NGUYỄN ĐÌNH THÚC – VŨ HẢI QUÂN
VĂN CHÍ NAM – ĐẶNG HẢI VÂN – LÊ PHONG

Giáo trình THỐNG KÊ MÁY TÍNH

Phiên bản 0.10

NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT
2009

Lời nói đầu

Đây là lời nói đầu của giáo trình này.

Mục lục

Chương 1
LÝ THUYẾT RA QUYẾT ĐỊNH BAYES

Một trong những phương pháp giải quyết bài toán phân lớp mẫu (Pattern Classification) là *lý thuyết ra quyết định Bayes (Bayes decision theory)* – nền tảng cho hướng tiếp cận thống kê. Trong hướng tiếp cận thống kê này, các độ đo xác suất được sử dụng nhằm đưa ra quyết định mẫu đang xét thuộc lớp nào.

Lấy một ví dụ về phân loại trái cây như sau. Trong một dây chuyền phân loại trái cây, một đầu người ta đưa vào một thùng trái cây với hai loại quả: táo và lê. Với mỗi quả, hệ thống phải phân loại quả đó là táo hay là lê để cho ra hai cổng khác nhau và mang đi đóng gói.

Giả sử như nhắm mắt bốc đại một quả trong thùng thì khả năng để có được quả táo là 0.8 (*P(táo) = 0.8*), khả năng có được quả lê là 0.2 (*P(lê) = 0.2*). Khi đó, với bất kỳ quả nào đưa vào, nếu ta đều phân loại là táo thì khả năng đúng sẽ là 0.8 và khả năng sai sẽ là 0.2 (tức là trung bình với 100 quả, ta phân loại đúng 80 quả, phân loại sai 20 quả). Rõ ràng là cách làm này chỉ dựa trên một thông tin đã được biết trước mà không dựa trên bất kỳ thông tin nào của đối tượng đang được xét.

Bây giờ, giả sử như *P(táo) = P(lê) = 0.5*. Lúc này việc phân loại như trên không có bất kỳ hiệu quả nào hết. Để ý một chút ta thấy *màu* của quả táo thường *đỏ* hoặc *xanh* và màu của quả lê thường *vàng*; điều đó có nghĩa là khả năng một quả có màu vàng là lê sẽ cao hơn nếu nó là táo. Khi đó, nếu phân loại dựa trên *màu* thì khả năng phân loại đúng sẽ được nâng cao. Như vậy, ta có thêm một cơ sở để phân loại: *màu*. Ta gọi màu là một *đặc trưng* để phân loại.

Xét trường hợp xấu hơn: vẫn có một số quả táo có màu vàng. Đặc trưng màu khó có thể giúp ích để phân loại những quả này. Vì vậy cần phải có thêm một đặc trưng khác. Để ý lần nữa, ta quan sát thấy rằng với những quả táo, phần gần cuống thường phình to hơn so với đầu bên kia, trong khi với những quả lê thì ngược lại. Các thông tin đó được chứa đựng trong đặc trưng *đường viền*. Như vậy, ta đã có thêm một đặc trưng.

Ví dụ trên đây cho thấy việc áp dụng hết sức đơn giản nhưng có phần cảm tính. Để làm rõ hơn (đặc biệt là cơ sở toán học), các phần tiếp theo của chương này sẽ trình bày chi tiết lý thuyết ra quyết định Bayes cũng như một số cách áp dụng.

Mục đích của chương Sinh viên sau khi học xong chương này cần phải

- nắm được định lý Bayes, các luật ra quyết định dựa trên lý thuyết ra quyết định Bayes và phương pháp xây dựng hệ phân loại bằng biệt hàm.
- có thể viết một chương trình đơn giản để xây dựng một hệ phân loại (ví dụ phân loại táo lê bằng đặc trưng màu).

I. ĐỊNH LÝ BAYES

Xét trong một tập các đối tượng chỉ gồm c lớp $\omega_1, \omega_2, ..., \omega_c$. Gọi $P(\omega_i), i = 1..c$ là xác suất phân bố của các lớp này trong tập các đối tượng đó. Xác suất này được gọi là *xác suất tiền định (priori probability)* vì nó cho biết khả năng một đối tượng thuộc về một lớp nào đó mà không dựa trên bất kỳ thông tin mô tả nào của đối tượng.

Giả sử để phân loại đối tượng, ta sử dụng vector đặc trưng $\mathbf{x}$ [1] (ở đây, tạm xét $\mathbf{x}$ liên tục trên $\Re^d$, việc $\mathbf{x}$ rời rạc sẽ được bàn ở Phần 5). Khi đó, khả năng để một đối tượng thuộc lớp ω_i có đặc trưng $\mathbf{x}$ được cho

[1] $\mathbf{x}$ bao gồm $d \geq 1$ đặc trưng (ví dụ như cường độ sáng, độ dài đường viền,…), với mỗi đặc trưng được thể hiện bởi một số thực; vì vậy, $\mathbf{x}$ là một vector trong không gian thực *d*-chiều. Để cho thuận tiện, $\mathbf{x}$ cũng được gọi là đặc trưng với lưu ý: vector được ký hiệu $\mathbf{x}$, số thực được ký hiệu x

bởi hàm mật độ xác suất có điều kiện $p(\mathbf{x}\,|\,\omega_i)$. Hàm này được gọi là *hàm likelihood*. Khả năng để một đối tượng có đặc trưng $\mathbf{x}$ thuộc lớp ω_i được cho bởi hàm xác suất $P(\omega_i\,|\,\mathbf{x})$. Xác suất này được gọi là *xác suất hậu định (posteriori probability)* vì nó cho biết khả năng một đối tượng thuộc về lớp nào dựa trên các đặc trưng của chính đối tượng đó. Xác suất này được tính như sau

$$P(\omega_i\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,\omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{1.1}$$

trong đó

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}\,|\,\omega_i)P(\omega_i) \tag{1.2}$$

Công thức trên được gọi là định lý Bayes. Phần mẫu số $p(\mathbf{x})$ chỉ mang ý nghĩa là đảm bảo cho tổng xác suất hậu định bằng 1. Vì vậy, đôi khi người ta chỉ viết

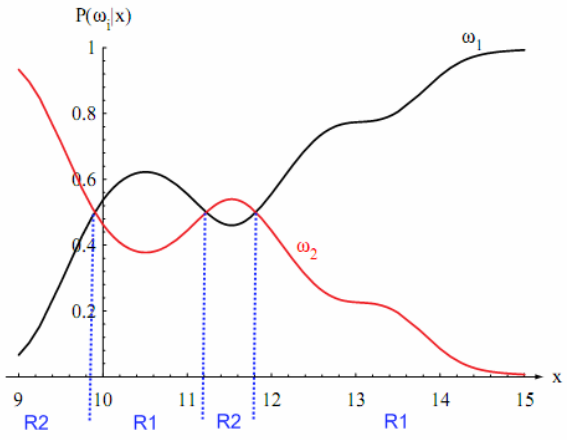
$$P(\omega_i\,|\,\mathbf{x}) \propto p(\mathbf{x}\,|\,\omega_i)P(\omega_i) \tag{1.3}$$

Định lý Bayes đơn giản nhưng có ý nghĩa to lớn. Thực vậy, xác suất hậu định gần như là không thể có được theo theo cách thống kê mẫu thông thường, trong khi với $p(\mathbf{x}\,|\,\omega_i)$ và $P(\omega_i)$ thì hoàn toàn có thể.

II. LÝ THUYẾT RA QUYẾT ĐỊNH BAYES

Dựa trên định lý Bayes, ta đã có được xác suất hậu định $P(\omega_i\,|\,\mathbf{x})$. Về mặt cảm tính, ta nhận thấy nếu $\omega_j = \max_{\omega_i, i=1..c} \arg P(\omega_i\,|\,\mathbf{x})$ thì *nên* phân đối tượng đang xét vào lớp ω_j. Trong phần này sẽ phân tích xem liệu nhận xét đó có đúng không và cơ sở toán học của nó là gì.

II.1 Trường hợp đơn giản



Hình 1 R1, R2 lần lượt là vùng đối tượng được phân lớp ω_1, ω_2. Vùng R1 tương ứng với khi $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$ và ngược lại.

Trong trường hợp này, để cho đơn giản, ta chỉ xét với hai lớp ω_1, ω_2 với luật ra quyết định như sau

Luật 1: chọn ω_1 nếu $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$, ngược lại thì chọn ω_2

Luật này chính là nhận xét đã nêu ra ở trên. Để đánh giá xem luật này có ý nghĩa thế nào, chúng ta xem thử tác động của nó lên *trung bình xác suất lỗi (average probability of error)*

$$P(error) = \int_{all\,\mathbf{x}} P(error, \mathbf{x}) d\mathbf{x} = \int_{all\,\mathbf{x}} P(error \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \qquad (1.4)$$

trong đó $P(error \mid \mathbf{x})$ được gọi là *xác suất lỗi (probability of error)* khi đưa ra quyết định và được tính bởi

$$P(error \mid \mathbf{x}) = \begin{cases} P(\omega_1 \mid \mathbf{x}) & \text{if we decide } \omega_2 \\ P(\omega_2 \mid \mathbf{x}) & \text{if we decide } \omega_1 \end{cases} \qquad (1.5)$$

Áp dụng Luật 1 chúng ta có

$$P(error \mid \mathbf{x}) = \min\left\{P(\omega_1 \mid \mathbf{x}), P(\omega_2 \mid \mathbf{x})\right\} \tag{1.6}$$

thay vào (1.4)

$$\begin{aligned} P(error) &= \int_{all\,\mathbf{x}} P(error \mid \mathbf{x})\, p(\mathbf{x}) d\mathbf{x} \\ &= \int_{all\,\mathbf{x}} \min\left\{P(\omega_1 \mid \mathbf{x}), P(\omega_2 \mid \mathbf{x})\right\} p(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{1.7}$$

Như vậy, rõ ràng là với Luật 1, chúng ta sẽ đạt được cực tiểu trung bình xác suất lỗi. **Hình 1** cho thấy một ví dụ về việc áp dụng Luật 1.

II.2 Trường hợp tổng quát

Trong trường hợp tổng quát, ta sẽ mở rộng vấn đề xa hơn như sau

- Số lớp là bất kỳ, nghĩa là có $c \geq 2$ lớp $\omega_1, \omega_2, ..., \omega_c$,
- Mở rộng việc phân loại thành a hành động (action) $\alpha_1, \alpha_2, ..., \alpha_a$. Phân loại là trường hợp đặc biệt của hành động: có $a = c$ hành động, hành động α_i phân đối tượng đang xét vào lớp ω_i,
- Sử dụng *hàm tiêu tốn (loss function)* $\lambda(.)$ để giúp tổng quát hóa cho xác suất lỗi, ví dụ như trong trường hợp đánh trọng số khác nhau cho việc phân loại sai vào các lớp khác nhau. $\lambda\left(\alpha_i \mid \omega_j\right)$ thể hiện cái giá phải trả khi thực hiện hành động α_i trong trường hợp đối tượng thuộc lớp ω_j.

Khi đó, tổng quát hóa xác suất lỗi bằng hàm *rủi ro có điều kiện (conditional risk)* như sau

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}) \tag{1.8}$$

Đẳng thức (1.8) cho thấy hàm rủi ro có điều kiện $R(\alpha_i \mid \mathbf{x})$ thể hiện cái giá phải trả khi thực hiện hành động α_i trong trường hợp đối tượng đang xét có đặc trưng $\mathbf{x}$.

Dựa trên hàm rủi ro có điều kiện, xác suất lỗi trung bình được tổng quát hóa bằng *rủi ro toàn bộ (overall risk)*

$$R = \int_{all\,\mathbf{x}} R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \tag{1.9}$$

trong đó $\alpha(\mathbf{x})$ là hàm ra quyết định nhận 1 trong a giá trị hành động $\alpha_1, \alpha_2, ..., \alpha_a$ đối với mỗi $\mathbf{x}$. Mục tiêu là phải đưa ra được $\alpha(\mathbf{x})$ để cực tiểu hóa rủi to toàn bộ.

Xét luật sau

Luật 2: chọn $\alpha(\mathbf{x}) = \arg\min_{\alpha_i, i=1..a} R(\alpha_i \mid \mathbf{x})$

Với cách chọn $\alpha(\mathbf{x})$ như Luật 2, rõ ràng rủi ro có điều kiện đạt giá trị cực tiểu, dẫn tới là rủi ro toàn bộ R cũng đạt giá trị cực tiểu R^* - giá trị cực tiểu này được gọi là rủi ro Bayes.

Bây giờ, xét hai trường hợp đặc biệt hành động phân loại; có nghĩa là có $a = c$ hành động, hành động α_i phân đối tượng đang xét vào lớp ω_i.

Số lớp bất kỳ với hàm tiêu tốn đối xứng (symmetrical loss function)

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1..c \tag{1.10}$$

Hàm tiêu tốn này mang ý nghĩa: sẽ không phải trả giá nếu phân đối tượng đang xét vào đúng lớp của nó; ngược lại, nếu phân sai thì mọi phân lớp sai sẽ chịu trả giá ngang nhau. Khi này, thay (1.10) vào (1.8) được hàm rủi ro có điều kiện

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) \\ &= 1 - P(\omega_i \mid \mathbf{x}) \end{aligned} \tag{1.11}$$

Để ý thấy rằng trong trường hợp này, hàm rủi ro có điều kiện giống với xác suất lỗi và rủi ro toàn bộ chính là xác suất lỗi trung bình.

Xét luật sau

Luật 3: chọn ω_i ***nếu*** $P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x})$ ***với mọi*** $j \neq i$

Luật 3 tương ứng với nhận xét đã được đưa ra ở đầu phần này. Rõ ràng là luật này làm cực tiểu hóa rủi ro có điều kiện $R(\alpha_i \mid \mathbf{x})$ và do đó làm cực tiểu hóa rủi ro toàn bộ.

Số lớp là 2 và hàm tiêu tốn bất kỳ. Lúc này, đặt $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$, từ (1.8) có rủi ro có điều kiện

$$\begin{aligned} R(\alpha_1 \mid \mathbf{x}) &= \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x}) \\ R(\alpha_2 \mid \mathbf{x}) &= \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x}) \end{aligned} \tag{1.12}$$

Từ luật 2 suy ra chọn hành động ω_1 nếu

$$\begin{aligned} & R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x}) \\ \Leftrightarrow\ & (\lambda_{21} - \lambda_{11}) P(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid \mathbf{x}) \\ \Leftrightarrow\ & (\lambda_{21} - \lambda_{11}) p(\mathbf{x} \mid \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x} \mid \omega_2) P(\omega_2) \end{aligned} \tag{1.13}$$

ngược lại chọn ω_2.

Ví dụ:

Xét một hệ phân lớp như sau

- có 2 lớp ω_1, ω_2 với $P(\omega_1) = 2/3$ và $P(\omega_2) = 1/3$,

- 3 hành động $\alpha_1, \alpha_2, \alpha_3$ trong đó α_1 là hành động "xếp vào lớp ω_1", α_2 là hành động "xếp vào lớp ω_2", α_3 là hành động "không phân lớp",
- giá trị của hàm tiêu tốn $\lambda(\alpha_i \mid \omega_j)$ được cho trong bảng sau

	α_1	α_2	α_3
ω_1	0	1	1/4
ω_2	1	0	1/4

- likelihood $p(x \mid \omega_1) = (2-x)/2$ và $p(x \mid \omega_2) = 1/2$ với đặc trưng x trong giới hạn $0 \le x \le 2$.

Trước tiên chúng ta xét xem với mẫu có đặc trưng x thì nên thực hiện hành động α_i nào. Từ các dữ kiện được cho, áp dụng định lý Bayes, chúng ta có được các xác suất

$$p(x) = p(x \mid \omega_1)P(\omega_1) + p(x \mid \omega_1)P(\omega_1) = \frac{2}{3}.\frac{2-x}{2} + \frac{1}{2}.\frac{1}{3} = \frac{5-2x}{6} \quad (1.14)$$

$$P(\omega_1 \mid x) = \frac{p(x \mid \omega_1)P(\omega_1)}{p(x)} = \frac{\frac{2-x}{2}.\frac{2}{3}}{\frac{5-2x}{6}} = \frac{4-2x}{5-2x} \quad (1.15)$$

$$P(\omega_2 \mid x) = \frac{p(x \mid \omega_2)P(\omega_2)}{p(x)} = \frac{\frac{1}{2}.\frac{1}{3}}{\frac{5-2x}{6}} = \frac{1}{5-2x} \quad (1.16)$$

Từ đó tính được các rủi ro có điều kiện

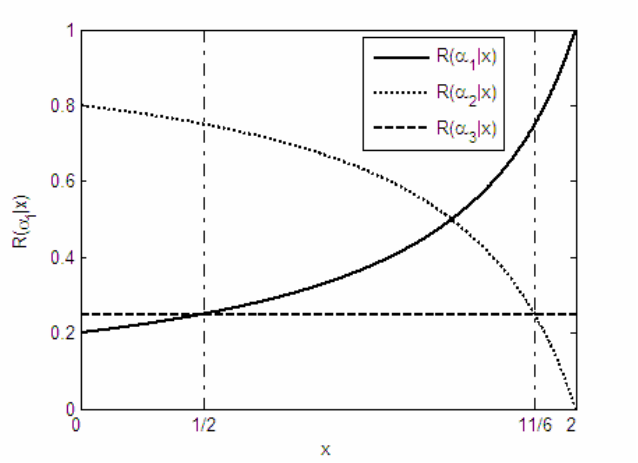
$$\begin{aligned} R(\alpha_1 \mid x) &= \lambda(\alpha_1 \mid \omega_1)P(\omega_1 \mid x) + \lambda(\alpha_1 \mid \omega_2)P(\omega_2 \mid x) \\ &= 0.\frac{4-2x}{5-2x} + 1.\frac{1}{5-2x} = \frac{1}{5-2x} \end{aligned} \quad (1.17)$$

$$R(\alpha_2|x) = \lambda(\alpha_2|\omega_1)P(\omega_1|x) + \lambda(\alpha_2|\omega_2)P(\omega_2|x)$$
$$= 1.\frac{4-2x}{5-2x} + 0.\frac{1}{5-2x} = \frac{4-2x}{5-2x} \quad (1.18)$$

$$R(\alpha_3|x) = \lambda(\alpha_3|\omega_1)P(\omega_1|x) + \lambda(\alpha_3|\omega_2)P(\omega_2|x)$$
$$= \frac{1}{4}.\frac{4-2x}{5-2x} + \frac{1}{4}.\frac{1}{5-2x} = \frac{1}{4} \quad (1.19)$$

Hình 2 cho thấy đồ thì của các rủi ro có điều kiện. Quan sát từ **Hình 2** kết hợp với Luật 2 ta có

- nếu $x \in [0,1/2]$ thì $R(\alpha_1|x)$ nhỏ nhất, do đó thực hiện hành động α_1,
- nếu $x \in (1/2,11/6)$ thì $R(\alpha_3|x)$ nhỏ nhất, do đó thực hiện hành động α_3,
- nếu $x \in [11/6,2]$ thì $R(\alpha_2|x)$ nhỏ nhất, do đó thực hiện hành động α_2.



Hình 2 Đồ thị biểu diễn các rủi ro có điều kiện

Rui ro tổng thể lúc này sẽ là cực tiểu với

$$
\begin{aligned}
R &= \int_0^2 \min\left\{R(\alpha_1 \mid x), R(\alpha_2 \mid x), R(\alpha_3 \mid x)\right\} p(x)dx \\
&= \int_0^{1/2} R(\alpha_1 \mid x)\, p(x)dx + \int_{1/2}^{11/6} R(\alpha_3 \mid x)\, p(x)dx + \int_{11/6}^{2} R(\alpha_2 \mid x)\, p(x)dx \\
&= \int_0^{1/2} \frac{1}{5-2x}.\frac{5-2x}{6}dx + \int_{1/2}^{11/6} \frac{1}{4}.\frac{5-2x}{6}dx + \int_{11/6}^{2} \frac{4-2x}{5-2x}.\frac{5-2x}{6}dx \\
&= \frac{1}{12} + \frac{4}{27} + \frac{1}{216} \simeq 0.236
\end{aligned}
$$

(1.20)

III. Phân lớp bằng biệt hàm (Discriminant function)

Trong phần này ta sẽ áp dụng lý thuyết ra quyết định Bayes để xây dựng một hệ phân lớp. Hệ phân lớp này sẽ được thể hiện dưới các *biệt hàm (discriminant function)*.

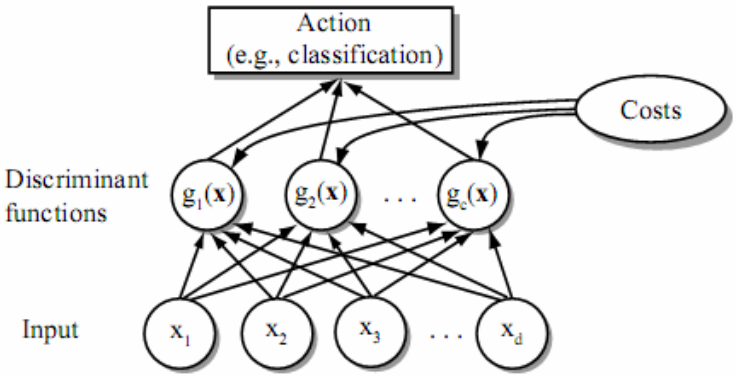
III.1 Biệt hàm và Vùng ra quyết định

III.1.1 *Biệt hàm*

Có nhiều cách để biểu diễn hệ phân lớp đối tượng. Một trong những cách hữu hiệu nhất là sử dụng các *biệt hàm* $g_i(\mathbf{x}), i = 1..c$. Khi đó, một đối tượng có đặc trưng $\mathbf{x}$ được phân vào lớp ω_i nếu

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \text{for all } j \neq i \tag{1.21}$$

Một cách hình tượng, hệ phân lớp như trên được xem như một mạng tính c biệt hàm và chọn lớp tương ứng với giá trị cao nhất như trong **Hình 3**.



