_{(35),1}, d_{(35),2}, d_{(35),4}$ ta được ma trận mới:

$$\begin{bmatrix} & (35) & 1 & 2 & 4 \\ (35)| & 0 & & & \\ 1| & 3 & 0 & & \\ 2| & 7 & 9 & 0 & \\ 4| & 8 & 6 & 5 & 0 \end{bmatrix}$$

Bước 2:

- Khoảng cách ngắn nhất trong ma trận trên là $d_{(3;5),1} = 3$. Vậy ta ghép $1;3;5$ thành cụm (135).
- Tiếp đó ta tính:

$$d_{(135),2} = min(d_{(35);2}, d_{1;2}) = min(7;9) = 7$$
$$d_{(135),4} = min(d_{(35);4}, d_{1;4}) = min(8;6) = 6$$

- Bỏ hàng và cột có các chỉ số (35) và 1, sau đó thêm vào hàng và cột với chỉ số (135) ta có:

$$\begin{bmatrix} & (135) & 2 & 4 \\ (135)| & 0 & & \\ 2| & 7 & 0 & \\ 4| & 6 & 5 & 0 \end{bmatrix}$$

Bước 3:

- Khoảng cách ngắn nhất trong ma trận trên là $d_{2,4} = 5$. Vậy ta ghép $2;4$ thành cụm (24).
- Tiếp đó ta tính:

$$d_{(24),(135)} = min(d_{2;(135)}, d_{4;(135)}) = min(6;7) = 6$$

- Bỏ hàng và cột có các chỉ số 2 và 4, sau đó thêm vào hàng và cột với chỉ số (24) ta có:

$$\begin{bmatrix} & (135) & (24) \\ (135)| & 0 & \\ (24)| & 6 & 0 \end{bmatrix}$$

Bước 4:
Cuối cùng kết hợp cụm (135) và (24) thành một cụm duy nhất (12345).

Chương 5

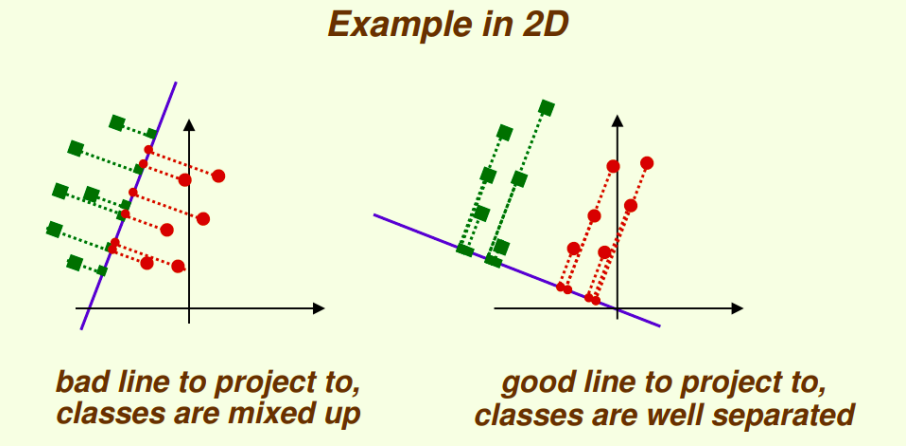
Phân biệt tuyến tính-Linear Discriminant Analysis(LDA)

Giảm chiều dữ liệu (Dimensionality Reduction) là một trong những kỹ thuật quan trọng trong phân tích dữ liệu. Các vector trạng thái trong các bài toán thực tế có thể có số chiều rất lớn, lên tới vài nghìn. Ngoài ra, số lượng các điểm dữ liệu cũng thường rất lớn.
$\Rightarrow$Khó khăn khi thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu.

Phân tích thành phần chính (PCA) là phương pháp tìm ra một hệ cơ sở mới sao cho thông tin của dữ liệu chủ yếu tập trung ở một vài toạ độ, phần còn lại chỉ mang một lượng nhỏ thông tin.

Nói cách khác, lượng dữ liệu sau khi giảm số chiều được giữ lại nhiều nhất có thể.

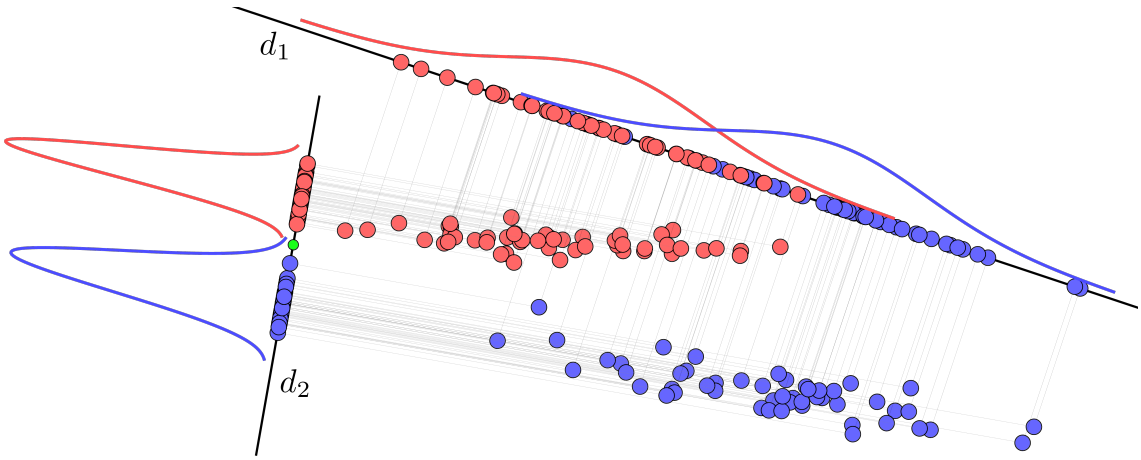
5.1 Ý tưởng



Tuy nhiên, trong nhiều trường hợp, ta không cần giữ lại lượng thông tin lớn nhất mà chỉ cần giữ lại thông tin cần thiết cho riêng bài toán.

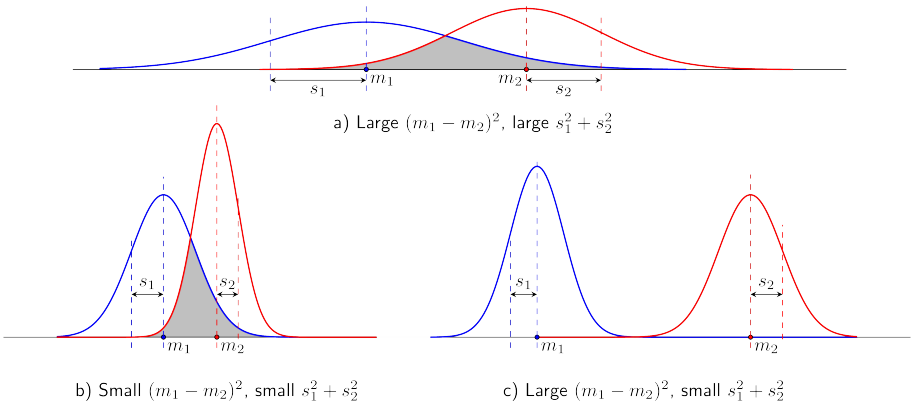
Phân biệt tuyến tính là một phương pháp giảm chiều dữ liệu sao cho việc phân lớp hiệu quả nhất.