Hình 3 Mô hình mạng cho hệ phân lớp được biểu diễn bằng các discriminant function.

Như vậy, áp dụng lý thuyết ra quyết định Bayes, ta nhận thấy

- trong trường hợp tính chi phí bằng xác suất lỗi trung bình, theo Luật 3, ta có thể áp dụng $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$,
- trong trường hợp tính chi phí bằng rủi ro toàn bộ, theo Luật 2, ta có thể áp dụng $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$.

Ở đây có một điều cần lưu ý. Nếu đã có một bộ biệt hàm $g_i(\mathbf{x}), i = 1..c$, ta có thể xây dựng một bộ biệt hàm mới $g_i^*(\mathbf{x}) = f(g_i(\mathbf{x})), i = 1..c$ với $f(.)$ là một hàm số đơn điệu tăng. Ví dụ như có thể đồng loạt cộng một hằng số hoặc nhân một số dương vào các biệt hàm. Sự mở rộng này, đôi lúc, giúp đơn giản hóa việc tính toán cũng như làm cho biệt hàm dễ hiểu và đơn giản hơn. Ví dụ, theo tiêu chuẩn cực tiểu hóa xác suất lỗi trung bình, các biệt hàm sau đều thỏa mãn

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{1.22}$$

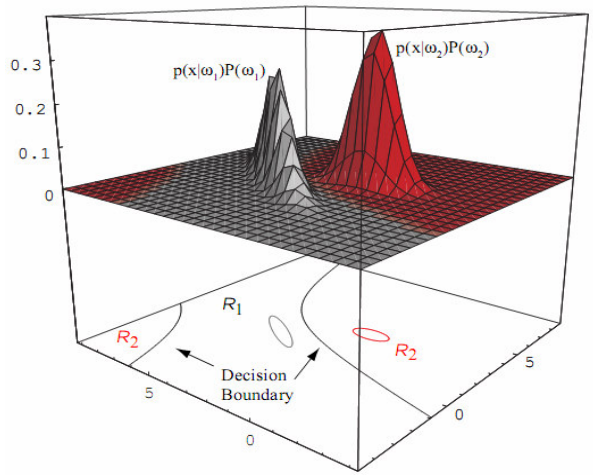
$$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i) \tag{1.23}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i) \tag{1.24}$$

Tuy nhiên, rõ ràng là biệt hàm ở (1.24) sẽ được tính toán đơn giản hơn so với ở (1.22).

III.1.2 *Vùng ra quyết định*

Dù được thể hiện như thế nào đi nữa thì mục tiêu cuối cùng của một hệ phân lớp vẫn là phân hoạch vùng không gian đặc trưng ra thành c vùng $R_1, R_2, ..., R_c$ trong đó $\mathbf{x} \in R_i$ khi và chỉ khi đối tượng có đặc trưng $\mathbf{x}$ được phân vào lớp ω_i. Chúng ta gọi $R_1, R_2, ..., R_c$ là các *vùng ra quyết định (decision region)*. Đường biên bao quanh các vùng ra quyết định được gọi là *biên ra quyết định (decision boundary)*. **Hình 4** cho thấy ví dụ về các vùng ra quyết định và biên ra quyết định.



Hình 4 R1 và R2 là 2 vùng ra quyết định. Biên ra quyết định là đường phân tách giữa các vùng ra quyết định.

III.2 Phân phối chuẩn

Phân phối chuẩn (Normal Distribution) đóng vai trò quan trọng trong bài toán phân lớp mẫu. Thực vậy, theo định lý hội tụ trung tâm (Central Limit Theorem) thì tổng kết hợp ảnh hưởng của một số lượng lớn các biến ngẫu nhiên nhỏ và độc lập sẽ dẫn tới phân phối chuẩn. Mặt khác, vì nhiều mẫu – từ cá, quả đến ký tự viết tay – đều có thể được xem như là mẫu biến dạng bởi một số lượng lớn các tiến trình ngẫu nhiên nên phân phối chuẩn là một mô hình tốt cho phân phối xác

suất thực sự. Chính vì tầm quan trọng của phân phối chuẩn nên phần này sẽ giới thiệu sơ lược về phân phối này. Ở phần sau sẽ xem xét việc sử dụng mô hình phân phối chuẩn cho bài toán phân lớp.

Phân phối chuẩn đơn biến (Univariate) Để cho đơn giản, trước tiên xét trường hợp đơn biến. Phân phối chuẩn được đặc trưng bởi hàm mật độ xác suất

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{1.25}$$

trong đó μ là *kỳ vọng (expected value)* (hay còn được gọi là *trung bình - mean*) của x được cho bởi

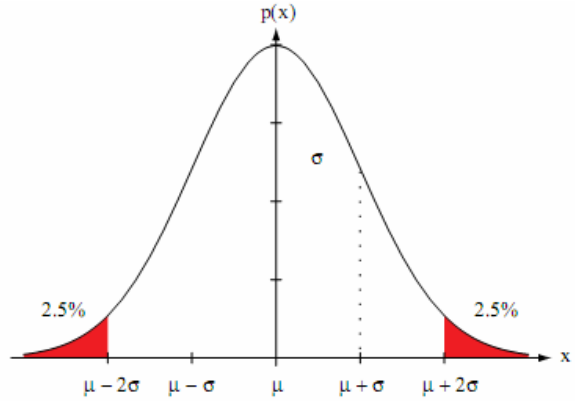
$$\mu \equiv E[x] = \int_{-\infty}^{\infty} xp(x)dx \tag{1.26}$$

và σ^2 là *phương sai (variance)* được cho bởi

$$\sigma^2 \equiv E\left[(x-\mu)^2 \right] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \tag{1.27}$$

Từ đẳng thức (1.25) chúng ta nhận thấy rằng phân phối chuẩn được đặc trưng bởi kỳ vọng μ và phương sai σ^2. Do đó, đôi khi, để đơn giản, người ta ký hiểu $p(x) \sim N(\mu, \sigma^2)$ để ám chỉ rằng x phân phối chuẩn theo kỳ vọng μ và phương sai σ^2.

Khi lấy mẫu từ phân phối chuẩn $N(\mu, \sigma^2)$ thì các mẫu có xu hướng tập trung ở quanh trung bình μ và trải ra xung quanh theo độ lệch chuẩn σ. **Hình 5** cho một ví dụ về phân phối chuẩn đơn biến.



Hình 5 Phân phối chuẩn đơn biến có xấp xỉ 95% **diện tích nằm trong khoảng** $|x-\mu| \le 2\sigma$ **. Đỉnh của phân phối này tại kỳ vọng** μ **và có giá trị**

$$p(\mu) = 1/\sqrt{2\pi}\sigma .$$

Phân phối chuẩn đa biến (Multivariate) Tổng quát hóa phân phối chuẩn đơn biến, ta có phân phối chuẩn đa biến trong không gian *d*-chiều với hàm mật độ như sau

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \qquad (1.28)$$

trong đó $\mathbf{x}$ là vector (cột) *d*-chiều, $\boldsymbol{\mu}$ là vector *d*-chiều kỳ vọng của $\mathbf{x}$ được cho bởi

$$\boldsymbol{\mu} \equiv E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \qquad (1.29)$$

$\mathbf{\Sigma}$ là ma trận *d* x *d hiệp phương sai (covariance matrix)* được cho bởi

$$\mathbf{\Sigma} \equiv E\left[(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\right] = \int (\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu}) p(\mathbf{x}) d\mathbf{x} \qquad (1.30)$$

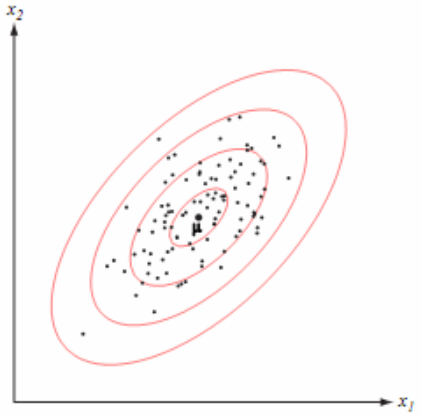
Tương tự như phân phối chuẩn đơn biến, phân phối chuẩn đa biến được đặc trưng bởi kỳ vọng $\boldsymbol{\mu}$ và hiệp phương sai $\mathbf{\Sigma}$, và do đó đôi khi người ta viết ngắn gọn là $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Xét $\mathbf{x} = [x_1, x_2, ..., x_d]^T$ thì từ đẳng thứ (1.29) được $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_d]^T$ với $\mu_i = E[x_i], i = 1..d$. Từ đẳng thứ (1.30) được $\boldsymbol{\Sigma}$ là ma trận có d dòng và d cột, giá trị $\sigma_{i,j}$ tại dòng thứ i cột thứ j là hiệp phương sai của x_i và x_j được cho bởi

$$\begin{aligned} \sigma_{i,j} &\equiv E[(x_i - \mu_i)(x_j - \mu_j)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_j - \mu_j)\,p(x_i, x_j)dx_i dx_j \end{aligned} \tag{1.31}$$

Từ đó dễ thấy rằng $\boldsymbol{\Sigma}$ là ma trận đối xứng.

Bây giờ, ta lấy mẫu theo phân phối $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Về mặt hình học, các mẫu sẽ nằm trong không gian thực d-chiều và chúng có xu hướng co cụm tại một vùng. Người ta chứng minh là cụm này có hình dáng là một hyperellipsoid có tâm tại $\boldsymbol{\mu}$, các trục nằm trên vector là các eigenvector của ma trận $\boldsymbol{\Sigma}$ với các bán kính là các eigenvalue tương ứng của ma trận $\boldsymbol{\Sigma}$. **Hình 6** cho một ví dụ trên không gian 2-chiều.



Hình 6 Các mẫu được lấy mẫu từ một phân phối chuẩn. Các mẫu này có xu hướng co cụm lại trong một vùng có tâm là kỳ vọng của phân phối. Các đường ellipse màu đỏ thể hiện sự bằng nhau của giá trị mật độ xác suất.

III.3 Biệt hàm cho phân phối chuẩn

Trong phần này, ta sẽ xây dựng hệ phân lớp theo biệt hàm với tiêu chuẩn là cực tiểu hóa xác suất lỗi trung bình. Ta sẽ xử dụng biệt hàm sau

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i),\ i = 1..c \tag{1.32}$$

cùng với giả thiết là $p(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Khi đó, từ (1.28) và (1.32) có được

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \tag{1.33}$$

3 trường hợp từ đơn giản đến tổng quát sau sau sẽ được xem xét.

III.3.1 *Trường hợp 1:* $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$

Trường hợp đơn giản nhất ở đây chính là khi các đặc trưng *đơn* độc lập với nhau về mặt thống kê và mỗi đặc trưng đơn có cùng phương sai σ^2. Vì vậy, ta suy ra được

$$|\boldsymbol{\Sigma}_i| = \sigma^{2d}, \quad \boldsymbol{\Sigma}_i^{-1} = \frac{1}{\sigma^2}\mathbf{I} \tag{1.34}$$

Suy ra $-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$ đều là hằng số đối với mỗi g_i. Do đó có biệt hàm đơn giản như sau

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)}{2\sigma^2} + \ln P(\omega_i) \tag{1.35}$$

Phân tích ra ta được

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \tag{1.36}$$

Do $\mathbf{x}^T\mathbf{x}$ là như nhau đối với mọi biệt hàm, nên ta có thể lược bỏ thành phần này để được biệt hàm tuyến tính đơn giản hơn

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[-2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] + \ln P(\omega_i) = \mathbf{w}_i^T\mathbf{x} + w_{i0} \qquad (1.37)$$

trong đó

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \qquad (1.38)$$

và

$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \ln P(\omega_i) \qquad (1.39)$$

Bây giờ, ta sẽ đi tìm biên ra quyết định giữa hai vùng ra quyết định có xác suất hậu định lớn nhất được đặc trưng bởi biệt hàm $g_i(.)$ và $g_j(.)$. Biên này được xác định bởi phương trình

$$g_i(\mathbf{x}) = g_j(\mathbf{x}) \qquad (1.40)$$

Suy ra

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \qquad (1.41)$$

trong đó

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \qquad (1.42)$$

và

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \qquad (1.43)$$

với $\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ là khoảng cách Euclide giữa hai điểm $\boldsymbol{\mu}_i$ và $\boldsymbol{\mu}_j$.

Phương trình (1.41) xác định một hyperplane Ω đi qua điểm $\mathbf{x}_0$ và vuông góc với vector $\mathbf{w}$. Từ (1.42) ta thấy rằng hyperplane Ω vuông góc với đường thẳng nối hai điểm $\boldsymbol{\mu}_i$ và $\boldsymbol{\mu}_j$. Xét hai trường hợp sau

- Nếu $P(\omega_i) = P(\omega_j)$: khi đó $\mathbf{x}_0 = (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)/2$, vì vậy mà hyperplane Ω sẽ vuông góc với đoạn nối hai điểm $\boldsymbol{\mu}_i$ và $\boldsymbol{\mu}_j$ ngay tại trung điểm. Trở lại với ví dụ phân loại táo-lê. Nếu như biết rằng số lượng táo và số lượng lê được đưa vào là bằng nhau thì rõ ràng là xác suất tiền định chẳng giúp ích được gì cho việc phân loại cả. Vì vậy, nếu quả đang xét có màu gần với màu trung bình của loại quả nào thì nên phân nó vào lớp đó.

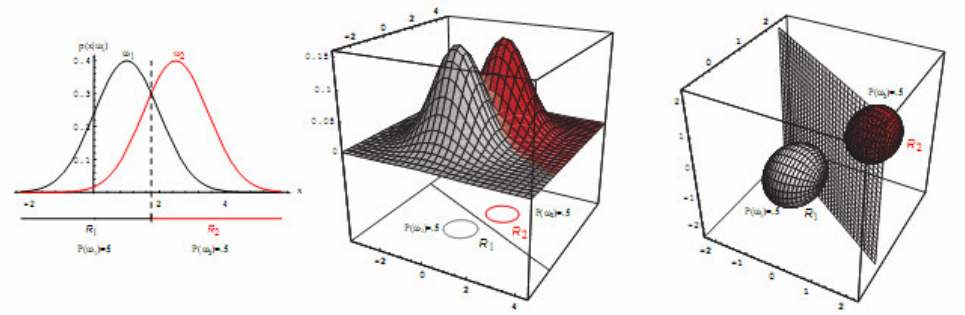
  Hình 7 cho một số ví dụ.

- Nếu $P(\omega_i) \neq P(\omega_j)$: khi đó điểm $\mathbf{x}_0$ sẽ dời ra xa khỏi vùng quyết định có xác suất tiền định lớn hơn. Điều đó cho thấy vùng có xác suất tiền định lớn hơn sẽ được mở rộng hơn. Bây giờ giả sử là *P(táo)=0.8 > P(lê) = 0.2*. Quả cần phân loại có màu nằm ở giữa độ trung bình màu của 2 loại quả. Khi này, nhờ biết về xác suất tiền định *P(táo) > P(lê)* nên suy ra được khả năng quả đang xét là táo sẽ cao hơn so với khả năng là lê.

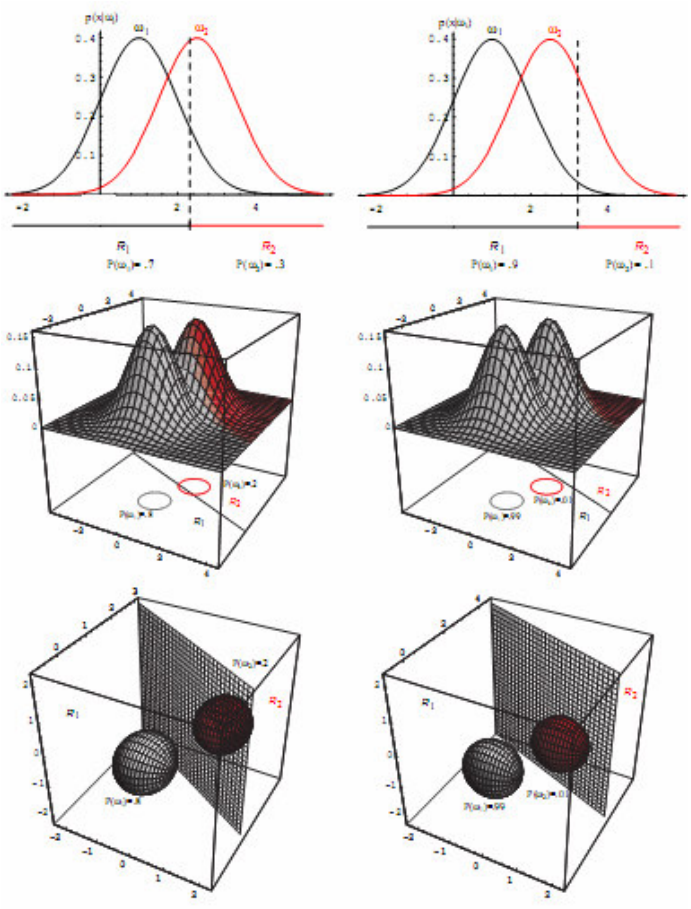
  Một điều đáng lưu ý là nếu $\sigma^2 \ll \left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|$ thì xác suất tiền định sẽ hầu như không gây tác động đến vị trí của biên ra quyết định.

  Hình 8 cho một số ví dụ.

Trong trường hợp đơn giản nhất là $P(\omega_i) = P(\omega_j), \forall i, j$, ta chỉ cần đơn giản là chọn lớp có kỳ vọng gần với $\mathbf{x}$ nhất. Hệ phân loại như vậy được gọi là *hệ phân loại bằng khoảng cách gần nhất (minimum distance classifier).*



Hình 7 Trường hợp hai hiệp phương sai của hai phân phối bằng nhau và tỷ lệ với ma trận I. Lúc này các phân phối được biểu diễn dưới dạng hình cầu trong không gian d-chiều và đường biên là một hyperplane của d-1 chiều vuông góc với đường nối 2 trung bình.



Hình 8 Trường hợp xác suất tiền định không bằng nhau. Biên ra quyết định không còn đi qua trung điểm đoạn nối 2 trung bình nữa.

Ví dụ:

Xét một ví dụ đơn giản sau. Ta cần xây dựng hệ phân lớp với 2 lớp ω_1, ω_2, đặc trưng x với likelihood sau

$$p(x|\omega_1) = N(4,1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-4)^2\right\}$$
$$p(x|\omega_2) = N(8,1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-8)^2\right\} \quad (1.44)$$

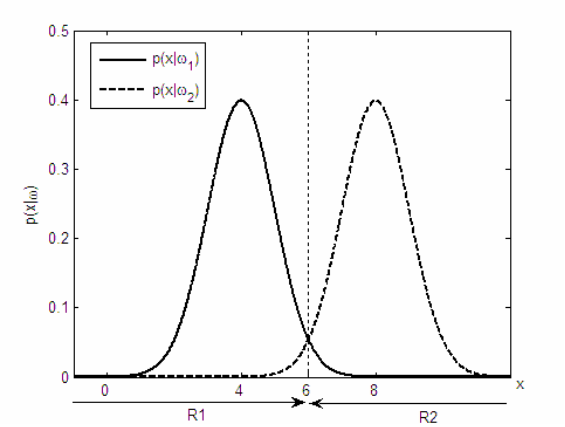
Từ (1.37) được 2 biệt hàm

$$\begin{aligned} g_1(x) &= 4x - 8 + \ln P(\omega_1) \\ g_2(x) &= 8x - 32 + \ln P(\omega_2) \end{aligned} \tag{1.45}$$

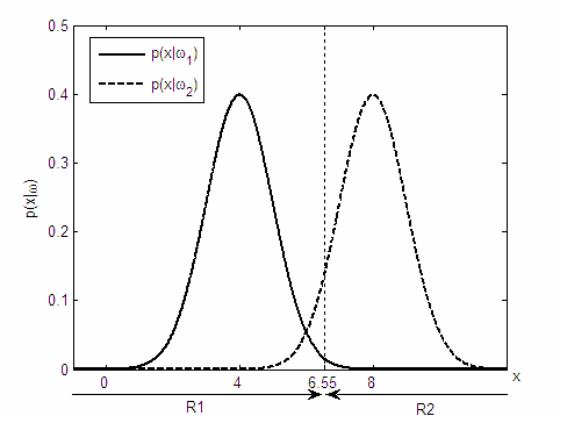
Suy ra điểm mốc phân biệt 2 vùng ra quyết định thỏa phương trình

$$\begin{aligned} & g_1(x) = g_2(x) \\ \Leftrightarrow\ & 4x - 8 + \ln P(\omega_1) = 8x - 32 + \ln P(\omega_2) \\ \Leftrightarrow\ & x = 6 + \frac{\ln P(\omega_1) - \ln P(\omega_2)}{4} \end{aligned} \tag{1.46}$$

Nếu $P(\omega_1) = P(\omega_2) = 0.5$ thì điểm mốc phân biệt 2 vùng ra quyết định là $x_0 = 6$. Nếu $P(\omega_1) = 0.9, P(\omega_2) = 0.1$ thì mốc phân biệt 2 vùng ra quyết định là $x_1 \simeq 6.55$. **Hình 9** và **Hình 10** cho thấy rõ về 2 trường hợp này.