Số chiều của dữ liệu mới $\leqslant C-1$ trong đó C là số lượng lớp.



Hình 5.1: Hình ảnh minh họa so sánh phép chiếu dữ liệu trong PCA và LDA

5.2 Bài toán phân loại 2 lớp



Hình 5.2: Khoảng cách giữa các kỳ vọng và tổng các phương sai ảnh hưởng tới độ discriminant của dữ liệu. a) Khoảng cách giữa hai kỳ vọng là lớn nhưng phương sai trong mỗi lớp cũng lớn, khiến cho hai phân phối chồng lấn lên nhau (phần màu xám). b) Phương sai cho mỗi lớp là rất nhỏ nhưng hai kỳ vọng quá gần nhau, khiến khó phân biệt 2 lớp. c) Khi phương sai đủ nhỏ và khoảng cách giữa hai kỳ vọng đủ lớn, ta thấy rằng dữ liệu discriminant hơn.

Nhận xét:

Hai lớp được gọi là phân biệt nếu hai lớp đó cách xa nhau (khoảng cách giữa hai kì vọng lớn) và dữ liệu trong mỗi lớp có xu hướng giống nhau (độ lệch chuẩn nhỏ).

Như vậy, phân tích tuyến tính là thuật toán đi tìm một phép chiếu sao cho tỉ lệ giữa khoảng cách kì vọng và tổng phương sai lớn nhất có thể.

Xây dựng hàm mục tiêu:

Giả sử rằng có N điểm dữ liệu $x_1, x_2, ..., x_n$ trong đó $N_1 < N$ điểm đầu tiên thuộc lớp 1, $N_2 = N - N_1$ điểm cuối cùng thuộc lớp 2. Phép chiếu dữ liệu xuống một đường thẳng có thể được mô tả bằng một vecto hệ số w, giá trị

tương ứng của dữ liệu được cho bởi:

$$y_n = w^T x_n \quad 1 \leq n \leq N$$

Between-class Variances:
Khoảng cách giữa hai kì vọng của 2 lớp sau phép chiếu được tính như sau:

$$m_1 - m_2 = \frac{1}{N_1} \sum_{i \in C_1} y_i - \frac{1}{N_2} \sum_{j \in C_1} y_j = w^T (m_1^{'} - m_2^{'})$$

với $m_1^{'}, m_2^{'}$ lần lượt là kì vọng của lớp 1, lớp 2.
Bình phương khoảng cách giữa hai kỳ vọng $(m_1 - m_2)^2$ được gọi là between-class variance.
Within-class Variances:
Khác với các bài trước, phương sai ở đây không được lấy trung bình như phương sai thông thường vì phương sai ở đây nên tỉ lệ thuận với số lượng điểm dữ liệu trong lớp đó. Ta gọi đó là within-class variances:

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

Hàm mục tiêu:
Phân biệt tuyến tính là thuật toán đi tìm giá trị lớn nhất của hàm mục tiêu:

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

Xét tử số:

$$(m_1 - m_2)^2 = w^T (m_1^{'} - m_2^{'})(m_1^{'} - m_2^{'})^T w = w^T S_B w$$

trong đó S_B được gọi là between-class covariance matrix.
Xét mẫu số:

$$s_1^2 + s_2^2 = \sum_{k=1}^{2} \sum_{n \in C_k} (w^T (x_n - m_k^{'}))^2 = w^T S_w w$$

trong đó S_w được gọi là within-class covariance matrix.
Từ đó, ta có thể biểu diễn hàm mục tiêu phụ thuộc vào w như sau:

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

Áp dụng các phương pháp tính cho đạo hàm nhiều biến, bài toán tối ưu cho LDA trở thành:

$$w = \alpha S_W^{-1} (m_1^{'} - m_2^{'})$$

với $\alpha \neq 0$ bất kỳ.

Biểu thức trên còn được gọi là Fisher's linear discriminant.
Giải bài toán tối ưu:

$$\frac{d}{dw}[J(w)] = \frac{d}{d(w)}\left[\frac{w^T S_B w}{w^T S_w w}\right] = 0$$

$$\Rightarrow [w^T S_w w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_w w]}{dw} = 0$$

$$\Rightarrow [w^T S_w w]2S_B w - [w^T S_B w]2S_w w = 0$$

Chia cho $w^T S_w w$:

$$\frac{[w^T S_w w]}{[w^T S_w w]}S_B w - \frac{[w^T S_B w]}{[w^T S_w w]}S_w w = 0$$

$$\Rightarrow S_B w - JS_w w = 0$$

$$\Rightarrow S_w^{-1} S_B w - Jw = 0$$

Ta có, bài toán giá trị riêng:

$$\boxed{w^* = \arg\max_w \left(\frac{w^T S_B w}{w^T S_w w}\right) = S_w^{-1}(\mu_1 - \mu_2)}$$

Ví dụ: Cho tập dữ liệu sau:
Lớp 1 có 5 mẫu $C_1 = [(1,2),(2,3),(3,3),(4,5),(5,5)]$
Lớp 2 có 6 mẫu $C_2 = [(1,0),(2,1),(3,1),(3,2),(5,3),(6,5)]$
Biểu diễn dữ liệu dưới dạng vector:

$$C_1 = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 3 & 2 \\ 5 & 3 \\ 6 & 5 \end{bmatrix}$$

Giải

Bước 1: Tính kì vọng của từng lớp

$$m_1 = E(C_1) = \begin{bmatrix} 3 & 3.6 \end{bmatrix}, \quad m_2 = E(C_2) = \begin{bmatrix} 3.3 & 2 \end{bmatrix}$$

Bước 2: Tính S_1 và S_2

$$S_1 = 4 * cov(C_1) = \begin{bmatrix} 10 & 8.0 \\ 8.0 & 7.2 \end{bmatrix}, \quad S_2 = 5 * cov(C_2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}$$

Bước 3: Tính S_w

$$S_w = s_1 + s_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$$

Bước 4: Tính S_w^{-1}

$$S_w^{-1} = inv(S_w) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$$

Bước 5: Tính vecto w

$$w = S_w^{-1}(m_1 - m_2) = \begin{bmatrix} -0.79 \\ 0.89 \end{bmatrix}$$

Bước 6: Tính y_1, y_2

$$y_1 = w^T C_1^T = \begin{bmatrix} -0.81 & 0.89 & 0.24 & 1.05 & 0.4 \end{bmatrix}$$

$$y_2 = w^T C_2^T = \begin{bmatrix} -0.65 & -0.57 & -1.22 & -0.49 & -1.06 & -0.25 \end{bmatrix}$$

5.3 Bài toán phân loại nhiều lớp

Với bài toán có C lớp, ta có thể giảm số chiều về $1, 2, 3, ..., C-1$ chiều.