Hình 9 Trường hợp $P(\omega_1) = P(\omega_2) = 0.5$, điểm phân biệt 2 vùng ra quyết định là $x_0 = 6$ bằng nằm ngay chính giữa 2 trung bình của 2 likelihood



Hình 10 Trường hợp $P(\omega_1) = 0.9, P(\omega_2) = 0.1$, điểm phân biệt 2 vùng ra quyết định nằm ở $x_1 \simeq 6.55$, chệch về phía điểm trung bình của likelihood của lớp có xác suất tiền định nhỏ hơn.

III.3.2 *Trường hợp 2:* $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

Ở trường hợp này, ma trận hiệp phương sai của mọi lớp đều như nhau và bất kỳ. Khi đó, vì $-\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$ đều là hằng số với mỗi g_i nên ta có được biệt hàm đơn giản hơn như sau

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i) \qquad (1.47)$$

Phân tích ra được

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \qquad (1.48)$$

Do $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ là như nhau đối với các biệt hàm nên ta có thể lược bỏ để có được biệt hàm tuyến tính đơn giản hơn

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[-2\boldsymbol{\mu}_i{}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu}_i{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right] + \ln P(\omega_i) = \mathbf{w}_i^T\mathbf{x} + w_{i0} \qquad (1.49)$$

trong đó

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \tag{1.50}$$

và

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{1.51}$$

Vì biệt hàm là tuyến tính nên biên ra quyết định sẽ là hyperplane. Nếu R_1, R_2 là hai vùng ra quyết định thì biên ra quyết định giữa chúng Ω được cho bởi phương trình

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \tag{1.52}$$

trong đó

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{1.53}$$

và

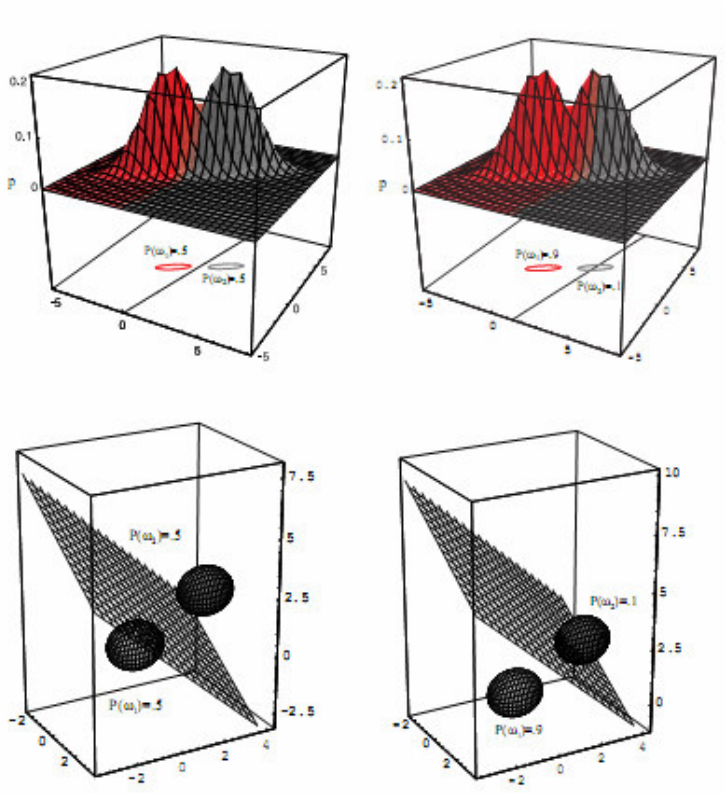
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{1.54}$$

Vì $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ thường sẽ không cùng hướng với vector $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ nên hyperplane Ω thường cũng sẽ không vuông góc với đường nối hai điểm $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$.

Tương tự với trường hợp 1, xét hai trường hợp sau

- Nếu $P(\omega_i) = P(\omega_j)$: khi đó $\mathbf{x}_0 = (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)/2$, vì vậy mà hyperplane Ω sẽ đi qua trung điểm đoạn nối hai điểm $\boldsymbol{\mu}_i$ và $\boldsymbol{\mu}_j$.
- Nếu $P(\omega_i) \neq P(\omega_j)$: khi đó điểm $\mathbf{x}_0$ sẽ dời ra xa khỏi vùng quyết định có xác suất tiền định lớn hơn. Điều đó cho thấy vùng có xác suất tiền định lớn hơn sẽ được mở rộng hơn.

Hình 11 cho một số ví dụ.



Hình 11 Trường hợp hiệp phương sai là bất kỳ nhưng bằng nhau với mọi phân phối.

III.3.3 *Trường hợp 3: Σ_i bất kỳ*

Đây là trường hợp tổng quát: các ma trận hiệp phương sai của các lớp là bất kỳ và không nhất thiết bằng nhau. Khi đó, chỉ có $-\frac{d}{2}\ln 2\pi$ là hằng số với mọi g_i. Lược bỏ phần đó đi ta được biệt hàm mới

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \qquad (1.55)$$

Phân tích ra được hàm bậc hai

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[\mathbf{x}^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right] - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (1.56)$$
$$= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

trong đó

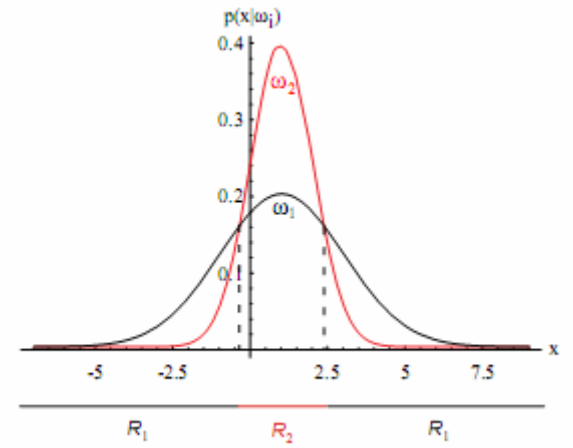
$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \quad (1.57)$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \quad (1.58)$$

và

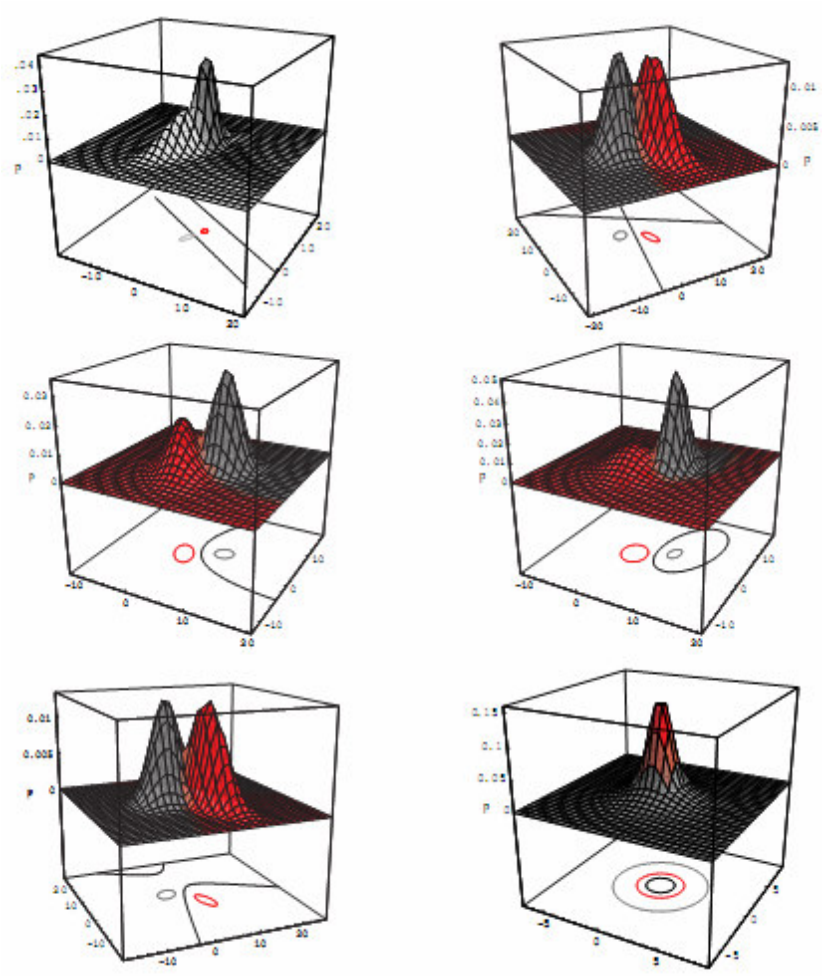
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \quad (1.59)$$

Khi này, biên ra quyết định là hyperquadric và có thể là bất kỳ dạng nạo: hyperplane, một cặp hyperplane, hypersphare, hyperellipsoid, hyperparaboloid và hyperhyperboloid. Thậm chí trong trường hợp *1*-chiều, với các hiệp phương sai bất kỳ, thì vùng ra quyết định cũng chưa chắc là liên thông (như được đưa ra trong **Hình 12**)

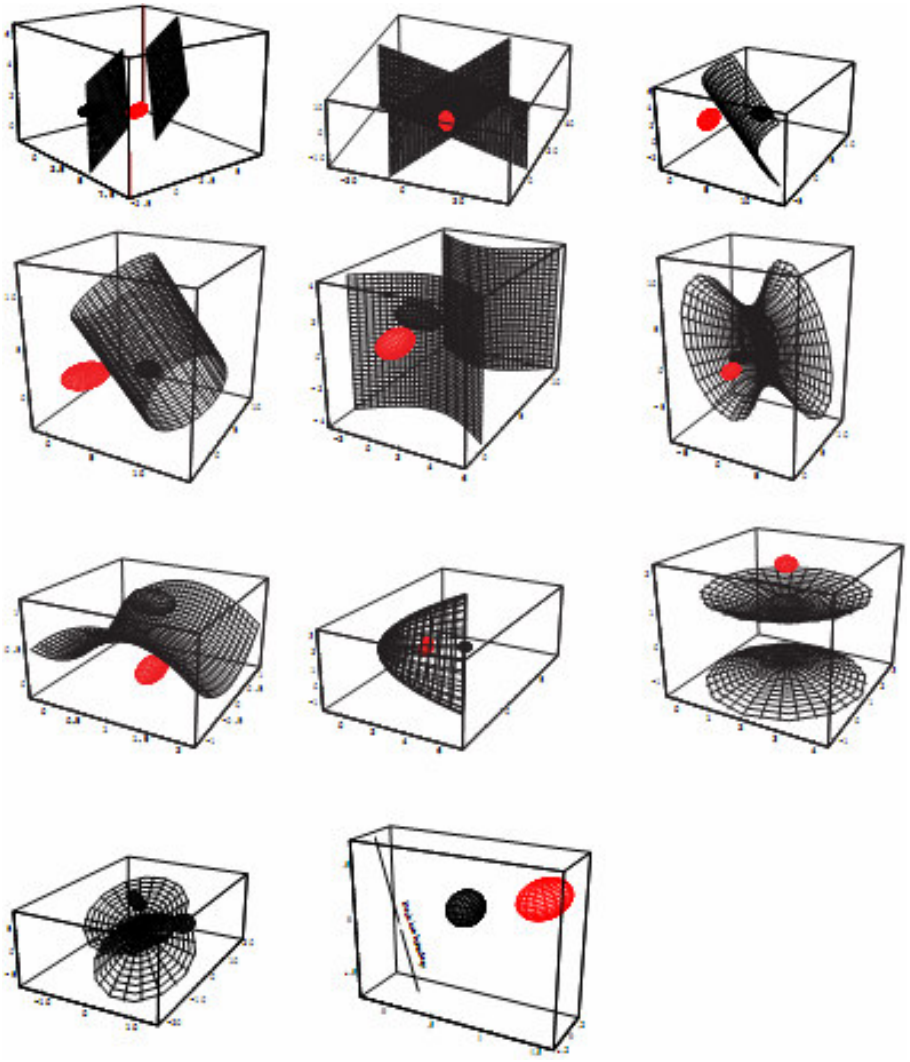


Hình 12 Vùng ra quyết định R1 được chia ra làm 2 bên không liên thông.

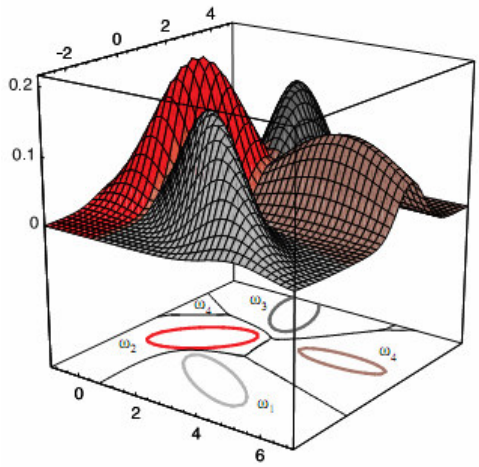
Hình 13 cho một số ví dụ về trường hợp *2*-chiều. **Hình 14** cho một số ví dụ về trường hợp *3*-chiều. **Hình 15** cho ví dụ về trường hợp có nhiều hơn hai lớp trong không gian *2*-chiều.



Hình 13 Trường hợp các hiệp phương sai khác nhau và bất kỳ trong không gian 2-chiều với 2 lớp.



Hình 14 Trường hợp các hiệp phương sai khác nhau và bất kỳ trong không gian 3-chiều với 2 lớp.



Hình 15 Trường hợp các hiệp phương sai khác nhau và bất kỳ trong không gian 2-chiều với 4 lớp.

Ví dụ:

Xét một ví dụ 1 chiều sau. Ta cần xây dựng hệ phân lớp với 2 lớp ω_1, ω_2 thỏa $P(\omega_1) = P(\omega_2) = 0.5$, đặc trưng x với likelihood sau

$$p(x|\omega_1) = N(0,3) = \frac{1}{\sqrt{2\pi}\sqrt{3}} \exp\left\{-\frac{1}{2}.\frac{x^2}{3}\right\}$$
$$p(x|\omega_2) = N(2,1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-2)^2\right\} \quad (1.60)$$

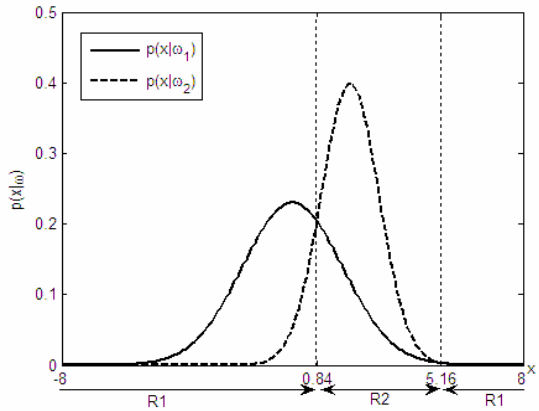
Từ (1.56) được 2 biệt hàm

$$g_1(x) = -\frac{1}{6}x^2 - \frac{1}{2}\ln 3 + \ln P(\omega_1)$$
$$g_2(x) = -\frac{1}{2}x^2 + 2x - 2 + \ln P(\omega_2) \quad (1.61)$$

Vì vùng ra quyết định R_1 thỏa $g_1(x) > g_2(x)$ tức là

$$-\frac{1}{6}x^2-\frac{1}{2}\ln 3+\ln P(\omega_1)>-\frac{1}{2}x^2+2x-2+\ln P(\omega_2)$$
$$\Leftrightarrow 2x^2-12x+12-3\ln 3+6\ln P(\omega_1)-6\ln P(\omega_2)>0 \qquad (1.62)$$
$$\Leftrightarrow x\in(-\infty,0.84)\cup(5.16,\infty)$$

nên $R_1=(-\infty,0.84)\cup(5.16,\infty)$. Từ đó dẫn tới $R_2=(0.84,5.16)$.



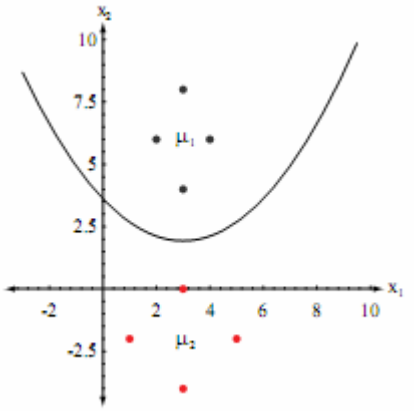
Hình 16 Vùng ra quyết định $R_1=(-\infty,0.84)\cup(5.16,\infty)$ và $R_2=(0.84,5.16)$.

Ví dụ:

Để làm rõ hơn nữa, hãy xét thêm một ví dụ 2 chiều như sau. Cho hai tập điểm như trong **Hình 17**, các điểm màu đen thuộc lớp ω_1, các điểm màu đỏ thuộc lớp ω_2. Giả sử là ta đã biết trước phân phối của hai tập điểm này là phân phối chuẩn với

$$\boldsymbol{\mu}_1=\begin{bmatrix}3\\6\end{bmatrix};\ \boldsymbol{\Sigma}_1=\begin{pmatrix}1/2 & 0\\ 0 & 2\end{pmatrix}\ \text{và}\ \boldsymbol{\mu}_2=\begin{bmatrix}3\\-2\end{bmatrix};\ \boldsymbol{\Sigma}_2=\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$$

và $P(\omega_1)=P(\omega_2)=0.5$.



Hình 17 Đường biên ra quyết định phân cách vùng ra quyết định được cho bởi hai tập điểm mẫu.

Từ (1.56), (1.57), (1.58) và (1.59) và cho $g_1(\mathbf{x}) = g_2(\mathbf{x})$ giải ra được phương trình biên ra quyết định như sau

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2 \qquad (1.63)$$

Phương trình này thể hiện một parabola đi qua đỉnh $\begin{pmatrix} 3 \\ 1.83 \end{pmatrix}$ như trong **Hình 17**.

IV. MỘT SỐ VẤN ĐỀ MỞ RỘNG

IV.1 Lý thuyết ra quyết định Bayes cho trường hợp đặc trưng rời rạc

Ở các phần trên, ta luôn giả thiết rằng $\mathbf{x}$ liên tục trong $\Re^d$. Trong phần này, ta sẽ xem xét với $\mathbf{x}$ chỉ nhận một trong m giá trị rời rạc $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m \in \Re^d$.

Khi này, ta sẽ thay hàm mật độ xác suất $p(\mathbf{x} \mid \omega_i)$ bằng hàm xác suất $P(\mathbf{x} \mid \omega_i)$, thay tích phân

$$\int p(\mathbf{x} \mid \omega_i) d\mathbf{x} \qquad (1.64)$$

bằng tổng lấy trên tất cả các giá trị rời rạc có thể có của $\mathbf{x}$

$$\sum_{\mathbf{x}} P(\mathbf{x} \mid \omega_i) \tag{1.65}$$

Định lý Bayes cũng sẽ phải thay đổi theo như sau

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_i) P(\omega_i)}{P(\mathbf{x})} \tag{1.66}$$

trong đó

$$P(\mathbf{x}) = \sum_{i=1}^{c} P(\mathbf{x} \mid \omega_i) P(\omega_i) \tag{1.67}$$

Công thức về rủi ro có điều kiện $R(\alpha_j \mid \mathbf{x})$ (1.8) sẽ không thay đổi. Vì vậy mà Luật 2 và Luật 3 cũng sẽ được bảo toàn.

Ví dụ:

Xét một ví dụ khi đặc trưng là rời rạc sau. Người ta được biết là 1% dân số bị một căn bệnh đang được tìm hiểu. Với phương pháp kiểm tra máu thì xác suất để một người bị bệnh được phát hiện có bệnh là 97%. Tuy nhiên, với phương pháp này, một người không mắc bệnh có 6% khả năng là bị chẩn đoán có bệnh.

Với ví dụ này, ta có 2 lớp ω

- *disease (bị bệnh)* là lớp những người bị bệnh với $P(\text{disease}) = 0.01$, và
- *non_disease (không bị bệnh)* là lớp những người không bị bệnh với $P(\text{non\_disease}) = 0.99$,

Đặc trưng x ở đây là kết quả chẩn đoán bệnh bằng phương pháp xét nghiệm máu, có thể nhận một trong 2 giá trị

- *positive (dương tính)*: xét nghiệm máu chẩn đoán bị bệnh,
- *negative (âm tính)*: xét nghiệm máu chẩn đoán không mắc bệnh.

Từ dữ kiện đã cho suy ra được likelihood $P(x \mid \omega)$ với các giá trị được cho trong bảng sau

	disease	***non_disease***
positive	0.97	0.06
negative	0.03	0.94

Xác suất để một người bất kỳ được chẩn đoán có bệnh và không có bệnh là

$$\begin{aligned} &P(\text{positive}) \\ &= P(\text{positive} \mid \text{disease}) P(\text{disease}) \\ &+ P(\text{positive} \mid \text{non\_disease}) P(\text{non\_disease}) \\ &= 0.97 \times 0.01 + 0.06 \times 0.99 = 0.0691 \\ &P(\text{negative}) = 1 - P(\text{positive}) = 1 - 0.0691 = 0.9309 \end{aligned} \tag{1.68}$$

Áp dụng định lý Bayes rời rạc ở trên suy ra được các xác suất hậu định sau

$$\begin{aligned} P(\text{disease} \mid \text{positive}) &= \frac{P(\text{positive} \mid \text{disease}) P(\text{disease})}{P(\text{positive})} \\ &= \frac{0.97 \times 0.01}{0.0691} \simeq 0.1404 \\ P(\text{non\_disease} \mid \text{positive}) &= 1 - 0.1404 = 0.8596 \end{aligned} \tag{1.69}$$

$$\begin{aligned} P(\text{disease} \mid \text{negative}) &= \frac{P(\text{negative} \mid \text{disease}) P(\text{disease})}{P(\text{negative})} \\ &= \frac{0.03 \times 0.01}{0.9309} \simeq 0.0003 \\ P(\text{non\_disease} \mid \text{negative}) &= 1 - P(\text{disease} \mid \text{negative}) = 0.9997 \end{aligned} \tag{1.70}$$

Điều đó có nghĩa là một người đi xét nghiệm máu cho kết quả positive thì khả năng người đó bị bệnh là 14.04% và không bị bệnh là 85.96%. Nếu áp dụng Luật 3 thì rõ ràng là dù cho xét nghiệm máu là dương tính thì người đó vẫn được cho là không có bệnh. Nguyên nhân chính

là do xác suất tiền định $P(\text{disease}) = 0.01 \ll P(\text{non\_disease}) = 0.99$. Quyết định dựa trên Luật 3 lúc này rõ ràng là không hợp lý.