Giả sử rằng chiều mà chúng ta muốn giảm về là $D^{'} < D$ và dữ liệu mới ứng với mỗi điểm dữ liệu x là:

$$y = W^T x, \qquad W \in R^{DxD^{'}}$$

trong đó W là ma trận chiếu.

Ma trận chiếu:

Ma trận chiếu được hình thành từ $(C-1)$ vectơ chiếu w_i và được biểu diễn như sau:

$$W = [w_1|w_2|...|w_{C-1}]$$

Nhận xét: Độ phân tán của một tập hợp dữ liệu có thể được coi như tổng bình phương khoảng cách từ mỗi điểm tới vector kỳ vọng của chúng. Nếu tất cả các điểm đều gần vector kỳ vọng của chúng thì độ phân tán của tập dữ liệu đó được coi là nhỏ. Ngược lại, nếu tổng này là lớn, tức trung bình các điểm đều xa trung tâm, tập hợp này có thể được coi là có độ phân tán cao.

Hàm mục tiêu:

$$J(W) = \frac{det(W^T S_B W)}{det(W^T S_w W)}$$

Within-class covariance matrix:

$$S_w = \sum_{i=1}^{C} S_i = \sum_{i=1}^{C} \sum_{x_k \in C_i} (x_k - m_i)(x_k - m_i)^T$$

Between-class covariance matrix:

$$S_B = \sum_{i=1}^{C} n_i (m_i - M)(m_i - M)^T$$

trong đó M là giá trị kì vọng của tất cả điểm dữ liệu và được biểu diễn như sau:

$$M = \frac{1}{N} \sum_{x_i} x_i = \frac{1}{N} \sum_{x \in C_i} n_i m_i$$

Nghiệm của bài toán tối ưu:
Tương tự với bài toán phân loại 2 lớp, ta có thể đưa được bài toán tối ưu về dạng:

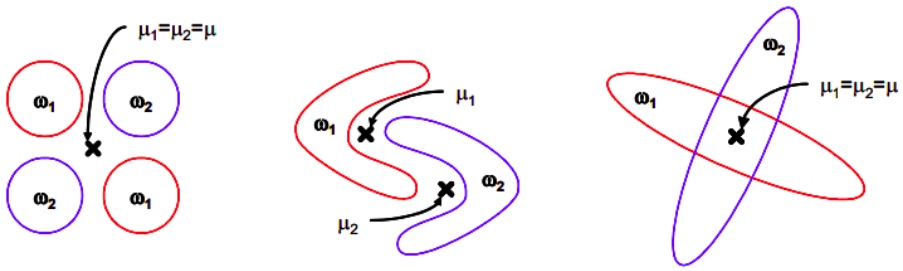
$$S_B W = \lambda S_w W$$

Từ đó suy ra, mỗi cột của W là một vecto riêng của $S_w^{-1} S_B$ ứng với trị riêng lớn nhất của ma trận này.
Nhận xét:

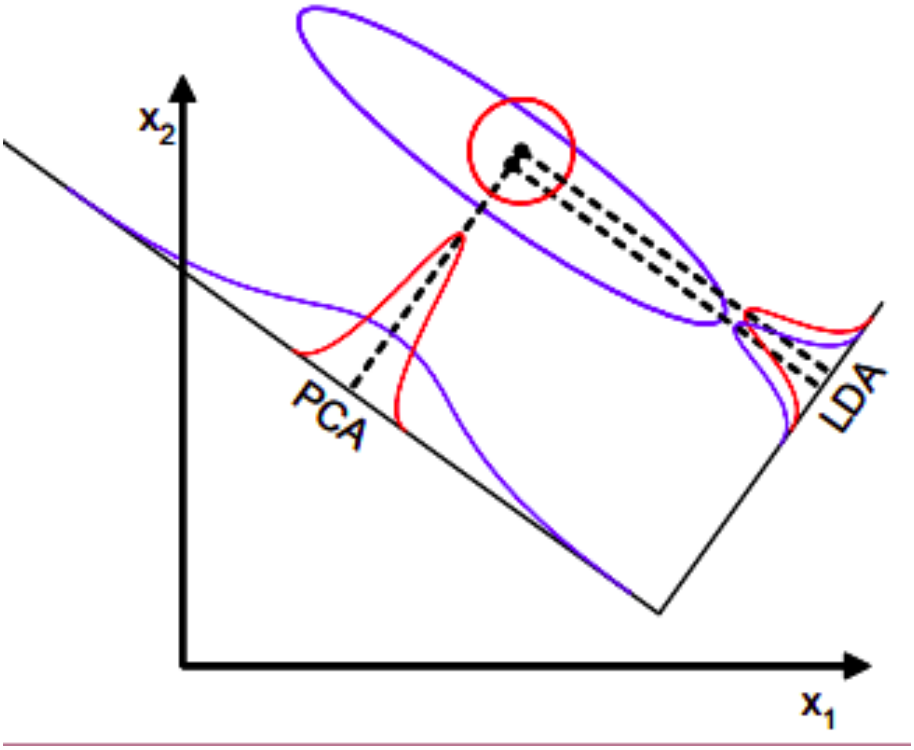
Các cột của W cần phải độc lập tuyến tính. Vì nếu không, dữ liệu trong không gian mới $y = W^T x$ sẽ phụ thuộc tuyến tính và có thể tiếp tục được giảm số chiều mà không ảnh hưởng gì. S_B là tổng của C ma trận bậc $\leq (C-1)$. Do đó, chỉ có $(C-1)\lambda_i$ là khác 0.

5.4 Nhược điểm của LDA

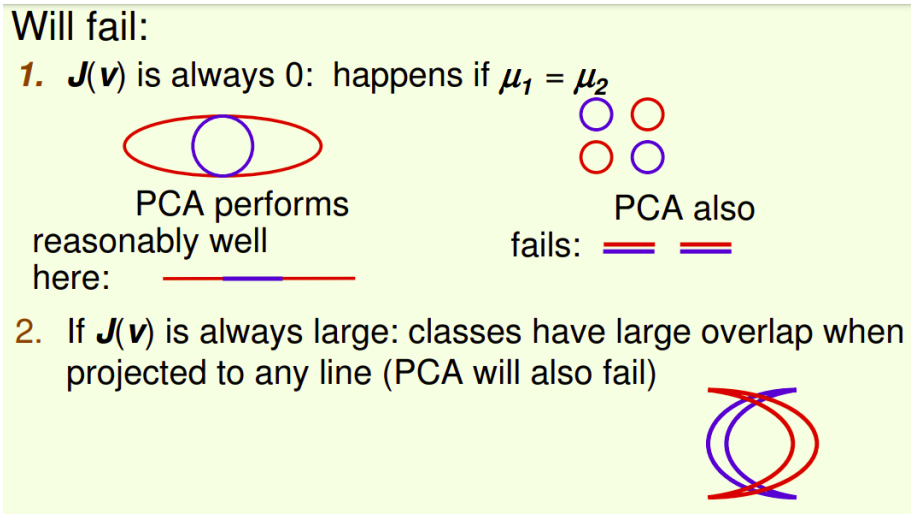
LDA chỉ có thể tạo ra nhiều nhất $C-1$ phép chiếu.