Bây giờ, khi bác sĩ nhận được kết quả xét nghiệm máu thì ông ấy có thể có một trong 3 hành động α sau

- *sick*: quyết định người được xét nghiệm là có bệnh,
- *healthy*: quyết định người được xét nghiệm là không có bệnh
- *test*: quyết định người được xét nghiệm cần được kiểm tra thêm

Nếu hành động của bác sỹ là đúng với tình trạng bệnh của người được xét nghiệm thì sẽ không phải trả giá gì cả. Tuy nhiên, nếu ngược lại thì giá phải trả sẽ là như sau

- nếu người được xét nghiệm có bệnh mà bác sỹ quyết định không có bệnh thì giá phải trả sẽ là 100,
- nếu người được xét nghiệm không có bệnh mà bác sỹ quyết định là có bệnh thì giá phải trả là 10,
- mặc cho người được xét nghiệm có bệnh hay không có bệnh, nếu bác sỹ quyết định cần có thêm kiểm tra thì giá phải trả là 5.

Khi đó hàm tiêu tốn $\lambda(\alpha \mid \omega)$ có giá trị được cho trong bảng sau

	sick	***healthy***	***test***
disease	0	100	5
non_disease	10	0	5

Nếu xét nghiệm máu cho kết quả dương tính thì

$$\begin{aligned} R(\text{sick}\,|\,\text{positive}) &= \lambda(\text{sick} \mid \text{disease})P(\text{disease} \mid \text{positive}) \\ &\quad + \lambda(\text{sick} \mid \text{non\_disease})P(\text{non\_disease} \mid \text{positive}) \\ &= 0 \times 0.1404 + 10 \times 0.8596 = 8.596 \end{aligned} \tag{1.71}$$

$$\begin{aligned} R(\text{healthy}\,|\,\text{positive}) &= \lambda(\text{healthy} \mid \text{disease})P(\text{disease} \mid \text{positive}) \\ &\quad + \lambda(\text{healthy} \mid \text{non\_disease})P(\text{non\_disease} \mid \text{positive}) \\ &= 100 \times 0.1404 + 0 \times 0.8596 = 14.04 \end{aligned} \tag{1.72}$$

$$R(\text{test}\,|\,\text{positive}) = \lambda(\text{test}\,|\,\text{disease})P(\text{disease}\,|\,\text{positive}) + \lambda(\text{test}\,|\,\text{non\_disease})P(\text{non\_disease}\,|\,\text{positive}) = 5\times 0.1404 + 5\times 0.8596 = 5 \quad (1.73)$$

Nếu xét nghiệm máu cho kết quả âm tính thì

$$R(\text{sick}\,|\,\text{negative}) = \lambda(\text{sick}\,|\,\text{disease})P(\text{disease}\,|\,\text{negative}) + \lambda(\text{sick}\,|\,\text{non\_disease})P(\text{non\_disease}\,|\,\text{negative}) = 0\times 0.0003 + 10\times 0.9997 = 9.997 \quad (1.74)$$

$$R(\text{healthy}\,|\,\text{negative}) = \lambda(\text{healthy}\,|\,\text{disease})P(\text{disease}\,|\,\text{negative}) + \lambda(\text{healthy}\,|\,\text{non\_disease})P(\text{non\_disease}\,|\,\text{negative}) = 100\times 0.0003 + 0\times 0.9997 = 0.03 \quad (1.75)$$

$$R(\text{test}\,|\,\text{negative}) = \lambda(\text{test}\,|\,\text{disease})P(\text{disease}\,|\,\text{negative}) + \lambda(\text{test}\,|\,\text{non\_disease})P(\text{non\_disease}\,|\,\text{negative}) = 5\times 0.0003 + 5\times 0.9997 = 5 \quad (1.76)$$

Như vậy, xét Luật 2, nếu kết quả xét nghiệm máu là dương tính thì bác sỹ nên kiểm tra thêm. Ngược lại, nếu kết quả xét nghiệm là âm tính thì bác sỹ nên quyết định người đó là không bị bệnh.

IV.2 Đặc trưng bị thiếu và biến dạng bởi nhiễu

Lý thuyết ra quyết định Bayes là một công cụ tối ưu khi ta đã biết đầy đủ các xác suất trong bài toán. Bây giờ, ta giả sử rằng việc xây dựng hệ phân lớp dựa vào một tập dữ liệu tốt (không bị sai lạc) nhưng dữ liệu đưa vào để kiểm tra thì bị sai lạc ở một khía cạnh nào đó mà ta đã biết. Vấn đề đặt ra là làm cách nào để giảm thiểu lỗi?

Trong phần này sẽ xét đến hai trường hợp: *đặc trưng bị thiếu* và *đặc trưng bị biến dạng bởi nhiễu có tính chất biết trước*. Phương pháp tiếp cận cơ bản là sẽ phục hồi lại càng nhiều càng tốt thông tin về phân phối cơ bản và sử dụng luật ra quyết định Bayes.

IV.2.1 *Đặc trưng bị thiếu*

Trở lại ví dụ phân loại táo-lê. Giả sử đang sử dụng hai đặc trưng là màu và đường viền (đường viền cho biết hình dạng của quả). Ta sẽ phải làm gì khi quả trong ảnh đang xét bị che mất một phần (ví dụ như phần cuống)? Rõ ràng là đặc trưng đường viền đã bị thiếu và không thể sử dụng để phân loại.

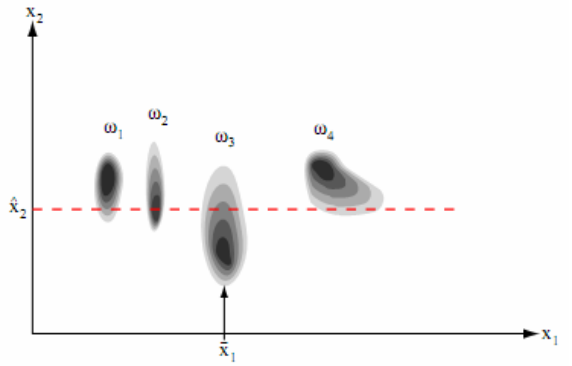
Bây giờ, xét đến một trường hợp cụ thể hơn với bốn lớp như trong **Hình 18.** Giả sử rằng đặc trưng x_1 bị thiếu và đặc trưng x_2 có giá trị là $\hat{x}_2$. Rõ ràng là nếu áp đặt cho đặc trưng x_1 lấy giá trị trung bình là $\overline{x}_1$ thì đối tượng đang xét sẽ được xếp vào lớp ω_3. Tuy nhiên, nếu so sánh về giá trị $p(\hat{x}_2 \mid \omega_i), i = 1..4$ và lấy cái cao nhất thì nên xếp vào lớp ω_2.

Ký hiệu vector đặc trưng là $\mathbf{x} = \left[\mathbf{x}_g \; \mathbf{x}_b\right]^T$ với $\mathbf{x}_g$ là vector đặc trưng tốt (không bị thiếu), $\mathbf{x}_b$ là vector đặc trưng xấu (bị thiếu). Một điều rõ ràng là không còn có thể ra quyết định dựa trên $P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b)$ được nữa, và vì vậy ra quyết định dựa trên $P(\omega_i \mid \mathbf{x}_g)$ là một ý kiến hợp lý. Ta có

$$\begin{aligned} P(\omega_i \mid \mathbf{x}_g) &= \frac{p(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}{p(\mathbf{x}_g)} \\ &= \frac{\int P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b) p(\mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}{p(\mathbf{x}_g)} \\ &= \frac{\int g_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}_b}{\int p(\mathbf{x}) d\mathbf{x}_b} \end{aligned} \qquad (1.77)$$

trong đó $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b)$ chính là một dạng của biệt hàm. Nói một cách ngắn gọn, ta sẽ lấy tích phân trên toàn bộ miền không gian của đặc trưng xấu.

Khi đã có $P(\omega_i \mid \mathbf{x}_g)$, ta sẽ áp dụng Luật 3 để phân lớp.



Hình 18 Ví dụ với hai đặc trưng và có bốn phân lớp. Các phân lớp đều có xác suất tiền định bằng nhau. Phân phối có điều kiện $p(\mathbf{x} \mid \omega_i)$ được cho bởi vùng xám đen (càng tối thì mật độ xác suất càng lớn).

IV.2.2 *Đặc trưng bị biến dạng bởi nhiễu*

Cũng với ví dụ phân loại táo-lê, nhưng bây giờ đặc trưng đường viền là tốt. Tuy nhiên, đặc trưng màu hay bị nhiễu do nhiều nguyên nhân: nguồn sáng bên ngoài, nhiễu từ sensor,… Những nhiễu này có thể sẽ gây tác động tiêu cực đến hệ phân loại. Vấn đề đặt ra là phải xây dựng được hệ phân loại có khả năng chịu được nhiễu.

Cũng với ký hiệu đặc trưng như ở phần trên nhưng với $\mathbf{x}_b$ là đặc trưng quan sát được đã bị biến dạng bởi nhiễu và $\mathbf{x}_t$ là giá trị thật của đặc trưng đó. Khi đó $p(\mathbf{x}_b \mid \mathbf{x}_t)$ thể hiện *mô hình nhiễu (noise model)*. Giả định rằng nếu đã biết $\mathbf{x}_t$ thì $\mathbf{x}_b$ độc lập với ω_i và $\mathbf{x}_g$. Từ đó được

$$P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b) = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) d\mathbf{x}_t}{p(\mathbf{x}_g, \mathbf{x}_b)}$$
$$= \frac{\int P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_b \mid \mathbf{x}_t) d\mathbf{x}_t}{\int p(\mathbf{x}_g, \mathbf{x}_t) p(\mathbf{x}_b \mid \mathbf{x}_t) d\mathbf{x}_t} \qquad (1.78)$$
$$= \frac{\int g_i(\mathbf{x}) p(\mathbf{x}) p(\mathbf{x}_b \mid \mathbf{x}_t) d\mathbf{x}_t}{\int p(\mathbf{x}) p(\mathbf{x}_b \mid \mathbf{x}_t) d\mathbf{x}_t}$$

trong đó $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}_g, \mathbf{x}_b)$ là một dạng của biệt hàm.

Để ý thấy là đẳng thức (1.78) khác với (1.77) ở chỗ tích phân được đánh trọng số tương ứng với mô hình nhiễu. Trong trường hợp xấu nhất khi $p(\mathbf{x}_b \mid \mathbf{x}_t)$ là đồng nhất trên toàn bộ không gian đặc trưng thì ta sẽ có được trường hợp tương tự với đặc trưng bị thiếu.

IV.3 Lý thuyết ra quyết định kết hợp Bayes và Ngữ cảnh

Trong các phần trên chỉ xem xét đến việc phân loại đối tượng dựa vào đặc trưng của chính bản thân nó mà không quan tâm đến những đối tượng đã được phân loại trước đó như thế nào. Đôi khi, chính những thông tin như vậy lại hữu ích. Vì vậy mà phần này sẽ xem xét đến tình huống là: lớp của đối tượng đang xét có mối quan hệ với lớp của những đối tượng đã xét trước đó. Ta xem thử có thể khai thác gì từ thông tin ràng buộc đó không. Ta gọi sự ràng buộc này là *ngữ cảnh (context)*. Trở lại ví dụ táo-lê. Ta sẽ chờ cho đủ n quả đi vào dây chuyền phân loại và sẽ đồng loạt phân lớp cho n quả đó.

Ký hiệu $\boldsymbol{\omega} = (\omega(1), \omega(2), ..., \omega(n))^T$ là vector xác định n lớp cho n đối tượng, trong đó mỗi $\omega(i)$ nhận một trong các giá trị $\omega_1, \omega_2, ..., \omega_c$. Ký hiệu $P(\boldsymbol{\omega})$ là xác suất tiền định cho cho n lớp. Ký hiệu $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ là ma trận cho biết n vector đặc trưng quan sát được từ n đối tượng. Cuối cùng, ký hiệu $p(\mathbf{X} \mid \boldsymbol{\omega})$ là xác suất có điều

kiện của $\mathbf{X}$ khi có được sự phân lớp $\boldsymbol{\omega}$. Áp dụng định lý Bayes được xác suất hậu định

$$P(\boldsymbol{\omega} \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \boldsymbol{\omega})P(\boldsymbol{\omega})}{p(\mathbf{X})} = \frac{p(\mathbf{X} \mid \boldsymbol{\omega})P(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega}} p(\mathbf{X} \mid \boldsymbol{\omega})P(\boldsymbol{\omega})} \tag{1.79}$$

Để phân lớp trong trường hợp này, tương tự như phương pháp ở phần 3, ta sẽ xây dựng được công thức rủi ro có điều kiện kết hợp và rủi ro toàn bộ kết hợp. Khi đó tương tự với Luật 2: ta tìm $\boldsymbol{\omega}$ làm cực tiểu hóa rủi ro có điều kiện kết hợp.

Về mặt lý thuyết thì đơn giản như vậy, tuy nhiên, vấn đề nảy sinh ở đây chính là chi phí thực hiện (1.79) trong thực tế: việc tính $P(\boldsymbol{\omega} \mid \mathbf{X})$ cần có lượng tính toán khổng lồ vì có c lớp nên $\boldsymbol{\omega}$ có thể nhận c^n giá trị. Vì vậy cần phải có những phương pháp để tránh việc phải tính $P(\boldsymbol{\omega} \mid \mathbf{X})$ với cả c^n giá trị của $\boldsymbol{\omega}$.

V. KẾT LUẬN

Nội dung chính trong chương này chính là lý thuyết ra quyết định Bayes. Ý tưởng của lý thuyết ra quyết định Bayes rất đơn giản: để cực tiểu hóa rủi ro toàn bộ, ta sẽ chọn hành động cực tiểu hóa rủi ro có điều kiện $R(\alpha \mid \mathbf{x})$. Trong trường hợp đặc biệt khi quan tâm đến cực tiểu hóa xác suất lỗi trung bình, ta nên chọn lớp làm cực đại hóa xác suất hậu định $P(\omega_i \mid \mathbf{x})$. Để tính xác suất hậu định $P(\omega_i \mid \mathbf{x})$, định lý Bayes là một công cụ hữu hiệu khi biết trước xác suất tiền định $P(\omega_i)$ và mật độ xác suất có điều kiện $p(\mathbf{x} \mid \omega_i)$.

Chương này cũng đã xem xét hệ phân lớp được biểu diễn bằng các biệt hàm. Trong đó, nếu áp dụng lý thuyết ra quyết định Bayes thì biệt hàm có thể được biểu diễn bằng $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$ hoặc $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$ tùy theo tiêu chuẩn rủi ro.

Bên cạnh đó, một số vấn đề mở rộng cũng đã được bàn tới: đặc trưng bị thiếu, đặc trưng bị biến dạng bởi nhiễu và lý thuyết ra quyết định kết hợp Bayes.

VI. BÀI TẬP

ĐỊNH LÝ BAYES

Bài 1

Xét với bài toán phân 2-lớp $(A, \overline{A})$ với đặc trưng nhị phân $(x, \overline{x})$. Giả sử xác suất tiền định $P(A) = 0.33$. Người ta lấy mẫu với kết quả phân phối cho trong bảng sau

	A	$\overline{A}$
x	248	167
$\overline{x}$	82	503

Áp dụng định lý Bayes tính xác suất hậu định của mỗi lớp.

Giải:

$$P(x) = P(A)P(x \mid A) + P(\overline{A})P(x \mid \overline{A})$$

$$P(x \mid A) = \frac{248}{248+82} \approx 0.7515$$

$$P(x \mid \overline{A}) \approx 0.2493, \quad P(\overline{A}) = 1 - P(A) = 0.67$$

$$P(A \mid x) = \frac{P(x \mid A)P(A)}{P(x)} \approx 0.5976$$

Tương tự

$$P(\overline{A} \mid x) \approx 0.4024, \quad P(A \mid \overline{x}) \approx 0.1402, \quad P(\overline{A} \mid \overline{x}) \approx 0.8598$$

Bài 2

Một ngân hàng xếp khách hàng ra 2 loại: tính dụng xấu và tín dụng tốt. Dựa trên thông tin trong quá khứ, ngân hàng thấy rằng 1% tính dụng tốt và 10% tín dụng xấu rút tiền quá số tiền gửi trong 1 tháng bất kỳ. Một khách hàng mới đến mở tài khoản tại ngân hàng này. Ngân hàng xác định rằng khả năng để khách hàng này trở thành tín dụng tốt là 70%.

1. Trong tháng đầu tiên, khách hàng này rút quá số tiền gửi. Hỏi ngân hàng sẽ xác định lại khả năng để khách hàng này trở thành tín dụng tốt là bao nhiêu?
2. Đến cuối tháng thứ 2, nếu khách hàng này không rút quá số tiền gửi thì ngân hàng sẽ xác định lại khả năng để khách hàng này trở thành tín dụng tốt là bao nhiêu?

Giải:

1. Gọi G là sự kiện khách hàng được xếp loại tín dụng tốt. O là sự kiện khách hàng rút quá số tiền gửi. Từ dữ kiện đề có được

$$P(O \mid G) = 0.01 \quad P(O \mid \bar{G}) = 0.1 \quad P(G) = 0.7$$

Áp dụng định lý Bayes

$$P(G \mid O) = \frac{P(O \mid G)P(G)}{P(O)} = \frac{P(O \mid G)P(G)}{P(O \mid G)P(G) + P(O \mid \bar{G})P(\bar{G})} \approx 0.189$$

2. Bước sang tháng thứ 2 thì $P(G) = 0.189$.

$$P(G \mid \bar{O}) = \frac{P(\bar{O} \mid G)P(G)}{P(\bar{O})} = \frac{P(\bar{O} \mid G)P(G)}{P(\bar{O} \mid G)P(G) + P(\bar{O} \mid \bar{G})P(\bar{G})} \approx 0.204$$

Bài 3

Trong chẩn đoán bệnh, một biến ngẫu nhiên A có thể có 2 giá trị là 0 (nghĩa là không có bệnh) và 1 (nghĩa là có bệnh). Biến ngẫu nhiên B thể hiện kết quả xét nghiệm, trong đó B nhận giá trị 0 (âm tính) và 1 (dương tính). Giả thiết rằng, dựa trên các hồ sơ y khoa, ta đã biết

$$P(B = 1 \mid A = 0) = 0.01, \; P(B = 1 \mid A = 1) = 0.9, \; P(A = 1) = 0.001$$

1. Sử dụng định lý Bayes hãy tính xác suất để một người có bệnh nếu biết rằng (a) kết quả xét nghiệm là âm tính và (b) kết quả xét nghiệm là dương tính.
2. Tương tự như câu trên nếu $P(A = 1) = 0.2$. Nhận xét xem $P(A \mid B)$ thay đổi thế nào.

Giải:

1. Từ dữ kiện có

$$P(B=1|A=0)=0.01 \qquad P(B=1|A=1)=0.9$$
$$P(B=0|A=0)=0.99 \qquad P(B=0|A=1)=0.1$$
$$P(A=1)=0.001 \qquad P(A=1)=0.999$$

Tính

$$P(B=0)=\sum_{i=0}^{1}P(B=0|A=i)P(A=i)=0.98911$$
$$P(B=1)=1-P(B=0)=0.01089$$

Áp dụng định lý Bayes được

$$P(A=1|B=0)=\frac{P(B=0|A=1)P(A=1)}{P(B=0)}=1.10^{-4}$$
$$P(A=1|B=1)=\frac{P(B=1|A=1)P(A=1)}{P(B=1)}=0.082645$$

2. Tương tự như trên ta được

$$P(A=1|B=0)=0.0246 \quad P(A=1|B=1)=0.957$$

Nhận xét: trong (1), do khả năng người bị bệnh là quá thấp ($P(A=1)=0.001$) dẫn tới cả 2 xác suất $P(A=1|B=0)$ và $P(A=1|B=1)$ đều thấp. Khi căn bệnh là phổ biến (ở (2)) thì 2 xác suất $P(A=1|B=0)$ và $P(A=1|B=1)$ cũng được nâng lên.

Bài 4

Trong 100000 người thì có khoảng 1 người bị nhiễm một loại virus nguy hiểm cần được chữa trị. Có 3 phòng thí nghiệm, được gọi là A, B và C, xét nghiệm về loại virus này. Mỗi xét nghiệm đưa ra 1 trong 2 kết quả: '+' nếu nguy cơ người được xét nghiệm có nguy cơ nhiễm cao

và '-' nếu nguy cơ thấp. Từ nhiều nghiên cứu, người ta xác định độ nhạy cảm (xác suất '+' khi người được xét nghiệm nhiễm virus) và độ đặc trưng (xác suất '-' khi người đó khỏe mạnh) của các xét nghiệm từ 3 phòng thí nghiệm trên là

Xét nghiệm	Độ nhạy cảm	Độ đặc trưng
A	0.8	0.9
B	0.9	0.8
C	0.8	0.5

Một người được xét nghiệm từ 3 phòng thí nghiệm A, B, C trên với kết quả lần lượt là '+-+'. Hãy xác định xác suất người đó bị nhiễm virus.

Giải:

Gọi S là biến ngẫu nhiên thể hiện trạng thái của người đó với 2 giá trị: 1 nếu người đó bị nhiễm virus, 0 nếu người đó khỏe mạnh; thì

$$P(S=0)=0.99999 \qquad P(S=1)=0.00001$$

Gọi $\mathbf{X}=[X_1,X_2,X_3]^T$ là vector đặc trưng ngẫu nhiên với $X_i, i=1,2,3$ lần lượt là biến ngẫu nhiên thể hiện kết quả xét nghiệm của phòng thí nghiệm A, B và C. Trong đó $X_i=0$ nếu kết quả là '-' và 1 nếu kết quả là '+'. Rõ ràng là độ nhạy cảm của xét nghiệm được cho bởi $P(X_i=1|S=1)$ và độ đặc trưng được cho bởi $P(X_i=0|S=0)$.

Do 3 phòng thí nghiệm xét nghiệm một cách độc lập nên có thể giả thiết là kết quả xét nghiệm từ 3 phòng trên là độc lập. Do đó

$$P(\mathbf{X}=\mathbf{x},S=0)=P(S=0)\prod_{i=1}^{3}P(X_i=x_i\mid S=0)=0.0399996$$

$$P(\mathbf{X}=\mathbf{x},S=1)=P(S=1)\prod_{i=1}^{3}P(X_i=x_i\mid S=1)=0.00000064$$

Suy ra

$$P(S=1|\mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x}, S=1)}{P(\mathbf{X}=\mathbf{x})}$$
$$= \frac{P(\mathbf{X}=\mathbf{x}, S=1)}{P(\mathbf{X}=\mathbf{x}, S=1) + P(\mathbf{X}=\mathbf{x}, S=0)} \approx 1.6\times 10^{-5}$$

LÝ THUYẾT RA QUYẾT ĐỊNH

Bài 5

Xét bài toán phân 2 lớp, 2 hành động với hàm tiêu tốn $\lambda_{11} = \lambda_{22} = 0$ và $\lambda_{12} = 10, \lambda_{21} = 1$. Hãy xây dựng luật quyết định tối ưu.

Giải:

Chọn hành động ω_1 nếu

$$(\lambda_{21} - \lambda_{11}) p(\mathbf{x}|\omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x}|\omega_2) P(\omega_2)$$
$$\Leftrightarrow p(\mathbf{x}|\omega_1) P(\omega_1) > 10\, p(\mathbf{x}|\omega_2) P(\omega_2)$$
$$\Leftrightarrow \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{10 P(\omega_2)}{P(\omega_1)}$$

ngược lại chọn ω_2.

Bài 6

Xét bài toán phân 2 lớp.

1. Với các hàm mật độ bất kỳ, nếu lấy $P(error|x) = 2P(\omega_1|x)P(\omega_2|x)$ thì có được chặn trên của trung bình xác suất lỗi.
2. Chứng minh rằng nếu lấy $P(error|x) = \alpha P(\omega_1|x)P(\omega_2|x)$ với $\alpha < 2$ thì không đảm bảo được chặn trên của trung bình xác suất lỗi.

Giải:

1.

$$\begin{aligned} P(error \mid x) &= 2P(\omega_1 \mid x)P(\omega_2 \mid x) \\ &= 2\min(P(\omega_1 \mid x)P(\omega_2 \mid x)).\max(P(\omega_1 \mid x)P(\omega_2 \mid x)) \\ &= 2.(1-\min(P(\omega_1 \mid x)P(\omega_2 \mid x)).\min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \\ &\geq \min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \end{aligned}$$

2. Nếu $\alpha < 2$ thì không đảm bảo $\alpha(1-\min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \geq 1$.

Bài 7

Xét bài toán phân 2 lớp.

1. Với các hàm mật độ bất kỳ, nếu lấy $P(error \mid x) = P(\omega_1 \mid x)P(\omega_2 \mid x)$ thì có được chặn dưới của trung bình xác suất lỗi.
2. Chứng minh rằng nếu lấy $P(error \mid x) = \beta P(\omega_1 \mid x)P(\omega_2 \mid x)$ với $\beta > 1$ thì không đảm bảo được chặn dưới của trung bình xác suất lỗi.

Giải:

1.

$$\begin{aligned} P(error \mid x) &= P(\omega_1 \mid x)P(\omega_2 \mid x) \\ &= \min(P(\omega_1 \mid x)P(\omega_2 \mid x)).\max(P(\omega_1 \mid x)P(\omega_2 \mid x)) \\ &= (1-\min(P(\omega_1 \mid x)P(\omega_2 \mid x)).\min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \\ &\leq \min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \end{aligned}$$

2. Nếu $\beta > 1$ thì không đảm bảo $\beta(1-\min(P(\omega_1 \mid x)P(\omega_2 \mid x)) \leq 1$.

Bài 8

Đặt $\omega_{\max}(\mathbf{x})$ là lớp thỏa $P(\omega_{\max} \mid \mathbf{x}) \geq P(\omega_i \mid \mathbf{x}), \forall i = 1,...,c$.

1. Chứng minh $P(\omega_{\max} \mid \mathbf{x}) \geq 1/c$
2. Chứng minh rằng với luật quyết định cực tiểu lỗi, trung bình xác suất lỗi là $P(\text{error}) = 1 - \int P(\omega_{\max} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$
3. Chứng minh $P(\text{error}) \leq (c-1)/c$. Đẳng thức xảy ra khi nào?

Giải:

1.

$$c.P(\omega_{\max} \mid \mathbf{x}) \geq \sum_{i=1}^{c} P(\omega_i \mid \mathbf{x}) = 1$$

2.

$$\begin{aligned} & P(\text{error}|\mathbf{x}) = 1 - P(\omega_{\max} \mid \mathbf{x}) \\ & \Rightarrow P(error) = \int P(\text{error}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ & = 1 - \int P(\omega_{\max}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

3.

$$P(error) = 1 - \int P(\omega_{\max}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \leq 1 - \int \frac{p(\mathbf{x})}{c} d\mathbf{x} = 1 - 1/c$$

Đẳng thức xảy ra khi $P(\omega_i \mid \mathbf{x}) = 1/c, \forall i = 1, ..., c$.

Bài 9

Xét bài toán phân c lớp: $\omega_i, i = 1...c$ và c+1 hành động $\alpha_i, i = 1..c+1$ với hành động α_i là phân đối tượng vào lớp $\omega_i, i = 1...c$, hành động α_{c+1} là không phân vào lớp nào hết (rejection). Hàm tiêu tốn được cho bởi

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \quad i, j = 1..c \\ \lambda_r & i = c+1 \\ \lambda_s & otherwise \end{cases}$$

1. Chứng minh rằng cực tiểu rủi ro có thể đạt được với luật sau: nếu $P(\omega_i \mid x) \geq P(\omega_j \mid x), \forall j$ và $P(\omega_i \mid x) \geq 1 - \lambda_r / \lambda_s$ thì quyết định ω_i, ngược lại thì không phân lớp.
2. Chuyện gì xảy ra nếu $\lambda_r > \lambda_s$, nếu $\lambda_r = 0$?

Giải:

1. Dựa vào $R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$, có

Nếu $i = c+1$ thì $R(\alpha_i \mid \mathbf{x}) = R(\alpha_{c+1} \mid \mathbf{x}) = \lambda_r$

Nếu $i < c+1$ thì $R(\alpha_i \mid \mathbf{x}) = \lambda_s \sum_{j \neq i}^{c} P(\omega_j \mid \mathbf{x}) = \lambda_s - \lambda_s P(\omega_i \mid \mathbf{x})$

Theo Luật 2, chọn ω_i với $i < c+1$ nếu

$$\begin{aligned} & R(\alpha_i \mid x) \leq R(\alpha_j \mid x), \forall j \\ \Leftrightarrow & \begin{cases} R(\alpha_i \mid x) \leq R(\alpha_j \mid x), \forall j = 1..c \\ R(\alpha_i \mid x) \leq R(\alpha_{c+1} \mid x) \end{cases} \\ \Leftrightarrow & \begin{cases} \lambda_s - \lambda_s P(\omega_i \mid \mathbf{x}) \leq \lambda_s - \lambda_s P(\omega_j \mid \mathbf{x}), \forall j = 1..c \\ \lambda_s - \lambda_s P(\omega_i \mid \mathbf{x}) \leq \lambda_r \end{cases} \\ \Leftrightarrow & \begin{cases} P(\omega_i \mid \mathbf{x}) \geq P(\omega_j \mid \mathbf{x}), \forall j = 1..c \\ P(\omega_i \mid x) \geq 1 - \lambda_r / \lambda_s \end{cases} \end{aligned}$$

2. Nếu $\lambda_r > \lambda_s$ thì không bao giờ quyết định việc không phân lớp. Nếu $\lambda_r = 0$ thì luôn quyết định việc không phân lớp trừ khi $P(\omega_i \mid x) = 1$.

PHÂN LỚP BẰNG BIỆT HÀM

Bài 10

Người ta xác định rằng cứ 3000 trẻ sơ sinh thì có 1 bị mất thính lực nghiêm trọng. Đây là bệnh nguy hiểm nếu để về sau nên cần phải chẩn đoán kịp thời. Với phương pháp xét nghiệm K nào đó, kết quả là một số thực. Sau khi biến đổi, kết quả xét nghiệm sẽ có phân phối chuẩn (Gaussian) với trung bình (TB) và độ lệch chuẩn (ĐLC) được cho trong bảng

Thính lực	**Xét nghiệm K**	
	TB	**ĐLC**
Bình thường	0	2
Suy giảm	3	2

1. Xây dựng một phân lớp để chẩn đoán (với xác suất lỗi nhỏ nhất) nếu biết rằng trẻ được xét nghiệm mất thính giác.
2. Xét nghiệm cho 1 trẻ được kết quả là 3.5. Trường hợp này cần chẩn đoán thế nào?

Giải:

1. Gọi S là biến ngẫu nhiên thể hiện trạng thái của trẻ với 2 giá trị: 0 nếu thính giác bình thường, 1 nếu thích giác bị tổn thương. Có

$$p_0 = P(S=0) = 1-1/3000,\ p_1 = P(S=1) = 1/3000$$

Gọi X là biến đặc trưng ngẫu nhiên thể hiện kết quả xét nghiệm. Có

$$P(X=x \mid S=0) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_0)^2}{2\sigma^2}\right\}$$

$$P(X=x \mid S=1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}$$

Sử dụng MAP để xây dựng phân lớp. Biệt hàm được xác định là

$$g(x) = \ln P(S=1 \mid X=x) - \ln P(S=0 \mid X=x) = \frac{(x-m)d}{\sigma^2} + \ln \frac{p_1}{p_0}$$

trong đó

$$d = \mu_1 - \mu_0 \qquad m = (\mu_1 + \mu_0)/2$$

Nếu $g(x) > 0$ thì chẩn đoán là trẻ bị mất thính lực.

2. $g(x) \approx -6.5 < 0$.

Bài 11

Tương tự như Bài 10. Nhưng do các phương pháp hiện tại không ổn định khi xét nghiệm ở những ngày đầu mới sinh nên người ta thường sử dụng cùng lúc 2 phương pháp để chẩn đoán. Mỗi xét nghiệm cho kết quả là một số thực. Sau khi biến đổi, kết quả xét nghiệm sẽ có phân phối chuẩn (Gaussian) với trung bình (TB) và độ lệch chuẩn (ĐLC) được cho trong bảng

Thính lực	**Xét nghiệm A**		**Xét nghiệm B**	
	TB	**ĐLC**	**TB**	**ĐLC**
Bình thường	0	1	0	2
Suy giảm	1	1	3	2

Cho biết 2 xét nghiệm từ 2 phương pháp độc lập xác suất.

1. Xây dựng một phân lớp để đoán (với xác suất lỗi nhỏ nhất) nếu biết rằng trẻ được xét nghiệm mất thính giác.
2. Xét nghiệm cho 1 trẻ được kết quả của phương pháp A là 1.5 và phương pháp B là 3.5. Trường hợp này cần chẩn đoán thế nào?

Giải:

1. Gọi S là biến ngẫu nhiên thể hiện trạng thái của trẻ với 2 giá trị: 0 nếu thính giác bình thường, 1 nếu thích giác bị tổn thương. Có

$$p_0 = P(S=0) = 1 - 1/3000,\ p_1 = P(S=1) = 1/3000$$

Gọi $\mathbf{X} = [X_A, X_B]^T$ là vector đặc trưng ngẫu nhiên với X_A, X_B lần lượt là biến ngẫu nhiên thể hiện kết quả xét nghiệm của phương pháp A và B.

X_A, X_B độc lập, do đó

$$P(\mathbf{X} = \mathbf{x} \mid S = 0)$$
$$= \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left\{ -\frac{(x_A - \mu_{0A})^2}{2\sigma_A{}^2} \right\} . \frac{1}{\sigma_B \sqrt{2\pi}} \exp\left\{ -\frac{(x_B - \mu_{0B})^2}{2\sigma_B{}^2} \right\}$$
$$P(\mathbf{X} = \mathbf{x} \mid S = 1)$$
$$= \frac{1}{\sigma_A \sqrt{2\pi}} \exp\left\{ -\frac{(x_A - \mu_{1A})^2}{2\sigma_A{}^2} \right\} . \frac{1}{\sigma_B \sqrt{2\pi}} \exp\left\{ -\frac{(x_B - \mu_{1B})^2}{2\sigma_B{}^2} \right\}$$

Sử dụng MAP để xây dựng phân lớp. Biệt hàm được xác định là

$$g(\mathbf{x}) = \ln P(S = 1 \mid \mathbf{X} = \mathbf{x}) - \ln P(S = 0 \mid \mathbf{X} = \mathbf{x})$$
$$= \frac{(x_A - m_A) d_A}{\sigma_A{}^2} + \frac{(x_B - m_B) d_B}{\sigma_B{}^2} + \ln \frac{p_1}{p_0}$$

trong đó

$$d_A = \mu_{1A} - \mu_{0A} \qquad d_B = \mu_{1B} - \mu_{0B}$$
$$m_A = (\mu_{1A} + \mu_{0A})/2 \qquad m_B = (\mu_{1B} + \mu_{0B})/2$$

Nếu $g(\mathbf{x}) > 0$ thì chẩn đoán là trẻ bị mất thính lực.

2. $g(\mathbf{x}) \approx -5.5 < 0$.

Bài 12

Xét bài toán phân 2 lớp ω_1, ω_2 với đặc trưng x. Cho $p(x \mid \omega_1)$ và $p(x \mid \omega_2)$ như sau

$$p(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad \forall x$$

$$p(x \mid \omega_2) = \begin{cases} \frac{1}{4}, & -2 < x < 2 \\ 0, x \geq 2 \text{ or } x \leq -2 \end{cases}$$

1. Tìm luật phân phối thỏa cực tiểu hóa xác suất lỗi biết rằng $P(\omega_1) = P(\omega_2) = 0.5$.
2. Tìm π_1^* sao cho nếu $P(\omega_1) > \pi_1^*$ thì phân lớp ở câu (1) luôn quyết định ω_1 bất chấp x.
3. Tại sao không có π_2^* sao cho nếu $P(\omega_2) > \pi_2^*$ thì phân lớp ở câu (1) luôn quyết định ω_2?

Giải:

1. Trong trường hợp $-2 < x < 2$, vì $P(\omega_1) = P(\omega_2) = 0.5$ nên xây dựng biệt hàm $g(x)$ sau

$$g(x) = \ln \frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} = \ln \frac{4}{\sqrt{2\pi}} - \frac{x^2}{2}$$

Nếu $g(x) > 0$ thì quyết định ω_1, ngược lại ω_2.

Trong trường hợp $x \geq 2$ hoặc $x \leq -2$ thì luôn quyết định ω_1.

2. Trong trường hợp $x \geq 2$ hoặc $x \leq -2$ thì luôn quyết định ω_1.

Trong trường hợp $-2 < x < 2$, xây dựng biệt hàm $g(x)$ sau

$$g(x) = \ln \frac{p(x \mid \omega_1)P(\omega_1)}{p(x \mid \omega_2)P(\omega_2)} = \ln \frac{4}{\sqrt{2\pi}} - \frac{x^2}{2} + \ln \frac{P(\omega_1)}{1 - P(\omega_1)}$$

$$\geq \ln \frac{4}{\sqrt{2\pi}} - \frac{(\pm 2)^2}{2} + \ln \frac{P(\omega_1)}{1 - P(\omega_1)}$$

Do đó xét tại $x = 2$ và $x = -2$ để tìm π_1^* sao cho

$$\ln \frac{4}{\sqrt{2\pi}} - \frac{(\pm 2)^2}{2} + \ln \frac{\pi_1^*}{1-\pi_1^*} = 0$$

Giải ra được $\pi_1^* \approx 0.8224$.

3. Do trong trường hợp $x \geq 2$ hoặc $x \leq -2$ thì luôn quyết định ω_1.

Bài 13

Xét bài toán phân 2 lớp với xác suất tiền định bằng nhau và các phân bố chuẩn sau

$$p(x \mid \omega_1) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a & c \\ c & b \end{bmatrix}\right), \qquad p(x \mid \omega_2) \sim N\left(\begin{bmatrix} d \\ e \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

trong đó $a.b - c.c = 1$.

1. Tìm biên ra quyết định.
2. Xác định a, b, c, d, e để biên ra quyết định là một đường thẳng.

Giải:

1. Do xác suất tiền định bằng nhau nên xây dựng hàm biệt hàm sau

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$$

Biên ra quết định được cho bởi phương trình

$$g_1(x) = g_2(x)$$
$$\Leftrightarrow (b-1)x_1^2 + (a-1)x_2^2 - 2cx_1x_2 + 2dx_1 + 2ex_2 - d^2 - e^2 = 0$$

2. Để biên ra quyết định là đường thẳng thì a=b=1 và c=0.

Bài 14

Trong một ứng dụng, nguồn tín hiệu có 1 trong 2 trạng thái: S=1 hoặc S=2. Các trạng thái tín hiệu nguồn xuất hiện với cùng xác suất bằng nhau. Người ta có thể quan sát vector đặc trưng $\mathbf{X} = [X_1, X_2]^T$. Tùy vào trạng thái tín hiệu nguồn S, vector đặc trưng có phân phối chuẩn với trung bình và ma trận phương sai như sau

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad C_1 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\mu_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

1. Xây dựng phân lớp tối ưu để đoán trạng thái tín hiệu nguồn với xác suất lỗi nhỏ nhất. Vẽ biên ra quyết định.
2. Chứng minh rằng quyết định tối ưu có thể được cho bằng cách sử dụng biến $Y = a|X_1| + b|X_2|$ và cơ chế lấy ngưỡng đơn giản. Xác định giá trị thích hợp cho a và b.

Giải:

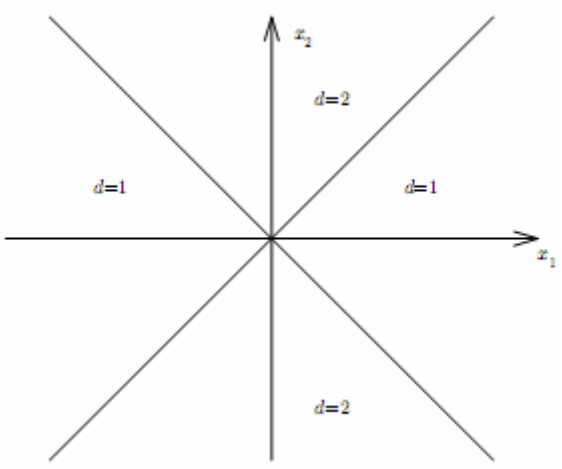
1. Do xác suất tiền định bằng nhau nên ta xây dựng biệt hàm như sau

$$g_i(\mathbf{x}) = \ln P(\mathbf{X} = \mathbf{x} \mid S = i) = -\frac{1}{2}\mathbf{x}^T C_i^{-1}\mathbf{x} - \ln 2\pi\sqrt{\det C_i}, \quad i = 1,2$$

Do chỉ có 2 lớp nên chỉ cần xây dựng biệt hàm sau

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = -x_1^2/6 - x_2^2/2 + x_1^2/2 + x_2^2/6$$
$$\propto x_1^2 - x_2^2$$

Nếu $g(\mathbf{x}) > 0$ thì quyết định trạng thái S=1. Vùng ra quyết định được cho bởi hình sau



2. Do $x_1^2 \geq x_2^2 \Leftrightarrow |x_1| \geq |x_2|$ nên nếu đặt $Y = |X_1| - |X_2|$ và sử dụng ngưỡng là 0 thì sẽ được một phân lớp hoàn toàn tương đương.

Bài 15

Xét bài toán phân 2 lớp với các phân phối chuẩn

$$p(x \mid \omega_1) \sim N(-1,1), \qquad p(x \mid \omega_2) \sim N(4,1)$$

và xác suất tiền định bằng nhau. Hỏi tại sao trung bình xác suất lỗi luôn nhỏ hơn 5%?