Hình 5.3: Nếu dữ liệu không có dạng phân phối chuẩn thì phép chiếu của LDA sẽ không giữ được cấu trúc dữ liệu và phân loại kém hiệu quả.



Hình 5.4: Khi việc phân biệt dữ liệu không nằm ở giá trị kì vọng mà phụ thuộc vào phương sai của dữ liệu.



Hình 5.5: Các nhược điểm khác của LDA

Chương 6

Mô hình Logistic

6.1 Mở đầu về hồi quy logistic

Chúng ta đã đề cập đến phân loại dựa trên các biến định lượng. Đối với các biến trong đó một số hoặc tất cả đều là định tính, ta có một cách tiếp cận mới, đó là sử dụng mô hình hồi quy logistic.

Trong trường hợp đơn giản nhất, ta xét hồi quy logistic nhị thức, tức là chỉ xảy ra hai khả năng, VD: "Không" hoặc "Có", "Thất bại" hoặc "Thành công". Ta có thể mã hóa giá trị cho hai khả năng này là 0 và 1.
Xét biến nhị phân Y tuân theo phân phối Bernoulli $Y \sim B(1,p), 0 \leq p \leq 1$.
Đặt: $P(Y=1)=p$, $P(Y=0)=1-p$
Khi đó: $E(Y)=1*p+0*(1-p)=p$,
$V(Y)=1^2*p+0^2*(1-p)-p^2=p(1-p)$
Giả sử Y phụ thuộc vào một biến dự đoán Z nào đó.
Ta có: $p(z)=P(Y=1|Z=z)=E(Y|Z=z)$
$\rightarrow$Ta thực hiện mô hình hóa xác suất bằng 1 với một mô hình tuyến tính như sau:

$$p=E(Y|z)=\beta_0+\beta_1 z+\varepsilon$$

6.2 Mô hình Logit

Xét tỷ lệ odds: $odds=\frac{p}{1-p}$ là tỷ lệ giữa xác suất của 1 với xác suất của 0, tỷ lệ này có thể lớn hơn 1.
Khi đó, mô hình logit được định nghĩa như sau:

$$logit(p)=ln(odds)=ln\left(\frac{p}{1-p}\right) \tag{6.1}$$

Logit là một hàm của xác suất p. Giả sử rằng, đồ thị logit là một đường thẳng đối với biến dự đoán Z, khi đó:

$$logit(p)=ln(odds)=ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 z \tag{6.2}$$

Lũy thừa cơ số e hai vế của (6.2), suy ra:

$$\theta(z) = \frac{p(z)}{1-p(z)} = exp(\beta_0 + \beta_1 z)$$

Giải phương trình $\theta(z)$ ta được:

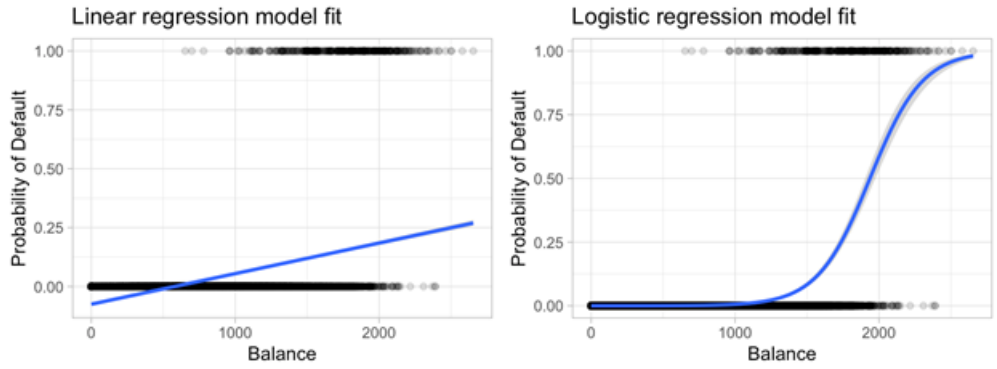
$$p(z) = \frac{exp(\beta_0 + \beta_1 z)}{1 + exp(\beta_0 + \beta_1 z)} = \frac{1}{1 + exp(-\beta_0 - \beta_1 z)} \tag{6.3}$$

p(z) được gọi là hàm logistic, hàm này lấy giá trị trong khoảng từ 0 đến 1 và đồ thị của nó là một đường cong hình chữ S.
Khi đó, hàm phân phối logistic có dạng như sau:

$$F(z) = logistic(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Hình vẽ dưới đây mô tả mối quan hệ giữa logit(p) với z và giữa p với z

$$logit(p) = \beta_0 + \beta z \qquad\qquad p = \frac{e^{\beta_0+\beta z}}{1+e^{\beta_0+\beta z}}$$



Hình 6.1: Mối quan hệ giữa logit(p) với z và p với z

6.3 Phân tích hồi quy Logistic

Gọi $(z_{j1}, z_{j2}, ..., z_{jr})$ là giá trị của r dự đoán cho lần quan sát thứ j.
Đặt $z_j = [1, z_{j1}, z_{j2}, ..., z_{jr}]'$, giả sử rằng quan sát Y_j tuân theo phân phối Bernoulli với xác suất thành công là $p(z_j)$ phụ thuộc vào giá trị của các biến dự đoán. Khi đó:

$$P(Y_j = y_j) = p^{y_j}(z_j)(1-p(z_j))^{1-y_j}; y_j = 0, 1$$

Vì vậy, $E(Y_j) = p(z_j), \qquad Var(Y_j) = p(z_j)(1-p(z_j))$
Dựa trên mô hình logit đối với một biến dự đoán đã được xây dựng ở mục 6.2, ta tổng quát hóa mô hình logit với r biến dự đoán như sau:

$$ln\left(\frac{p(z)}{1-p(z)}\right) = \beta_0 + \beta_1 z_1 + ... + \beta_r z_r = \beta' z_j \tag{6.4}$$

ở đó, $\beta = [\beta_0, \beta_1, ..., \beta_r]'$
Để ước lượng cho β, ta sẽ sử dụng phương pháp ước lượng hợp lý cực đại.

6.3.1 Ước lượng hợp lý cực đại

Gọi L là hàm hợp lý phụ thuộc vào các tham số $b_0, b_1, ..., b_r$ qua r biến dự đoán được đánh giá tại n lần quan sát. Khi đó, ta có:

$$L(b_0, b_1, ..., b_r) = \prod_{j=1}^{n} p^{y_j}(z_j)(1-p(z_j))^{1-y_j} = \frac{\prod_{j=1}^{n} e^{y_j(b_0+b_1 z_{j1}+...+b_r z_{jr})}}{\prod_{j=1}^{n}(1+e^{b_0+b_1 z_{j1}+...+b_r z_{jr}})} \tag{6.5}$$

Giá trị của các tham số hợp lý cực đại không thể được biểu diễn trong một nghiệm dạng đóng tốt như trong trường hợp mô hình tuyến thông thường. Thay vào đó, chúng phải được xác định bằng số, xuất phát với một phỏng đoán ban đầu và lặp lại đến mức cực đại của hàm hợp lý. Về mặt kỹ thuật, quy trình này được gọi là phương pháp bình phương nhỏ nhất được tái trọng số lặp đi lặp lại.
Gọi vectơ $\widehat{\beta}$ là số các giá trị thu được của các ước lượng hợp lý cực đại.