Giải:

Do 2 độ lệch chuẩn bằng nhau nên biên ra quyết định là $x = (\mu_1 + \mu_2)/2 = 1.5$. Nhận thấy $|x - \mu_1| = |x - \mu_2| > 2\sigma$, do đó trung bình xác suất lỗi luôn nhỏ hơn 5%.

Chương 2

CỰC ĐẠI LIKELIHOOD VÀ ƯỚC LƯỢNG BAYES

Trong chương trước, ta đã biết cách xây dựng một hệ phân lớp tối ưu nếu được cho trước xác suất tiền định $P(\omega_i)$ và likelihood $p(\mathbf{x}|\omega_i)$. Tuy nhiên, với những ứng dụng nhận dạng thực tế, ta rất khó có được đầy đủ những dữ liệu xác suất của bài toán. Thông thường ta chỉ có kiến thức rất chung chung về vấn đề và một lượng hữu hạn dữ liệu huấn luyện (những dữ liệu này đại diện cho những mẫu mà ta muốn huấn luyện để phân lớp). Bài toán được đặt ra là tìm một phương pháp sử dụng những thông tin này để xây dựng hay huấn luyện một hệ thống phân lớp.

Một phương pháp để giải quyết vấn đề này là sử dụng những mẫu có sẵn để định lượng những giá trị xác suất chưa biết và sử dụng những kết quả này để xây dựng hệ thống phân lớp. Trong những bài toán phân lớp có giám sát (supervised pattern classification problems), việc định lượng xác suất tiền định không gặp khó khăn. Tuy vậy, định lượng likelihood lại là một vấn đề khá nan giải. Số lượng những mẫu có sẵn quá nhỏ, và vấn đề thực sự xuất hiện khi số chiều của vector $\boldsymbol{x}$ quá lớn. Nếu ta biết trước số lượng tham số và bài toán cho ta khả năng lượng hoá những mật độ xác suất có điều kiện thì độ phức tạp của bài toán sẽ được giảm đáng kể. Chẳng hạn như, ta có thể cho rằng $p(\mathbf{x}|\omega_i)$ là hàm mật độ chuẩn với trung bình $\boldsymbol{\mu}_i$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}_i$, mặc dù ta không biết chính xác giá trị của những đại lượng này. Việc giả định này làm đơn giản bài toán từ định lượng một hàm số không biết $p(\boldsymbol{x}|\omega_i)$ thành định lượng hai giá trị $\boldsymbol{\mu}_i$ và $\boldsymbol{\Sigma}_i$.

Bài toán định lượng tham số là một vấn đề kinh điển trong ngành thống kê và có rất hiều hướng tiếp cận. Ta sẽ xem xét hai phương

pháp là: ước lượng cực đại likelihood (maximum likelihood estimation) và ước lượng Bayesian (Bayesian estimation). Mặc dù cách tiếp cận của hai phương pháp này khác nhau, kết quả khi sử dụng chúng lại gần giống nhau. Ước lượng cực đại likelihood xem giá trị tham số là cố định nhưng chưa biết: những giá trị này được ước lượng là tốt nhất khi khả năng có được mẫu quan sát hiện có là lớn nhất. Ngược lại, ước lượng Bayesian xem những tham số như các biến ngẫu nhiên với một phân bố tiền định đã biết. Kết hợp với các mẫu quan sát, ta suy ra được phân bố hậu định và từ đó điều chỉnh lại giá trị thực của những tham số này. Trong ước lượng Bayesian, thêm một mẫu sẽ làm "nhọn" thêm mật độ xác suất hậu định, làm cho đỉnh cực đại của hàm gần với giá trị thực của tham số. Hiện tượng này được gọi là "học Bayesian". Trong cả hai phương pháp, ta đều sẽ sử dụng xác suất hậu định cho luật phân lớp.

Mục đích của chương Sinh viên sau khi học xong chương này cần phải

- Nắm được cơ sở lý thuyết của phương pháp ước lượng cực đại likelihood và ước lượng Bayes
- Từ một bộ dữ liệu cho trước, có thể thực hiện việc ước lượng các tham số thống kê dựa vào 2 phương pháp trên và mô hình cho trước.

I. ƯỚC LƯỢNG CỰC ĐẠI LIKELIHOOD

Phương pháp ước lượng cực đại likelihood có một số ưu điểm nổi trội

- khi số lượng mẫu huấn luyện tăng lên thì nó gần như là sẽ cho kết quả chính xác hơn,
- đơn giản hơn so với các phương pháp khác như ước lượng Bayes.

Các phần tiếp sau sẽ bàn chi tiết về phương pháp này.

I.1 Nguyên lý chung

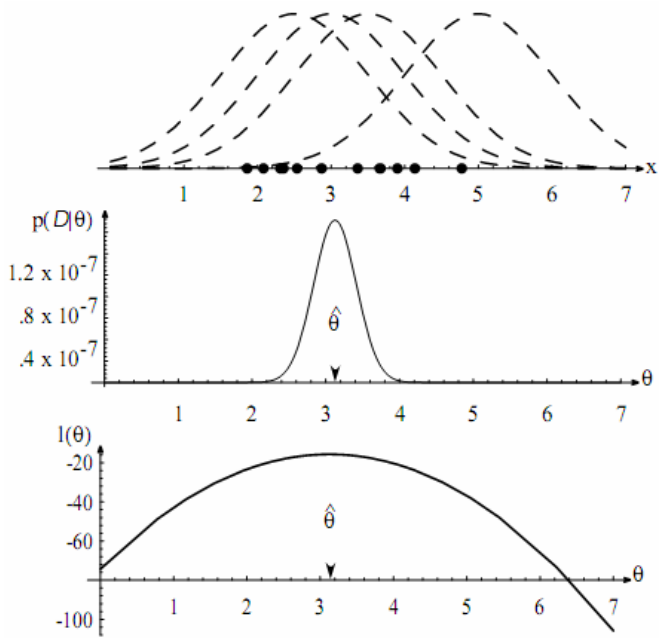
Giả sử đã tách tập mẫu ra thành các lớp, vậy ta sẽ có *c* tập hợp, $D_1,\ldots D_c$, với những mẫu ở lớp D_j đã được chọn một cách độc lập dựa vào luật xác suất $p(\mathbf{x}|\omega_i)$. Ta gọi những mẫu này là độc lập phân bố đồng nhất (i.d.d – Independent Identically Distributed). Giả sử rằng $p(\mathbf{x}|\omega_j)$ có một dạng tham số đã biết và được xác định duy nhất bởi giá trị của vector tham số $\boldsymbol{\theta}_j$. Chẳng hạn, ta có $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \sum_j)$, trong đó $\boldsymbol{\theta}_j$ bao gồm các thành phần của $\boldsymbol{\mu}_j$ và $\sum_j$. Để thể hiện sự phụ thuộc của $p(\mathbf{x}|\omega_j)$ vào $\boldsymbol{\theta}_j$, ta viết $p(\mathbf{x}|\omega_j,\boldsymbol{\theta}_j)$. Bài toán cần giải quyết là sử dụng những thông tin được cung cấp bởi tập huấn luyện để có được một ước lượng tốt cho các vector tham số chưa biết $\boldsymbol{\theta}_1,\ldots, \boldsymbol{\theta}_c$ tương ứng với các phân lớp.

Để đơn giản hoá lời giải cho bài toán, ta giả sử những mẫu trong tập D_i không cho bất kỳ thông tin gì về $\boldsymbol{\theta}_j$ (với $i \neq j$); nghĩa là, ta giả sử những tham số cho những lớp khác nhau là hoàn toàn độc lập. Điều này cho phép ta thao tác trên các lớp một cách độc lập và rời rạc. Với giả thiết này, ta có *c* bài toán rời rạc có dạng sau: Sử dụng tập *D* gồm các quan sát được lấy mẫu một cách độc lập từ hàm mật độ xác suất $p(\mathbf{x}|\boldsymbol{\theta})$ để ước lượng vector tham số chưa biết $\boldsymbol{\theta}$.

Giả sử *D* có *n* mẫu, $\mathbf{x}_1,\ldots, \mathbf{x}_n$, và vì những mẫu này được lấy một cách độc lập, ta có

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) \qquad (2.1)$$

Nhắc lại từ chương trước, ta xem $p(D|\boldsymbol{\theta})$ như một hàm theo $\boldsymbol{\theta}$ và được gọi là likelihood của $\boldsymbol{\theta}$ đối với tập các mẫu tương ứng. Ước lượng cực đại likelihood của $\boldsymbol{\theta}$ là giá trị $\hat{\boldsymbol{\theta}}$ làm cho $p(D|\boldsymbol{\theta})$ đạt giá trị cực đại. Bằng trực giác, ta thấy ước lượng này phù hợp bởi giá trị $\boldsymbol{\theta}$ giúp mô tả chính xác được tập huấn luyện được quan sát (Hình 19)



Hình 19 Hình trên cùng cho thấy tập các điểm 1-chiều có phân bố Gauss, các đường đứt nét là một tập ví dụ của vô hạn các đồ thị mật độ Gauss ứng cử. Hình giữa cho thấy đồ thị biểu diễn likelihood *p(D|θ)* như một hàm của trung bình. Nếu ta có càng nhiều điểm mẫu thì đồ thị càng hẹp. Hình cuối là đồ thị của hàm log của likelihood. $\hat{\boldsymbol{\theta}}$ là giá trị làm cực đại likelihood.

Logarith của likelihood thường được sử dụng trong mục đích phân tích vì nó dễ tính toán hơn so với likelihood. Bởi vì Logarith là hàm đơn điệu tăng, giá trị $\hat{\boldsymbol{\theta}}$ làm cực đại log-likelihood cũng sẽ làm cực đại hàm likelihood. Nếu $p(D|\boldsymbol{\theta})$ là một hàm khả vi của $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ có thể tìm được bằng phương pháp vi phân đại số. Nếu số lượng tham số cần xác định là p, ta sử dụng $\boldsymbol{\theta}$ để ký hiệu cho vector gồm p thành phần $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_p)^T$, và $\nabla_{\boldsymbol{\theta}}$ là phép gradient:

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial}{\partial \theta_p} \end{bmatrix} \tag{2.2}$$

Ta định nghĩa $l(\boldsymbol{\theta})$ là hàm log-likelihood

$$l(\boldsymbol{\theta}) = \ln\ p(D \mid \boldsymbol{\theta}) \tag{2.3}$$

Lời giải cho bài toán sẽ là giá trị của $\boldsymbol{\theta}$ sao cho hàm log-likelihood đạt cực đại

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \tag{2.4}$$

Từ đẳng thức (2.1), ta có

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\ p(\mathbf{x}_k \mid \boldsymbol{\theta}) \tag{2.5}$$

và

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln\ p(\mathbf{x}_k \mid \boldsymbol{\theta}) \tag{2.6}$$

Vì vậy, tập hợp những phương trình cần thiết để ước lượng cực đại likelihood cho $\boldsymbol{\theta}$ có thể biểu diễn bằng tập p phương trình

$$\nabla_{\boldsymbol{\theta}} l = 0 \tag{2.7}$$

Lời giải $\hat{\boldsymbol{\theta}}$ cho phương trình (2.7) có thể cho ra cực đại toàn cục, cực đại hay cực tiểu địa phương, hay là một điểm uốn của $l(\boldsymbol{\theta})$. Vì vậy cần phải kiểm tra để tìm ra cực đại toàn cục thực sự. Ngoài ra cũng cần phải kiểm tra cực đại có xuất hiện tại hai điểm biên của vùng tham số.

Một điều ta cần lưu ý là: $\hat{\boldsymbol{\theta}}$ chỉ là một sự ước lượng – ta chỉ có một giới hạn của một tập lớn vô hạn số các điển huấn luyện và ta kỳ vọng sự ước lượng này sẽ bằng với giá trị thực của hàm được sinh ra.

I.2 Trường hợp Gauss: không biết $\boldsymbol{\mu}$

Để thấy kết quả của cực đại likelihood áp dụng cho một trường hợp cụ thể, ta giả sử các mẫu được lấy từ một phân bố chuẩn đã biết với trung bình $\boldsymbol{\mu}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}$. Để cho đơn giản, trước tiên, ta xem xét trường hợp khi giá trị trung bình $\boldsymbol{\mu}$ chưa biết. Xem xét tại một mẫu $\mathbf{x}_k$ bất kỳ ta nhận thấy

$$\ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \qquad (2.8)$$

và

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) = \Sigma^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \qquad (2.9)$$

Đồng nhất $\boldsymbol{\theta}$ với $\boldsymbol{\mu}$, từ (2.9), ta thấy ước lượng cực đại likelihood cho $\boldsymbol{\mu}$ phải thoả

$$\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0 \qquad (2.10)$$

Nhân với Σ và rút gọn, ta được

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \qquad (2.11)$$

Kết quả này rất hợp lý. Nó có ý nghĩa là ước lượng cực đại likelihood của một tập chưa biết chỉ là giá trị trung bình số học của tập các mẫu huấn luyện – trung bình mẫu – đôi khi được ký hiệu là $\hat{\boldsymbol{\mu}}_n$ để làm rõ sự phụ thuộc vào số lượng mẫu. Về mặt hình học, nếu ta xem *n* mẫu như là một "đám mây" các điểm thì trung bình mẫu là tâm của "đám mây".

I.3 Trường hợp Gauss: Không biết μ và ∑

Trong các bài toán tổng quát (và là những trường hợp thông thường), cả $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$ đều chưa được biết. Như vậy, hai tham số này tạo nên những thành phần của vector tham số $\boldsymbol{\theta}$. Trước tiên ta xem xét trường hợp đơn biến với $\theta_1 = \mu$ và $\theta_2 = \sigma^2$. Log-likelihood của một mẫu là

$$\ln p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2 \tag{2.12}$$

và đạo hàm của nó là

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{2.13}$$

Áp dụng (2.7) cho log-likelihood sẽ được điều kiện sau

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \tag{2.14}$$

và

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \tag{2.15}$$

Với $\hat{\theta}_1$ và $\hat{\theta}_2$ là hai ước lượng cực đại likelihood cho θ_1 và θ_2 tương ứng. Bằng cách thay $\hat{\mu} = \hat{\theta}_1$, $\hat{\sigma}^2 = \hat{\theta}_2$ và rút gọn, ta được ước lượng cực đại likelihood cho μ và σ^2

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{2.16}$$

và

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{2.17}$$

Với trường hợp đa biến thì hoàn toàn tương tự, ta cũng có thể dự đoán kết quả của ước lượng likelihood cực đại cho $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$ được cho bởi các công thức sau

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \tag{2.18}$$

và

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T \tag{2.19}$$

Vì vậy, lần này ta cũng tìm được ước lượng cực đại likelihood cho trung bình là trung bình mẫu. Ước lượng cực đại likelihood cho ma trận hiệp phương sai là trung bình đại số của n ma trận $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$. Vì ma trận hiệp phương sai là kỳ vọng của $(\mathbf{x} - \hat{\boldsymbol{\mu}})(\mathbf{x} - \hat{\boldsymbol{\mu}})^T$ nên đây cũng là một kết quả rất thoả đáng.

I.4 Độ lệch

Ước lượng cực đại likelihood cho biến σ^2 bị lệch *(biased)*, có nghĩa là giá trị kỳ vọng trên toàn bộ tập dữ liệu kích thước n mẫu không bằng với giá trị thực của phương sai:

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{2.20}$$

Ước lượng cực đại likelihood cho một ma trận hiệp phương sai, tương tự, cũng bị lệch.

Ước lượng không lệch *(unbiased)* cơ bản của σ^2 và Σ được cho bởi

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=\sigma^2 \tag{2.21}$$

$$\mathbf{C}=\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^T \tag{2.22}$$

trong đó **C** được gọi là ma trận hiệp phương sai mẫu. Nếu một ước lượng là không lệch cho tất cả các phân bố, chẳng hạn như ở (2.21), sẽ được gọi là không lệch tuyệt đối *(absolutely unbiased)*. Nếu một ước lượng có xu hướng không lệch khi số lượng mẫu càng lớn thì ta gói nó là tiệm cận không lệch *(asymptotically unbiased)*. Trong nhiều bài toán về nhận dạng mẫu với tập dữ liệu huấn luyện lớn, ước lượng tiệm cận không lệch có thể chấp nhận được.

Rõ ràng là $\hat{\boldsymbol{\Sigma}}=\left[(n-1)/n\right]\mathbf{C}$ tiệm cận không lệch – hai ước lượng này sẽ đồng nhất khi $n\to\infty$. Mặc dù vậy, sự tồn tại của hai ước lượng ma trận hiệp phương sai tương tự nhau nhưng tuyệt đối khác nhau cũng có thể gây mâu thuẫn và cũng tự nhiên khi câu hỏi đặt ra là kết quả nào "đúng". Đương nhiên, với *n > 1*, câu trả lời là những ước lượng đó không đúng cũng không sai, chúng chỉ "khác nhau". Sự tồn tại cùng lúc hai ước lượng cho thấy rằng không có một ước lượng đơn lẻ nào có thể có hết mọi tính chất mà ta kỳ vọng. Với mục đích xây dựng hệ phân lớp tốt nhất, những tính chất được kỳ vọng nhất khá phức tạp. Mặc dù việc xây dựng những hệ phân lớp bằng cách thay thể ước lượng cực đại likelihood cho những tham số chưa biết là rất hợp lý và đúng đắn, nhưng ta vẫn luôn thắc mắc liệu rằng những ước lượng khác có thể dẫn cho ra kết quả tốt hơn hay không. Phần tiếp theo trình bày câu hỏi này dưới góc nhìn của Bayes.

Nếu ta có một mô hình đáng tin cây cho những phân phối và chúng phụ thuộc vào tham số $\boldsymbol{\theta}$, mô hình phân lớp cực đại likelihood rõ ràng sẽ cho ra kết quả tốt nhất. Nhưng nếu như mô hình của ta là sai - liệu ta có được một mô hình phân lớp tốt nhất với tập những mô hình đã giả định sẵn? Chẳng hạn như, sẽ ra sao nếu ta giả định rằng một phân bố theo $N(\mu, 1)$ nhưng thực tế nó lại theo $N(\mu, 10)$? Như vậy thì giá trị ta tìm được cho $\theta = \mu$ bằng cực đại likelihooa có cho ra kết quả tốt

nhất với dạng thức từ $N(\mu, 1)$? Rất tiếc câu trả lời là "không" khi mà sai số của mô hình lớn đáng chú ý. Từ đây, ta nhận thức được rằng ta cần có những thông tin đáng tin cậy về mô hình – nếu mô hình giả định là sai thì ta không thể đảm báo kết quả từ mô hình đó là tốt nhất, ngay cả trong tập mô hình cho trước.

II. ƯỚC LƯỢNG BAYES

Trong phần này ta sẽ xem xét ướng lượng Bayes hay còn được gọi là phương pháp học Bayes trong bài toán phân loại. Mặc dù lời giải có được từ phương pháp này khá giống với lời giải từ phương pháp cực đại likelihood, chúng vẫn khác nhau về mặt khái niệm: với phương pháp likelihood, trong khi ta giả thiết vector tham số $\boldsymbol{\theta}$ cần tìm là cố định thì trong phương pháp học Bayes, ta xem $\boldsymbol{\theta}$ như một biến ngẫu nhiên và dữ liệu huấn luyện giúp ta biến đổi phân bố tiền định trên biến này thành mật độ phân bố hậu định.

II.1 Hàm mật độ khi biết phân lớp (class-conditional density)

Việc tính xác suất hậu định $P(\omega_i|\mathbf{x})$ là phần lõi của phương pháp phân loại Bayesian. Công thức Bayes cho phép ta tính toán những xác xuất này từ những xác suất tiền định $P(\omega_i)$ và mật độ khi biết phân lớp $p(\mathbf{x}|\omega_i)$, nhưng làm sao để tính khi mà ngay những lượng này cũng chưa biết? Câu trả lời chung: cách tốt nhất ta có thể tính $P(\omega_i|\mathbf{x})$ là sử dụng những thông tin sẵn có. Một phần của thông tin này có thể là những kiến thức đã được biết trước, như thông tin về loại hàm của những mật độ chưa biết và khoảng giá trị của những tham số chưa biết. Một phần của thông tin này có thể nằm trong tập các mẫu huấn luyện. Nếu ta đặt D là tập các mẫu, ta nhấn mạnh đến vai trò của các mẫu bằng cách xác định mục đích là tính các xác suất hậu định $P(\omega_i|\mathbf{x},D)$. Từ những xác suất này, ta có thể thực hiện phân loại Bayes.

Cho biết mẫu D, công thức Bayes trở thành

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D)P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D)P(\omega_j \mid D)} \qquad (2.23)$$

Công thức trên gợi ý rằng: ta có thể sử dụng những thông tin được cung cấp bởi các mẫu huấn luyện để xác định mật độ khi biết phân lớp và xác suất tiền định.

Mặc dù ta có thể duy trì tính tổng quát của vấn đề, ta nên xem xét việc xác suất tiền định được biết hoặc có thể dễ dàng tính toán được; sau đó, ta sẽ thay thế $P(\omega_i) = P(\omega_i|D)$. Hơn nữa, với những trường hợp học có giám sát, ta có thể tách riêng những mẫu huấn luyện thành những lớp vào c tập con $D_1,\ldots, D_c$ với những mẫu trong D_i thuộc về lớp ω_i. Như đã đề cập khi nói về phương pháp cực đại likelihood, trong đại đa số các trường hợp (và trong tất cả những trường hợp mà ta sẽ xem xét), những mẫu trong D_i sẽ không ảnh hưởng lên $p(\mathbf{x}|\omega_j,D)$ (với $i \neq j$). Điều này sẽ dẫn tới được hai hệ quả đơn giản sau đây. Trước hết, nó cho phép ta làm việc với từng lớp riêng biệt, chỉ sử dụng những mẫu ở D_i để tính $p(\mathbf{x}|\omega_i,D)$. Kết hợp với giả thiết về xác suất tiền định được biết, công thức (2.23) sẽ được viết thành

$$P(\omega_i \mid \mathbf{x}, D) = \frac{p(\mathbf{x} \mid \omega_i, D_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, D_j)P(\omega_j)} \qquad (2.24)$$

Thứ hai, vì mỗi lớp được tính toán riêng biệt và độc lập, ta có thể miễn sự phân biệt giữa các lớp và đơn giản hoá những ký hiệu. Như vậy, ta sẽ có c bài toán riêng biệt có dạng sau: sử dụng một tập D của những mẫu được lấy một cách độc lập, dựa vào xác suất cố định nhưng chưa biết $p(\mathbf{x})$ để tính $p(\mathbf{x}|D)$. Đây là vấn đề cốt lõi của bài toán học Bayes.

II.2 Phân bố xác suất của tham số

Mặc dù mật độ xác suất $p(\mathbf{x})$ là không biết, ta giả sử rằng nó có một dạng thức cho trước. Chỉ duy nhất một thứ chưa biết đó là giá trị của vector tham số $\boldsymbol{\theta}$. Ta nhấn mạnh về việc $p(\mathbf{x})$ không biết nhưng lại có một dạng thức đã biết bằng cách đặt hàm $p(\mathbf{x}|\boldsymbol{\theta})$ và hàm này hoàn toàn không biết. Bất kỳ thông tin gì ta có thể biết trước về $\boldsymbol{\theta}$ để quan sát mẫu sẽ xem như được chứa trong mật độ tiền định $p(\boldsymbol{\theta})$. Kết hợp với

các mẫu, ta sẽ có mật độ hậu định $p(\boldsymbol{\theta}|D)$ với hy vọng là đỉnh của nó sẽ ở lân cận giá trị thực của $\boldsymbol{\theta}$.

Lưu ý rằng ta đã chuyển bài toán học có giám sát thành bài toán ước lượng mật độ trong học không giám sát. Cho đến đây, mục đích cơ bản của ta là tính $p(\mathbf{x}|D)$ càng chính xác càng tốt để chúng ta có thể có được $p(\mathbf{x})$. Ta thực hiện việc này bằng cách lấy tích phân của $p(\mathbf{x},\boldsymbol{\theta}|D)$ trên toàn bộ không gian tham số $\boldsymbol{\theta}$

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D) d\boldsymbol{\theta} \qquad (2.25)$$

Bây giờ, vì $p(\mathbf{x},\boldsymbol{\theta}|D) = p(\mathbf{x}|\boldsymbol{\theta},D)p(\boldsymbol{\theta}|D)$ và vì sự các mẫu trong D được lấy một cách độc lập nên $p(\mathbf{x}|\boldsymbol{\theta},D) = p(\mathbf{x}|\boldsymbol{\theta})$ (có nghĩa là sự phân bố $\mathbf{x}$ được biết khi biết giá trị của vector tham số). Và vì vậy, đẳng thức (2.25) được viết thành

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta} \qquad (2.26)$$

Nếu $p(\boldsymbol{\theta}|D)$ có đỉnh rất nhọn tại một giá trị $\hat{\boldsymbol{\theta}}$ nào đó, ta được $p(\mathbf{x} \mid D) \approx p(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$, có nghĩa là kết quả có thể có được bằng cách thay thế giá trị $\hat{\boldsymbol{\theta}}$ vào vector tham số. Kết quả này dựa trên giả định rằng $p(\mathbf{x}|\boldsymbol{\theta})$ trơn, và phần đuôi của tích phân không quan trọng. Những điều kiện này khá phổ biết nhưng không phải là bất biến. Nhìn chung, nếu ta càng ít chắc chắn về giá trị chính xác của $\boldsymbol{\theta}$, đẳng thức này sẽ dẫn tới giá trị trung bình $p(\mathbf{x}|\boldsymbol{\theta})$ trên toàn bộ giá trị có thể có của $\boldsymbol{\theta}$. Như vậy, khi mật độ chưa được xác định có một dạng tham số đã biết, các mẫu sẽ ảnh hưởng lên $p(\mathbf{x}|\mathbf{D})$ thông qua xác suất hậu định $p(\boldsymbol{\theta}|D)$. Trong thực tế tích phân ở (2.26) sẽ được tính toán bằng số học (chẳng hạn như bằng phương pháp Monte-Carlo).

III. ƯỚC LƯỢNG THAM SỐ BAYAES TRONG TRƯỜNG HỢP GAUSS

Trong phần này, ta sử dụng phương pháp ước lượng Bayes để tính toán mật độ hậu định $p(\boldsymbol{\theta}|D)$ và mật độ xác suất $p(\mathbf{x}|D)$ cho trường hợp $p(\mathbf{x}\,|\,\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

III.1 Trường hợp đơn biến: p(μ|D)

Xem xét trường hợp $\boldsymbol{\mu}$ là tham số chưa biết. Để cho đơn giản, trước tiên, ta xem nó như trường hợp đơn biến, có nghĩa là

$$p(x\,|\,\mu) \sim N(\mu, \sigma^2) \tag{2.27}$$

trong đó chỉ có trung bình μ là chưa biết. Giả sử rằng mọi điều ta có thể biết được về μ đều có thể biểu diễn dưới dạng một mật độ tiền định $p(\mu)$. Sau đó, ta tiến đến một giả thiết khác

$$p(\mu) \sim N(\mu_0, \sigma_0^2) \tag{2.28}$$

với μ_0 và σ_0^2 đều đã biết. Nói nôm na là μ_0 thể hiện giá trị tiên đoán tốt nhất của μ và σ_0^2 là độ đo sự không chắc chắn về điều tiên đoán này. Giả thiết về tiền phân bố cho μ là chuẩn sẽ giúp đơn giản hoá những vấn đề toán học về sau.

Chọn mật độ tiền định cho μ sẽ giúp ta có cái nhìn về bài toán như sau. Giả sử một giá trị được lấy ra cho μ trong miền giá trị được xem xét theo xác suất $p(\mu)$. Khi giá trị này được rút ra, nó trở thành giá trị thực của μ và xác định mật độ của x. Giả sử n mẫu $x_1,\ldots, x_n$ được lấy độc lập. Đặt $D = \{x_1, \ldots, x_n\}$, sử dụng công thức Bayes để có

$$p(\mu\,|\,D) = \frac{p(D\,|\,\mu)\,p(\mu)}{\int p(D\,|\,\mu)\,p(\mu)\,d\mu} = \alpha \prod_{k=1}^{n} p(x_k\,|\,\mu)\,p(\mu) \tag{2.29}$$

trong đó α là tác nhân chuẩn hóa phụ thuộc vào D nhưng lại độc lập với μ. Công thức này cho ta thấy những mẫu huấn luyện ảnh hưởng như thế nào đến giá trị thực của μ; nó tạo ra mối quan hệ từ mật độ

tiền định $p(\mu)$ đến mật độ hậu định $p(\mu|D)$. Vì $p(x \mid \mu) \sim N(\mu, \sigma^2)$ và $p(\mu) \sim N(\mu_0, \sigma_0^2)$, ta có

$$p(\mu|D) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu - x_k}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

(2.30)

Trong đó những thừa số không phụ thuộc vào μ đã được gom lại thành hằng số α, α' và α''. Nhận thấy $p(\mu|D)$ là một hàm mũ của một hàm bậc 2 theo μ, như vậy đây là một phân bố chuẩn. Bởi vì điều này đúng cho bất kỳ độ lớn nào của tập huấn luyện, $p(\mu|D)$ vẫn luôn là phân bố chuẩn khi số lượng mẫu tăng và $p(\mu|D)$ được gọi là mật độ tái tạo *(reproducing density)* và $p(\mu)$ được gọi là liên hợp tiền định *(conjugate prior)*. Nếu viết $p(\mu \mid D) \sim N(\mu_n, \sigma_n^2)$, μ_n và σ_n^2 có thể tìm được bằng cách tính hệ số ở (2.30) với dạng biểu diễn Gaussian

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] \tag{2.31}$$

Xác định hệ số bằng cách này cho ta

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \tag{2.32}$$

và

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma_2}\overline{x}_n + \frac{\mu_0}{\sigma_0^2} \tag{2.33}$$

trong đó $\overline{x}_n$ là trung bình mẫu

$$\overline{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{2.34}$$

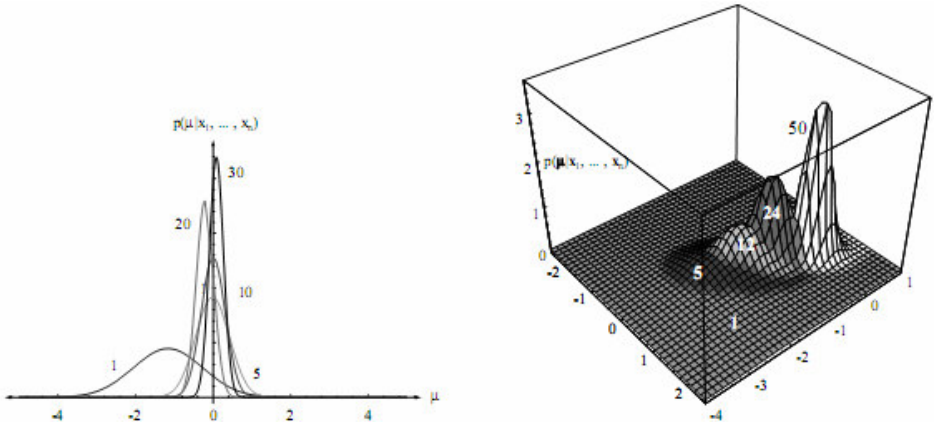
Giải riêng cho μ_n và σ_n^2 ta được

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\overline{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \tag{2.35}$$

và

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{2.36}$$

Những đẳng thức này cho thấy những thông tin được biết trước kết hợp với những thông tin thực nghiệm trong các mẫu như thế nào để có được mật độ hậu định $p(\mu|D)$. Nói nôm na là μ_n biểu diễn diễn sự tiên đoán tốt nhất của ta về μ sau khi quan sát n mẫu, và σ_n^2 đo mức độ không chính xác của việc tiên đoán này. Vì σ_n^2 giảm ngặt với n – tiến đến σ^2/n khi n tiến đến vô cực – mỗi một khi quan sát thêm một mẫu thì mức độ không chắc chắn về kết quả thực của μ sẽ giảm. Khi n tăng, $p(\mu|D)$ sẽ có đỉnh nhọn hơn, tiệm cận tới hàm Dirac delta khi n tiến đến vô cực. Điều này được gọi chung là học Bayes (Hình 20).



Hình 20 Học Bayes của trung bình của phân phối chuẩn trong 1 và 2 chiều. Mỗi ước lượng phân phối hậu định được đánh dấu bằng số lượng mẫu được sử dụng.

Nhìn chung, μ_n là kết hợp tuyến tính của $\bar{x}_n$ và μ_0 với các hệ số không âm có tổng bằng 1. Như vậy μ_n luôn có giá trị nằm giữa $\bar{x}_n$ và μ_0. Nếu $\sigma \neq 0$, μ_n tiến đến trung bình mẫu khi n tiến đến vô cực. Nếu $\sigma_0 = 0$, lúc đó sự chắc chắn về $\mu = \mu_0$ mạnh đến nỗi với mọi tập mẫu quan sát đều không thể thay đổi quan điểm của ta được. Ở thái cực khác, nếu $\sigma_0 >> \sigma$, ta khá không chắc về tiên đoán $\mu_n = \bar{x}_n$, và do đó chỉ có thể sử dụng tập mẫu để ước lượng μ. Nói chung, sự cân bằng tương đối giữa những điều biết trước và dữ liệu thử nghiệm được quyết định bởi tỉ lệ của σ^2 trên σ_0^2, được gọi là tỉ lệ võ đoán *(dogmatism)*. Nếu tỉ lệ này hữu hạn thì sau một lượng đủ các mẫu được lấy, giá trị chính xác giả định cho μ_0 và σ_0^2 sẽ không quan trọng và μ_n sẽ hội tụ về trung bình mẫu.

III.2 Trường hợp đơn biến: p(x|D)

Sau khi đã tìm được mật độ hậu định cho trung bình, $p(\mu|D)$, ta chỉ còn phải tìm mật độ khi biết phân lớp $p(x|D)$. Từ (2.26), (2.27) và (2.31) ta có

$$p(x|D) = \int p(x|\mu)\,p(\mu|D)\,d\mu$$
$$= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu$$
$$= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n)$$

(2.37)

Trong đó

$$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu\,.$$

Như vậy, *p(x|D)* có phân phối chuẩn với trung bình μ_n và phương sai $\sigma^2+\sigma_n^2$

$$p(x|D) \sim N(\mu_n, \sigma^2+\sigma_n^2) \qquad (2.38)$$

Nói cách khác, để có được *p(x|D)*, có dạng thức được biết là $p(x|\mu) \sim N(\mu,\sigma^2)$, ta chỉ việc thay μ bằng μ_n và σ^2 bằng $\sigma^2+\sigma_n^2$. Kết quả cuối cùng: *p(x|D)* là mật độ mong muốn *p(x|ω_i,D_j)* và cùng với mật độ *P(ω_j)* sẽ cho ta những thông tin xác suất cần thiết để xây dựng hệ phân lớp. Điều này trái ngược với phương pháp cực đại likelihood chỉ tạo những điểm ước lượng cho $\hat{\mu}$ và $\hat{\sigma}^2$, thay vì ước lượng phân bố cho *p(x|D)*.

III.3 Trường hợp đa biến

Phương pháp giải quyết trường hợp đa biến trong đó chỉ có $\boldsymbol{\mu}$ chưa biết là một trường hợp tổng quát của đơn biến. Vì vậy, ở đây, ta sẽ chỉ mô tả những điểm khác. Như trước đó đã giả sử rằng

$$p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \qquad p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \qquad (2.39)$$

Với Σ, Σ_0, và μ_0 đã biết trước. Sau khi quan sát tập D của n mẫu độc lập $\mathbf{x}_1, \ldots, \mathbf{x}_n$, sử dụng công thức Bayes để được

$$p(\boldsymbol{\mu}|D) = \alpha \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\mu}) p(\boldsymbol{\mu})$$
$$= \alpha' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^T \left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \right)\boldsymbol{\mu} - 2\boldsymbol{\mu}\left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^{n} \mathbf{x}_k + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \right) \right) \right]$$

(2.40)

có dạng

$$p(\boldsymbol{\mu} \mid D) = \alpha'' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu} - \boldsymbol{\mu}_n \right)^T \boldsymbol{\Sigma}_n^{-1} \left(\boldsymbol{\mu} - \boldsymbol{\mu}_n \right) \right] \qquad (2.41)$$

Như vậy, $p(\boldsymbol{\mu} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, và, một lần nữa, ta lại có mật độ tái tạo. Đồng nhất các hệ số, ta được

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \qquad (2.42)$$

và

$$\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \qquad (2.43)$$

Trong đó $\hat{\boldsymbol{\mu}}_n$ là trung bình mẫu

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \qquad (2.44)$$

Lời giải cho những phương trình này cho $\boldsymbol{\mu}$ và Σ_n được đơn giản hoá bằng ma trận

$$\left(\mathbf{A}^{-1} + \mathbf{B}^{-1} \right)^{-1} = \mathbf{A}\left(\mathbf{A} + \mathbf{B} \right)^{-1} = \mathbf{B}\left(\mathbf{A} + \mathbf{B} \right)^{-1} \mathbf{A} \qquad (2.45)$$

điều này có hiệu lực cho tất cả những cặp ma trận **A** và **B** *nonsingular*, kích thước $d \times d$. Sau một vài tính toán, ta có được kết quả cuối cùng

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \hat{\boldsymbol{\mu}}_n + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0 \qquad (2.46)$$

và

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma} \qquad (2.47)$$

Chứng minh $p(\mathbf{x} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$ có thể có được bằng việc tính toán tích phân

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid D) d\boldsymbol{\mu} \qquad (2.48)$$

Nhưng, kết quả này có thể có được bằng cách đơn giản hơn: xem **x** như là một tổng của biến ngẫu nhiên độc lập: gồm một vector ngẫu nhiên $\boldsymbol{\mu}$ với $p(\mathbf{x} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ và một vector ngẫu nhiên $\boldsymbol{\gamma}$ với $p(\boldsymbol{\gamma}) \sim N(0, \Sigma)$. Bởi vì tổng của hai vector ngẫu nhiên độc lập có phân bố chuẩn sẽ là một vector ngẫu nhiên có phân bố chuẩn, với trung bình là tổng của hai trung bình và ma trận hiệp phương sai là tổng của hai ma trận hiệp phương sai. Ta có

$$p(\mathbf{x} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n) \qquad (2.49)$$

và việc tổng quát hoá hoàn tất.

IV. ƯỚC LƯỢNG THAM SỐ BAYES: NGUYÊN LÝ TỔNG QUÁT

Ta đã xem xét phương pháp Bayes để có được mật độ $p(\mathbf{x}|D)$ trong trường hợp đặc biệt – trường hợp phân phối chuẩn đa biến. Phương pháp này có thể được tổng quát hoá để áp dụng cho mọi trường hợp mà mật độ không biết có thể tham số hoá. Những giả định cơ bản được tóm tắt như sau:

- Dạng của mật độ $p(\mathbf{x}|\boldsymbol{\theta})$ được xem là đã biết, nhưng giá trị của tham số vector $\boldsymbol{\theta}$ thì không biết chính xác.

- Những thông tin ban đầu về $\boldsymbol{\theta}$ được cho là có chứa trong một mật độ tiền định $p(\boldsymbol{\theta})$ đã biết.
- Những thông tin còn lại về $\boldsymbol{\theta}$ có chứa trong một tập hợp D gồm n mẫu $\mathbf{x}_1,\ldots,\mathbf{x}_n$ được lấy độc lập dựa vào mật độ xác suất $p(\mathbf{x})$ không biết.

Bài toán cơ bản là tìm mật độ hậu định $p(\boldsymbol{\theta}|D)$, từ đó sử dụng đẳng thức (2.26) để tính $p(\mathbf{x}|D)$

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta} \tag{2.50}$$

Bằng công thức Bayes, ta có

$$p(\boldsymbol{\theta} \mid D) = \frac{p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{2.51}$$

và với giả thiết độc lập

$$p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}) \tag{2.52}$$

Điều này thiết lập lời giải chi tiết bài toán, và đẳng thức (2.51), (2.52) mô tả mối quan hệ của nó với phương pháp cực đại likelihood. Giả sử $p(D|\boldsymbol{\theta})$ đạt được đỉnh nhọn tại $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Nếu mật độ tiền định $p(\boldsymbol{\theta})$ khác 0 và không thay đổi nhiều xung quanh miền lân cận, thì $p(\boldsymbol{\theta}|D)$ cũng có đỉnh tại vị trí đó. Như vậy, đẳng thức (2.26) cho ta thấy $p(\mathbf{x}|D)$ sẽ xấp xỉ $p(\mathbf{x}|\hat{\boldsymbol{\theta}})$, kết quả có thể có được bằng ước lượng cực đại likelihood. Nếu đỉnh của $p(D|\boldsymbol{\theta})$ rất nhọn, ảnh hưởng của những thông tin có được lên độ không chắc chắn về giá trị của $\boldsymbol{\theta}$ có thể được bỏ qua. Trong trường hợp này và ngay cả trong trường hợp tổng quát, phương pháp Bayes cho ta biết cách sử dụng những thông tin có sẵn để tính toán mật độ $p(\mathbf{x}|D)$.

Đến đây, một số câu hỏi được đặt ra. Thứ nhất liên quan đến độ phức tạp tính toán. Cái còn lại là sự hội tụ của $p(\mathbf{x}|D)$ về $p(\mathbf{x})$. Ta sẽ xem xét

về sự hội tụ một cách ngắn gọn và sau đó sẽ quay lại vấn đề độ phực tạp tính toán.

Gọi $D^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ là tập các mẫu có trong một lớp. Từ (2.52), nếu $n > 1$

$$p(D^n \mid \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(D^{n-1} \mid \boldsymbol{\theta}) \tag{2.53}$$

Thay thế công thức này vào (2.51) và sử dụng công thức Bayes, ta thấy mật độ hậu định thoả mãn quan hệ hồi quy

$$p(\boldsymbol{\theta} \mid D^n) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D^{n-1}) d\boldsymbol{\theta}} \tag{2.54}$$

Ta có $p(\boldsymbol{\theta}|D^0) = p(\boldsymbol{\theta})$, sử dụng lặp lại phương trình này cho ta những mật độ tiếp theo $p(\boldsymbol{\theta})$, $p(\boldsymbol{\theta}|\mathbf{x}_1)$, $p(\boldsymbol{\theta}|\mathbf{x}_1,\mathbf{x}_2)$,… (Cũng khá rõ ràng từ đẳng thức (2.54), ta thấy $p(\boldsymbol{\theta}|D^n)$ chỉ phụ thuộc vào những điểm trong D^n, chứ không phải quá trình mà nó được chọn.) Điều này được gọi là phương pháp ước lượng hồi quy Bayes. Đây cũng là ví dụ đầu tiên về phương pháp học trực tuyến *(on-line)* khi mà việc học được diễn ra cùng lúc với việc thu thập dữ liệu. Kết quả của mật độ sẽ hội tụ về hàm Dirac delta có tâm tại giá trị tham số thực (Ví dụ 1).