6.3.2 Khoảng tin cậy cho các tham số

Khi kích thước mẫu lớn, $\widehat{\beta}$ là xấp xỉ chuẩn với giá trị trung bình β. Khi đó:

$$\widehat{Cov}(\widehat{\beta}) \approx \left[\sum_{j=1}^{n} \widehat{p}(z_j)(1-\widehat{p}(z_j)) z_j z_j'\right]^{-1} \tag{6.6}$$

Căn bậc hai của các phần tử đường chéo của ma trận này là mẫu lớn được ước lượng độ lệch chuẩn hoặc sai số chuẩn hóa (SE) của các ước lượng $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_r$ tương ứng. Khoảng tin cậy mẫu lớn 95% cho β_k là:

$$\widehat{\beta}_k \pm 1.96 SE(\widehat{\beta}_k), k = 0, 1, ..., r \tag{6.7}$$

6.3.3 Kiểm định tỷ lệ hợp lý

Đối với mô hình có r biến dự đoán cộng với hằng số, thị hàm hợp lý cực đại được biểu thị như sau $L_{max} = L(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_r)$.
Nếu giả thuyết không là $H_0 : \beta_k = 0$, khi đó:

$$L_{max,Reduced} = L(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta_{k-1}}, \widehat{\beta_{k+1}}, ..., \widehat{\beta}_k)$$

Kiểm định H_0 bằng cách tính giá trị:

$$-2ln\left(\frac{L_{max,Reduced}}{L_{max}}\right) \tag{6.8}$$

được gọi là deviance (tạm dịch là độ lệch). H_0 bị bác bỏ nếu giá trị của độ lệch (6.8) lớn.
Một kiểm định định khác có thể được sử dụng ở đây là kiểm định của Wald. Kiểm định Wald của H_0: $\beta_k = 0$ sử dụng kiểm định thống kê $Z = \frac{\widehat{\beta}_k}{SE(\widehat{\beta}_k)}$ hoặc khi bình phương Z^2 của nó với 1 bậc tự do. Tuy nhiên, kiểm định tỷ lệ hợp lý sẽ hiệu quả hơn kiểm định của Wald vì mức của kiểm định tỷ lệ hợp lý thường gần với mức ý nghĩa α hơn.

6.3.4 Hồi quy logistic nhị thức trong trường hợp tổng quát

Xét trường hợp tổng quát trong đó một số lần chạy được thực hiện ở cùng giá trị của các biến dự đoán z_j và có tổng số m tập khác nhau trong đó các biến dự đoán này là không đổi. Tiến hành n_j phép thử độc lập với các biến dự đoán z_j. Gia sử Y_j được mô hình hóa dưới dạng 1 phân phối nhị thức với xác suất $p(z_j) = P(Success|z_j)$.
Vì Y_j được giả định là độc lập nên hàm hợp lý là tích số

$$L(\beta_0, \beta_1, ..., \beta_r) = \prod_{j=1}^{m} \binom{n_j}{y_j} p^{y_j}(z_j)(1 - p(z_j))^{n_j - y_j} \tag{6.9}$$

trong đó xác suất p^{y_j} tuân theo mô hình logit (6.4).
Khi tổng kích thước mẫu lớn, ta có:

$$\widehat{Cov}(\widehat{\beta}) \approx \left[\sum_{j=1}^{m} n_j \widehat{p}(z_j)(1 - \widehat{p}(z_j)) z_j z_j^{'}\right]^{-1} \tag{6.10}$$

và phần tử đường chéo thứ i là ước lượng phương sai của $\widehat{\beta}_{i+1}$. Căn bậc hai là ước lượng của sai số chuẩn hóa mẫu lớn $SE(\widehat{\beta}_{i+1})$.
Ước lượng phương sai của xác suất $p(z_j)$ trong trường hợp mẫu lớn bằng:

$$\widehat{Var}(\widehat{p}(z_k)) \approx (\widehat{p}(z_k)(1 - \widehat{p}(z_k)))^2 \left[\sum_{j=1}^{m} n_j \widehat{p}(z_j)(1 - \widehat{p}(z_j)) z_j z_j^{'}\right]^{-1} z_k$$

*** Phần dư và kiểm định Goodness-of-Fit:** Phần dư có thể được kiểm tra để tìm các mẫu cho thấy sự thiếu phù hợp của dạng mô hình logit và sự lựa chọn của các biến dự đoán. Trong hồi quy logistic phần dư không được định nghĩa rõ ràng như trong các mô hình hồi quy bội.
-Deviance residuals (tạm dịch là độ lệch phần dư) (d_j):

$$d_j = \pm\sqrt{2\left[y_j ln\left(\frac{y_j}{n_j\widehat{p}(z_j)}\right) + (n_j - y_j) ln\left(\frac{n_j - y_j}{n_j(1-\widehat{p}(z_j))}\right)\right]}$$

trong đó dấu của d_j trùng dấu của $y_j - n_j\widehat{p}(z_j)$ và

+Nếu $y_j = 0$, khi đó

$$d_j = -\sqrt{2n_j\left|ln(1-\widehat{p}(z_j))\right|}$$

+Nếu $y_j = n_j$, khi đó

$$d_j = -\sqrt{2n_j\left|ln(\widehat{p}(z_j))\right|} \tag{6.11}$$

-Phần dư Pearson (r_j):

$$r_j = \frac{y_j - n_j\widehat{p}(z_j)}{\sqrt{n_j\widehat{p}(z_j)(1-\widehat{p}(z_j))}} \tag{6.12}$$

-Phần dư Pearson được chuẩn hóa r_{sj}:

$$r_{sj} = \frac{r_j}{\sqrt{1-h_{jj}}} \tag{6.13}$$

với h_{jj} là phần tử thứ (j,j) trong ma trận "mũ" H được cho bởi:

$$H = V^{\frac{-1}{2}}Z(Z'V^{-1}Z)^{-1}Z'V^{\frac{-1}{2}} \tag{6.14}$$

trong đó, V^{-1} là ma trận đường chéo với (j , j) phần tử $n_j\widehat{p}(z_j)(1-\widehat{p}(z_j))$.

$V^{\frac{-1}{2}}$ là ma trận đường chéo với (j, j) phần tử $\sqrt{n_j\widehat{p}(z_j)(1-\widehat{p}(z_j))}$

Giá trị phần dư lớn hơn khoảng 2,5 cho thấy lack of fit (tạm dịch là thiếu sự phù hợp) ở biến dự đoán z_j cụ thể.

Một kiểm định tổng thể của goodness of fit được ưu tiên đặc biệt đối với các kích thước mẫu nhỏ hơn - được cung cấp bởi thống kê khi bình phương của Pearson

$$X^2 = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{n} \frac{(y_j - n_j\widehat{p}(z_j))^2}{n_j\widehat{p}(z_j)(1-\widehat{p}(z_j))} \tag{6.15}$$

Lưu ý rằng thống kê khi bình phương (6.15) là tổng bình phương của các phần dư Pearson. Việc kiểm tra phần dư Pearson cho phép kiểm tra chất lượng của fit trên toàn bộ mẫu của các các biến dự báo.

6.4 Quy tắc phân loại sử dụng hồi quy logistic

Gọi biến phản hồi Y là 1 nếu đơn vị quan sát thuộc quần thể 1 và 0 nếu nó thuộc quần thể 2. Khi một hàm hồi quy logistic đã được thiết lập và sử dụng các tập trainning (tạm dịch là tập huấn luyện) cho mỗi nhóm trong hai quần

thể, chúng ta có thể tiến hành phân loại. Và quy tắc phân loại được phát biểu như sau:

+Gán z cho quần thể 1 nếu tỷ lệ odds được ước lượng lớn hơn 1, tức là:

$$\frac{\widehat{p}(z)}{1-\widehat{p}(z)} = exp(\widehat{\beta}_0 + \widehat{\beta}_1 z_1 + ... + \widehat{\beta}_r z_r) > 1$$

+Gán z cho quần thể 1 nếu phân biệt tuyến tính lớn hơn 0, tức là:

$$ln\left(\frac{\widehat{p}(z)}{1-\widehat{p}(z)}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 z_1 + ... + \widehat{\beta}_r z_r > 0$$

Chương 7

Bài toán thực tế: Bài toán dự đoán bệnh tim

Giới thiệu bài toán

- Bệnh tim bao gồm các bệnh về mạch máu, các bệnh về nhịp tim (loạn nhịp tim) và các dị tật tim bẩm sinh. Bệnh tim mạch thường liên quan đến các mạch máu bị thu hẹp hoặc tắc nghẽn có thể dẫn đến đau tim, đau ngực (đau thắt ngực) hoặc đột quỵ, những bệnh ảnh hưởng đến cơ, van hoặc nhịp tim,...
- Khoảng 610.000 người chết vì bệnh tim ở Hoa Kỳ hàng năm tức cứ 4 người thì có 1 người chết vì bệnh tim. Bệnh tim là nguyên nhân gây tử vong hàng đầu cho cả nam và nữ. Hơn một nửa số ca tử vong do bệnh tim trong năm 2009 là ở nam giới, trong đó, bệnh mạch vành là loại bệnh tim phổ biến nhất, giết chết hơn 370.000 người hàng năm.
- Khoảng 735.000 người Mỹ bị đau tim mỗi năm. Trong số này, 525.000 trường hợp là cơn đau tim đầu tiên và 210.000 trường hợp xảy ra ở những người đã từng bị đau tim.
- Bệnh tim rất khó để xác định vì một số yếu tố khác góp phần như tiểu đường, huyết áp cao, cholesterol cao, rối loạn nhịp tim,...Vì thế việc phát hiện ra bệnh tim từ sớm đã trở thành một bài toán quan trọng cần được giải quyết.
- Do đó, các nhà khoa học đã áp dụng các phương thức tiếp cận hiện đại để dự đoán bệnh. Và cụ thể ở đây là thuật toán hồi quy logistic và phân tích phân biệt để phân loại người bệnh vào các cấp độ xuất hiện triệu chứng bệnh tim từ 0 (không có riệu chứng bệnh tim) tới 4.

Giới thiệu dữ liệu

*Link dataset: https://archive.ics.uci.edu/ml/datasets/heart+disease
*Bộ dữ liệu chứa 303 bản ghi và 14 trường dữ liệu:

- Age: tuổi
- Sex: giới tính (0 = nữ; 1 = nam)
- Chest-pain type (cp): hình thức đau ngực
  + 0 = đau thắt tiệm cận

+ 1 = đau thắt ngực không điển hình
+ 2 = đau thắt ngực không do tim
+ 3 = đau thắt ngực điển hình

- Resting Blood Pressure (trestbps): giá trị huyết áp của một người (mmHg)
- Serum Cholestrol (chol): lượng cholesterol trong huyết thanh (mg/dl)
- Fasting Blood Sugar (fbs): so sánh giá trị đường huyết một người với 120mg/dl. ($1 > 120; 0 \leq 120$)
- Resting ECG (restecg): kết quả điện tâm đồ
  + 0 = phì đại tâm thất trái
  + 1 = bình thường
  + 2 = có bất thường sóng ST-T
- Max heart rate achieved (thalach): nhịp tim tối đa đạt được
- Exercise induced angina (exang): có bị triệu chứng đau thắt ngực khi tập luyện thể dục không (0 = không; 1 = có)
- ST depression induced by exercise relative to rest (oldpeak): chênh xuống tại đoạn ST lúc tập thể dục với lúc nghỉ ngơi, là một số nguyên hoặc số thực.
- Peak exercise ST segment (slope): đoạn dốc ST tại thời điểm đỉnh điểm của bài tập
  + 0 = xuống dốc
  + 1 = phẳng
  + 2 = đi lên
- Number of major vessels (0–4) colored by flourosopy (ca): số động mạch được nhuộm bằng phương pháp flourosopy (một thủ tục X-quang sử dụng thuốc nhuộm và máy ảnh đặc biệt), là một số nguyên hoặc thực
- Thal: bệnh Thalassemia (thiếu máu Địa Trung Hải)
  + 0 = NULL
  + 1 = có khuyết tật cố định
  + 2 = bình thường
  + 3 = có khuyết tật đảo ngược
- Diagnosis of heart disease : kết quả chẩn đoán bệnh tim (0 = hoàn toàn không có dấu hiệu mắc bệnh tim; 1 = có dấu hiệu mắc bệnh).

Code chạy bài toán

Ngôn ngữ: Python 3.7
Phần mềm: Jupyter Notebook

Bước 1: Import các thư viện và định nghĩa hàm random cần thiết

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
import plotly.tools as tls
import plotly.express as px
import plotly.figure_factory as ff

from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot

%matplotlib inline
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")

from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis


from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import random

def random_colors(number_of_colors):
    color = ["#"+''.join([random.choice('0123456789ABCDEF') for j in range(6)])
                 for i in range(number_of_colors)]
    return color
```

Bước 2: Xuất ra một vài thông tin về dữ liệu

```
train = pd.read_csv("heart.csv")
```

```
table = ff.create_table(train.head().round(3))
iplot(table,filename='jupyter-table1')
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63.0	1.0	3.0	145.0	233.0	1.0	0.0	150.0	0.0	2.3	0.0	0.0	1.0	1.0
37.0	1.0	2.0	130.0	250.0	0.0	1.0	187.0	0.0	3.5	0.0	0.0	2.0	1.0
41.0	0.0	1.0	130.0	204.0	0.0	0.0	172.0	0.0	1.4	2.0	0.0	2.0	1.0
56.0	1.0	1.0	120.0	236.0	0.0	1.0	178.0	0.0	0.8	2.0	0.0	2.0	1.0
57.0	0.0	0.0	120.0	354.0	0.0	1.0	163.0	1.0	0.6	2.0	0.0	2.0	1.0

```
train.columns
```

```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

```
train.shape
```

```
(303, 14)
```

```
iplot(ff.create_table(train.dtypes.to_frame().reset_index().round(3)),filename='jupyter-table2')
```

index	0
age	int64
sex	int64
cp	int64
trestbps	int64
chol	int64
fbs	int64
restecg	int64
thalach	int64
exang	int64
oldpeak	float64
slope	int64
ca	int64
thal	int64
target	int64

Bước 3: Tính toán các thông số liên quan

```
iplot(ff.create_table(train.describe().reset_index().round(3)),filename='jupyter-table2')
```

index	age	sex	cp	trestbpschol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0	303.0
mean	54.366	0.683	0.967	131.624	246.264	0.149	0.528	149.647	0.327	1.04	1.399	0.729	2.314	0.545
std	9.082	0.466	1.032	17.538	51.831	0.356	0.526	22.905	0.47	1.161	0.616	1.023	0.612	0.499
min	29.0	0.0	0.0	94.0	126.0	0.0	0.0	71.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	47.5	0.0	0.0	120.0	211.0	0.0	0.0	133.5	0.0	0.0	1.0	0.0	2.0	0.0
50%	55.0	1.0	1.0	130.0	240.0	0.0	1.0	153.0	0.0	0.8	1.0	0.0	2.0	1.0
75%	61.0	1.0	2.0	140.0	274.5	0.0	1.0	166.0	1.0	1.6	2.0	1.0	3.0	1.0
max	77.0	1.0	3.0	200.0	564.0	1.0	2.0	202.0	1.0	6.2	2.0	4.0	3.0	1.0

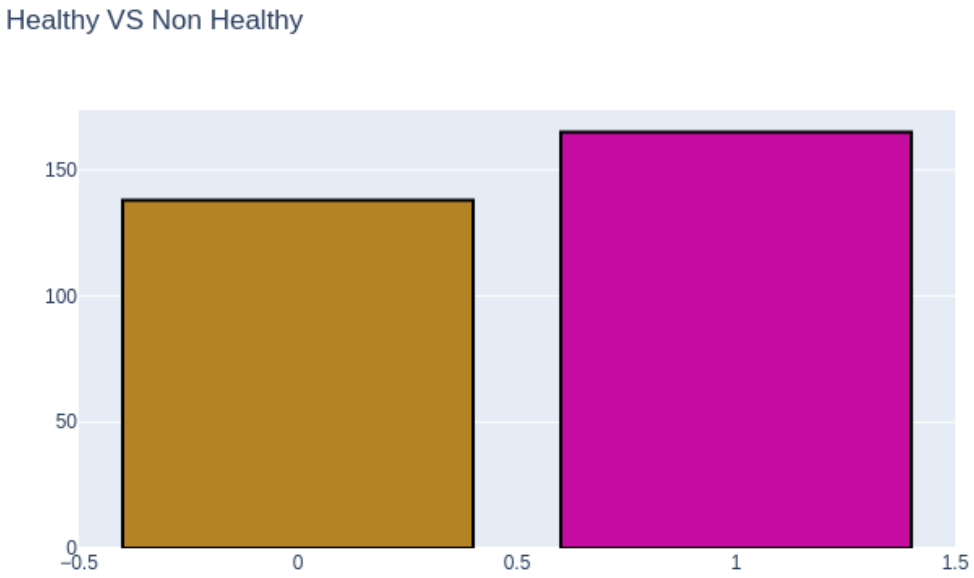
Bước 4: Khai phá dữ liệu

```
train.isnull().sum()
```

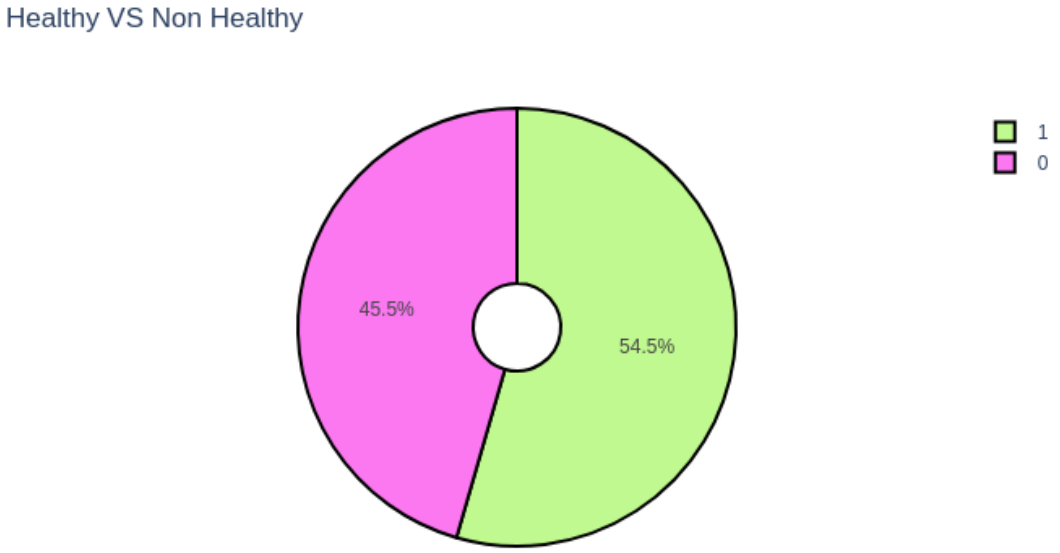
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Phân tích trường ’Target’

```
species_count = train['target'].value_counts()
data = [go.Bar(
    x = species_count.index,
    y = species_count.values,
    marker = dict(color = random_colors(3),line=dict(color='#000000', width=2))
)]

layout = go.Layout(
   {
      "title":"Healthy VS Non Healthy",
   }
)

fig = go.Figure(data=data,layout = layout)
iplot(fig)
```

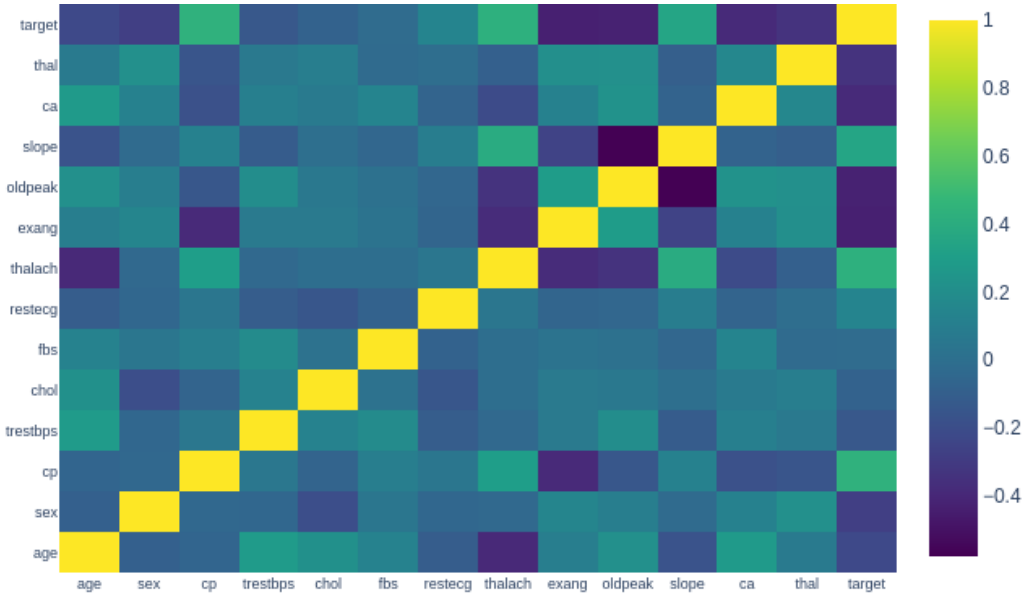



```
trace = go.Pie(labels = list(train.target.unique()), values = list(train.target.value_counts()),
                               hole = 0.2,
               marker=dict(colors = random_colors(3),
                           line=dict(color='#000000', width=2)
                           ))
data = [trace]
layout = go.Layout(
   {
      "title":"Healthy VS Non Healthy",
   }
)

fig = go.Figure(data=data,layout = layout)
iplot(fig)
```



Sự tương quan giữa các trường dữ liệu

```
data = [go.Heatmap(z = np.array(train.corr().values),
                   x = np.array(train.corr().columns),
                   y = np.array(train.corr().columns),
                     colorscale='Viridis',)
       ]
layout = go.Layout(dict(title = 'Correlation Matrix for variables',
                        #autosize = False,
                        #height  = 1400,
                        #width   = 1600,
                        margin  = dict(r = 0 ,l = 100,
                                       t = 0,b = 100,
                                     ),
                        yaxis   = dict(tickfont = dict(size = 9)),
                        xaxis   = dict(tickfont = dict(size = 9)),
                       )
                  )

fig = go.Figure(data=data,layout = layout)
iplot(fig)
```



Bước 5: Chuẩn bị dữ liệu

```
X = train.iloc[:,:-1].values    # loại bỏ cột kết quả 'Target' và trả về mảng array
y = train.iloc[:,-1].values     # chỉ giữ lại cột kết quả 'Target' và trả về mảng array
encoder = LabelEncoder()
y = encoder.fit_transform(y)

# Chia bộ dữ liệu thành 2 bộ nhỏ hơn: Training và Test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)
```

Tại đây, ta tách cột ’Target’ ra thành một bộ dữ liệu riêng: bộ nhãn dữ liệu. Tiếp đó ta lấy ngẫu nhiên các mẫu trong bộ dữ liệu ban đầu và chia ra thành 2 bộ luyện và test với tỉ lệ 7:3.

Bước 6: Logistic Regression

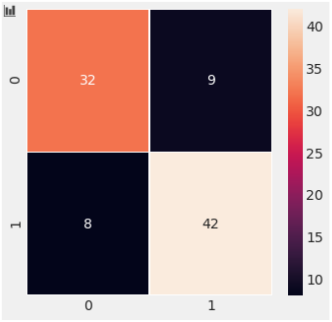
```
Model = LogisticRegression()
Model.fit(X_train, y_train)
y_pred = Model.predict(X_test)

# Tổng hợp kết quả thu được từ bộ phân loại
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.80      0.78      0.79        41
           1       0.82      0.84      0.83        50

    accuracy                           0.81        91
   macro avg       0.81      0.81      0.81        91
weighted avg       0.81      0.81      0.81        91
```

Confussion matrix

```
f,ax = plt.subplots(figsize = (5,5))
sns.heatmap(confusion_matrix(y_test, y_pred),annot = True,linewidth = 1,fmt =".0f",ax = ax)
# Accuracy score
print('accuracy is',accuracy_score(y_pred,y_test))
```

```
accuracy is 0.8131868131868132
```



Từ hình ảnh trên ta thấy với phương pháp LR, ta phân loại với lớp 0 (khỏe mạnh, Healthy) đúng 32 mẫu, sai 8 mẫu, còn với lớp 1 (có dấu hiệu bệnh, Non-Healthy) đúng 42 mẫu, sai 9 mẫu.

Bước 7: Linear Discrimination Analysis

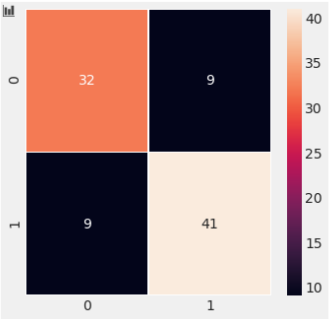
```
# Linear Discriminant Analysis

Model = LinearDiscriminantAnalysis()
Model.fit(X_train, y_train)

y_pred = Model.predict(X_test)

# Tổng hợp kết quả thu được từ bộ phân loại
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.78      0.78      0.78        41
           1       0.82      0.82      0.82        50

    accuracy                           0.80        91
   macro avg       0.80      0.80      0.80        91
weighted avg       0.80      0.80      0.80        91
```

Confussion matrix

```
f,ax = plt.subplots(figsize = (5,5))
sns.heatmap(confusion_matrix(y_test, y_pred),annot = True,linewidth = 1,fmt =".0f",ax = ax)
# Accuracy score
print('accuracy is',accuracy_score(y_pred,y_test))
```

```
accuracy is 0.8021978021978022
```



Từ hình ảnh trên ta thấy với phương pháp LR, ta phân loại với lớp 0 (khỏe mạnh, Healthy) đúng 32 mẫu, sai 9 mẫu, còn với lớp 1 (có dấu hiệu bệnh, Non-Healthy) đúng 41 mẫu, sai 9 mẫu.

Tài Liệu Tham Khảo

1. Nguyễn Văn Hữu - Nguyễn Hữu Dư, *Phân tích thống kê và dự báo*, NXB Đại Học Quốc Gia Hà Nội, xuất bản năm 2003.

2. Richard A.Johnson and Dean W.Wechern, *Applied Multivariate Statistical Analysis*, sixth edition, 2007.

3. Th.S Lê Xuân Lý, Bài giảng môn phân tích số liệu.

4. Iris dataset https://archive.ics.uci.edu/ml/datasets/iris, https://en.wikipedia.org/wiki/Iris-flower-data-set

5. https://www.mghassany.com/MLcourse/logistic-regression.htmllogreg-examps

6. https://machinelearningcoban.com/2017/01/27/logisticregression/

7. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

8. http://www.facweb.iitkgp.ac.in/sudeshna/courses/ml08/lda.pdf