Về nguyên tắc, đẳng thức (2.54) yêu cầu ta phải lưu lại tất cả những điểm huấn luyện trong D^{n-1} để tính $p(\boldsymbol{\theta}|D^n)$, nhưng với một số phân bố ta chỉ cần quan tâm đến một ít các tham số liên quan đến $p(\boldsymbol{\theta}|D^{n-1})$ chứa những thông tin cần thiết.

Ví dụ 1:

Giả sử tất cả các mẫu có cùng 1 dạng thức phân bố

$$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & otherwise \end{cases},$$

Nhưng ban đầu ta chỉ biết tham số bị chặn. Cụ thể, ta giả sử $0 < \theta \le 10$. Sử dụng phương pháp hồi quy Bayes để ước lượng θ và mật độ xác suất từ tập dữ liệu $D = \{4, 7, 2, 8\}$, đã được chọn ngẫu nhiên từ phân bố nền. Trước khi có những dữ liệu mới, ta có $p(\theta|D^0)$

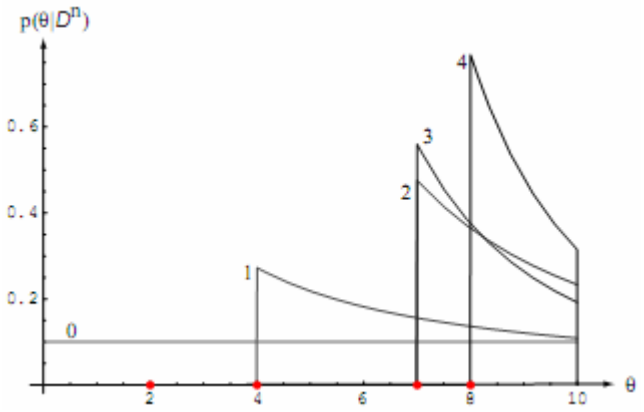
$= p(\theta) = U(0, 10)$. Khi có dữ liệu đầu tiên $x_1 = 4$, sử dụng (2.54) để có ước lượng tốt hơn:

$$p(\theta \mid D^1) \propto p(x \mid \theta) p(\theta \mid D^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & otherwise \end{cases}$$

ở đây ta bỏ qua tham số chuẩn hóa. Khi có $x_2 = 7$, ta được:

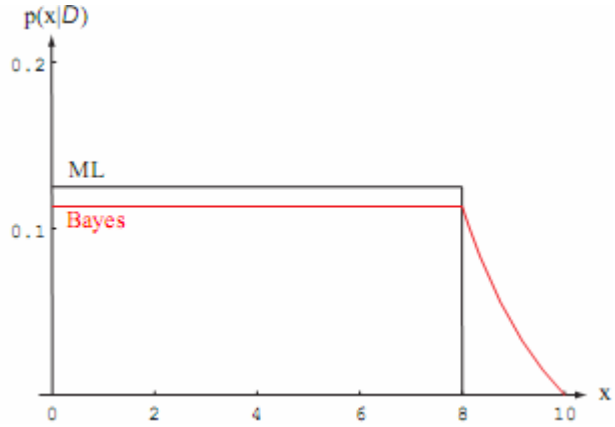
$$p(\theta \mid D^2) \propto p(x \mid \theta) p(\theta \mid D^1) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & otherwise \end{cases}$$

Tương tự cho những điểm mẫu còn lại. Cũng nên hiểu rõ rằng vì mỗi bước thành công sẽ thêm một thừa số $1/\theta$ vào $p(x|\theta)$, và phân bố khác 0 chỉ đúng với những giá trị x lớn hơn điểm mẫu lớn nhất trong tập D, dạng chung của lời giải sẽ là $p(\theta \mid D^n) \propto 1/\theta^n$ cho $\max_x[D^n] \le \theta \le 10$, như thể hiện trong Hình 21. Áp dụng phương pháp cực đại likelihood cho tập D ta sẽ được $\hat{\theta} = 8$, và điều này ngụ ý một phân phối đồng nhất *(uniform)* $p(x|D) \sim U(0,8)$.



Hình 21 Mật độ hậu định $p(\theta \mid D^n)$ cho mô hình và n điểm.

Dựa vào phương pháp ước lượng Bayes như ở đẳng thức (2.50), mật độ là đồng nhất cho đến $x = 8$, nhưng lại có đuôi ở các giá trị cao hơn – một dấu hiệu cho biết ảnh hưởng của $p(\theta)$ trước đó đã chưa bị lấn át bởi tập huấn luyện.



Hình 22 Kết quả ước lượng của phương pháp cực đại likelihood và ước lượng Bayes

Trong khi phương pháp cực đại likelihood ước lượng 1 điểm trong không gian của θ, thì Bayes lại ước lượng một phân bố. Về mặt kỹ thuật, ta không thể so sánh hai ước lượng này. Chỉ khi nào ta tính toán được phân bố $p(x|D)$ như Hình 22 thì việc so sánh mới hợp lý.

Với đa số những mật độ xác suất hay gặp $p(\mathbf{x}|\boldsymbol{\theta})$, chuỗi xác suất hậu định sẽ hội tụ về hàm delta. Nói nôm na, với một số lượng lớn các mẫu, chỉ có một giá trị $\boldsymbol{\theta}$ cho $p(\mathbf{x}|\boldsymbol{\theta})$ thích hợp với dữ liệu, tức là $\boldsymbol{\theta}$ xác định duy nhất từ $p(\mathbf{x}|\boldsymbol{\theta})$. Trong trường hợp này, ta nói $p(\mathbf{x}|\boldsymbol{\theta})$ là khả đồng nhất *(identifiable)*.

Tuy vậy, có những trường hợp nhiều giá trị của $\boldsymbol{\theta}$ có thể cho ta cùng một $p(\mathbf{x}|\boldsymbol{\theta})$. Trong những trường hợp này, $\boldsymbol{\theta}$ không thể được xác định duy nhất từ $p(\mathbf{x}|\boldsymbol{\theta})$ và $p(\mathbf{x}|D^n)$ sẽ có các đỉnh ở gần tất cả những giá trị $\boldsymbol{\theta}$ thích hợp với dữ liệu. May mắn là sự nhập nhằng này sẽ được giải quyết bằng tích phân trong đẳng thức (2.26), vì $p(\mathbf{x}|\boldsymbol{\theta})$ là giống nhau cho mọi giá trị này của $\boldsymbol{\theta}$. Như vậy, $p(\mathbf{x}|D^n)$ sẽ hội tụ về $p(\mathbf{x})$ dù cho $p(\mathbf{x}|\boldsymbol{\theta})$ có khả đồng nhất hay không.

IV.1 Khi nào thì phương pháp cực đại likelihood và Bayes khác nhau?

Trong trường hợp tổng quát, phương pháp cực đại likelihood và phương pháp ước lượng Bayes là tương đương nhau khi tập dữ liệu

huấn luyện là lớn vô tận. Tuy nhiên, trên thực tế những bài toán khá đa dạng và có một tập huấn luyện giới hạn, vì vậy, cũng khá tự nhiên khi có câu hỏi đặt ra là: lúc nào thì cực đại likelihood và ước lượng Bayes cho ra kết quả khác nhau và phương pháp nào tốt hơn.

Có rất nhiều yếu tố ảnh hưởng đến sự chọn lựa. Thứ nhất là độ phức tạp tính toán (xem V.2), và với lý do này thì phương pháp cực đại likelihood được sử dụng nhiều hơn do nó chỉ cần tính toán vi phân hoặc *gradient search* cho $\hat{\boldsymbol{\theta}}$, trong khi phương pháp ước lượng Bayes thì lại cần tích phân đa chiều khá phức tạp. Từ đó dẫn tới một yếu tố khác: khả dịch *(interpretability)*. Trong đa số trường hợp, cực đại likelihood khá dễ hiểu và rõ ràng vì nó cho ra một mô hình tốt nhất trong một tập các mô hình cho trước. Ngược lại, phương pháp Bayes cho ra kết quả là một trung bình trọng số của những mô hình (tham số), dẫn đến một lời giải phức tạp hơn và khó hiểu hơn. Phương pháp Bayes phản ánh độ không chắc chắn trong những mô hình khả thi.

Một điều nữa cần cân nhắc là độ tin cậy với những thông tin cho trước, như là dạng của phân bố $p(\mathbf{x}|\boldsymbol{\theta})$. Lời giải cực đại likelihood $p(\mathbf{x}|\hat{\boldsymbol{\theta}})$ phải bao gồm một giả thiết về dạng tham số; nhưng Bayes thì không cần như vậy. Ta thấy sự khác nhau này ở ví dụ 1, khi mà lời giải Bayes không có dạng tham số như đã giả thiết ban đầu, tức là phân phối đồng nhất $p(\mathbf{x}|D)$. Tổng quát, qua việc sử dụng toàn bộ phân bố của $p(\boldsymbol{\theta}|D)$, phương pháp Bayes sử dụng nhiều thông tin của bài toán hơn là phương pháp cực đại likelihood. (Chẳng hạn ở ví dụ 1, điểm huấn luyện thứ 3 không làm thay đổi kết quả của cực đại likelihood, nhưng với Bayes thì có.) Nếu như những thông tin là đáng tin cậy, Bayes sẽ cho ra kết quả tốt hơn. Hơn nữa, nếu không có thông tin tiền định tường minh thì phương pháp Bayes tương đương với phương pháp cực đại likelihood.

Khi $p(\boldsymbol{\theta}|D)$ rộng hoặc bất đối xứng xung quanh $\hat{\boldsymbol{\theta}}$, hai phương pháp này sẽ có thể cho ra kết quả phân bố $p(\mathbf{x}|D)$ khác nhau. Nếu như sự bất đối xứng khá lớn thì nhìn chung sẽ có thông tin về sự phân bố, ví dụ như là sự bất đối xứng của ngưỡng $\boldsymbol{\theta}$ trong ví dụ 1. Phương pháp Bayes sẽ khai thác được những thông tin như vậy, nhưng với likelihood thì không (ít ra là không trực tiếp). Hơn nữa, phương pháp

Bayes làm rõ hơn về vấn đề của độ lệch và phương sai – nói nôm na là sự cân bằng giữa độ chính xác và sự biến thiên dựa trên độ lớn của tập dữ liệu huấn luyện.

Khi thiết kế một hệ thống phân lớp bằng một trong hai cách này, ta xác định mật độ xác suất hậu định cho việc phân loại vào từng lớp và phân loại một điểm bằng việc cực đại hóa xác suất này. Có 3 nguyên nhân tạo ra lỗi cho hệ thống:

- ***Lỗi Bayes hay là lỗi không phân biệt:*** lỗi này do sự chồng lên nhau của các mật độ $p(\mathbf{x}|\omega_i)$ cho những giá trị i khác nhau. Lỗi này thừa hưởng từ tính chất của bài toán và không thể loại bỏ được.
- ***Lỗi mô hình:*** Lỗi này do mô hình sai. Lỗi này chỉ có thể loại bỏ được nếu người thiết kế chọn một mô hình cụ thể mà bao gồm mô hình thật đã tạo ra dữ liệu. Người thiết kế thường chọn mô hình dựa trên kiến thức về bài toán hơn là dựa trên những phương pháp ước lượng sau đó, và do vậy lỗi mô hình thường hiếm khi khác nhau trong cực đại likelihood và Bayes.
- ***Lỗi ước lượng:*** Lỗi do những tham số được ước lượng từ một tập hữu hạn các mẫu. Lỗi này có thể giảm được bằng cách tăng kích thước của tập dữ liệu huấn luyện.

Sự phân bố tương đối của những lỗi này phụ thuộc vào bài toán. Với một tập dữ liệu huấn luyện vô hạn, lỗi ước lượng sẽ được lại bỏ và tổng số lỗi phân loại sẽ như nhau cho cả hai mô hình.

Tóm lại, dù rằng trong thực tế cực đại likelihood là đơn giản hơn, lại có những quan điểm về lý thuyết và phương pháp luận ủng hộ cho phương pháp Bayes, và khi được sử dụng cho thiết kế những bộ phân lớp, phương pháp này có thể cho ra những kết quả gần chính xác.

V. VẤN ĐỀ VỀ SỐ CHIỀU

Trong những ứng dụng thực tiễn về phân loại đa lớp, việc gặp phải những bài toán liên quan đến năm mươi hay một trăm đặc trưng không phải là bất bình thường, đặc biệt là với những đặc trưng nhị phân. Ta có thể tin rằng từng đặc trưng là hữu dụng cho ít nhất một

vài phân loại; trong khi đó ta cũng nghi ngờ là từng đặc trưng cung cấp những thông tin độc lập, và những đặc trưng thừa không được tính. Và do đó ta phải đương đầu với hai vấn đề. Vấn đề quan trọng là độ chính xác của việc phân loại phụ thuộc như thế nào vào số chiều (và số lượng dữ liệu huấn luyện); thứ hai là độ phức tạp tính toán cho việc thiết kế hệ thống phân loại này.

V.1 Độ chính xác, số chiều và kích thước tập mẫu huấn luyện

Nếu những đặc trưng là độc lập về mặt thống kê thì một vài kết quả lý thuyết cho biết ta có thể có được kết quả phân loại rất tốt. Ví dụ, xem xét hai lớp đa biến phân bố chuẩn cùng hiệp phương sai $p(\mathbf{x} \mid \omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1,2$. Nếu xác suất tiền định bằng nhau, thì cũng không khó để chỉ ra rằng xác suất lỗi được cho bởi

$$P(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du \tag{2.55}$$

với r^2 là bình phương của khoảng cách Mahalanobis

$$r^2 = (\boldsymbol{\mu}_1 - B_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{2.56}$$

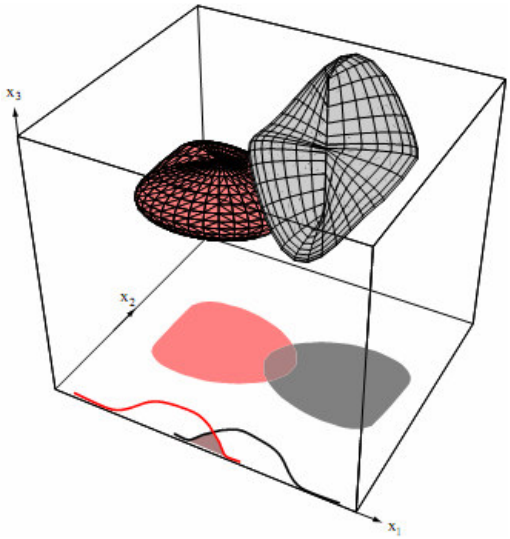
Như vậy, xác suất lỗi giảm khi r tăng, tiệm cận với 0 khi r tiến đến vô cực. Trong trường hợp độc lập có điều kiện, $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_d^2)$

$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2 \tag{2.57}$$

Điều này cho thấy từng đặc trưng đóng góp cho việc giảm xác suất lỗi như thế nào. Một cách tự nhiên, những đặc trưng hữu dụng nhất là những đặc trưng có hiệu trung bình lớn trong mối tương quan với độ lệch chuẩn. Ngoài ra, không có đặc trưng nào là vô dụng nếu trung bình của nó cho hai lớp là khác nhau. Một cách dễ thấy để giảm tỉ lệ lỗi là thêm những đặc trưng độc lập mới.

Nhìn chung, nếu kết quả đạt được với một bộ đặc trưng cho sẵn không đáp ứng đủ, việc xem xét thêm vào những đặc trưng mới là khá tự

nhiên, cụ thể là với những đặc trưng có khả năng tường sự phân biệt trong các trường hợp nhập nhằng. Mặc dù việc tăng thêm số đặc trưng đồng nghĩa với tăng thêm độ phức tạp tính toán cho việc lấy đặc trưng và cả việc phân lớp, nhưng cũng rất có lý khi tin rằng việc này sẽ tăng hiệu quả của việc phân loại. Sau cùng, nếu như các thông tin xác suất của bài toán hoàn toàn được biết trước thì rủi ro Bayes không thể tăng lên khi thêm đặc trưng mới. Trong trường hợp xấu nhất, hệ phân lớp Bayes sẽ bỏ qua đặc trưng mới; nhưng nếu đặc trưng mới cung cấp thêm một thông tin gì đó mới thì hiệu quả ắt sẽ tăng (Hình 23).



Hình 23 Hai phân bố xác suất trong không gian 3-chiều có mật độ không chồng lập lên nhau và do đó lỗi Bayes bị triệt tiêu. Tuy nhiên, nếu chiếu lên mặt phẳng 2-chiều thì sẽ xuất hiện trùng lấp và làm cho lỗi Bayes trở nên lớn.

Không may thay, thực tế lại thường cho thấy, sự thêm vào những đặc trưng mới lại dẫn tới việc hiệu quả xấu hơn. Nghịch lý này cho thấy một vấn đề quan trọng cho việc thiết kế hệ thống phân loại. Nguồn gốc cơ bản của sự khó khăn có thể là từ mô hình sai – như giả định phân bố Gauss hay giả định có điều kiện là sai – hay là số lượng mẫu thiết kế hay mẫu huấn luyện là giới hạn từ đó dẫn đến mật độ không được ước lượng chính xác.

V.2 Độ phức tạp tính toán

Một trong những tác nhân ảnh hưởng đến việc thiết kế hệ phân lớp chính là độ phức tạp tính toán. Đầu tiên, ta cần biết về bậc của một hàm $f(x)$: Chúng ta nói rằng $f(x)$ có bậc của $h(x) - f(x) = O(h(x))$ – khi mà tồn tại một hằng số c_0 và x_0 sao cho $|f(x)| \leq c_0|h(x)|$ với mọi $x > x_0$. Ví dụ, cho $f(x) = a_0 + a_1x + a_2x^2$; trong trường hợp đó ta có $f(x) = O(x^2)$. Rõ ràng là từ định nghĩa trên, bậc của một hàm số không duy nhất. Chẳng hạn với $f(x)$ đã cho sẽ có $O(x^2), O(x^3), O(x^4), O(x^2 lnx)$.

Vì sự không duy nhất này, ta cần cụ thể hơn trong việc mô tả bậc của hàm. Chúng ta nói rằng $f(x) = \Theta(h(x))$ nếu có hằng số x_0, c_1, và c_2 sao cho với $x > x_0$, $f(x)$ luôn luôn nằm giữa $c_1h(x)$ và $c_2h(x)$. Như vậy, hàm bậc hai ở trên chỉ có thể đúng với $f(x) = \Theta(x^2)$, nhưng lại không đúng với $f(x) = \Theta(x^3)$.

Để mô tả độ phức tạp của một thuật toán, ta thường chú ý đến số lượng những phép toán cơ bản như là cộng, nhân và chia cần thực hiện hay là thời gian và bộ nhớ cần trên một máy tính. Để minh hoạ khái niệm này, ta xem xét độ phức tạp của ước lượng cực đại likelihood của tham số cho một hệ phân lớp với các phân bố tiền định là Gauss có trong d-chiều, với n mẫu huấn luyện cho mỗi c lớp. Với mỗi lớp, ta cần tính biệt hàm của (2.58). Độ phức tạp tính toán trong việc tìm trung bình mẫu là $O(nd)$, vì với mỗi chiều, ta cần cộng n giá trị thành phần. Việc chia cho n để lấy trung bình là một phép toán đơn giản và độc lập với số lượng mẫu, nên không ảnh hưởng đến độ phức tạp. Cứ mỗi $d(d+1)/2$ thành phần độc lập của ma trận hiệp phương sai mẫu $\hat{\boldsymbol{\Sigma}}$, thì có n phép nhân và phép cộng (xem (2.19)), cho ta độ phức tạp là $O(d^2n)$. Khi $\hat{\boldsymbol{\Sigma}}$ đã được tính, nó cần $O(d^2)$ phép toán, ta có thể dễ dàng kiểm tra bằng cách đếm số phép toán trong phương pháp quét ma trận. Nghịch đảo có thể được tính trong $O(d^3)$ phép toán, ví dụ như sử dụng phương pháp tối giản Gauss. Đẳng thức (2.58) minh họa những thành phần riêng biệt này trong bài toán thiết lập những tham số cho phân bố chuẩn bằng phương pháp cực đại likelihood

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \overbrace{\hat{\boldsymbol{\mu}}}^{O(dn)})^T \overbrace{\hat{\boldsymbol{\Sigma}}^{-1}}^{O(nd^2)} (\mathbf{x} - \hat{\boldsymbol{\mu}}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \overbrace{\frac{1}{2}\ln|\hat{\boldsymbol{\Sigma}}|}^{O(d^2n)} + \overbrace{\ln P(\omega)}^{O(n)} \quad (2.58)$$

Cũng tự nhiên khi giả sử là $n > d$ (nếu ngược lại thì ma trận hiệp phương sai sẽ không có nghịch đảo), và như vậy đối với những bài toán lớn thì độ phức tạp toàn phần sẽ chủ yếu là từ $O(d^2n)$ trong (2.58) . Điều này áp dụng cho tất cả các phân lớp. Như vậy, độ phức tạp toàn phần cho việc học phân lớp Bayes là $O(cd^2n)$. Vì c là một hằng số nhỏ hơn d^2 và n rất nhiều, nên ta có thể nói độ phức tạp là $O(d^2n)$. Như đã biết ở phần trên là thông thường ta muốn có nhiều dữ liệu huấn luyện hơn nếu không gian có nhiều chiều hơn; việc phân tích độ phức tạp này cho ta thấy chi phí tăng rất nhanh nếu làm như vậy.

Tiếp theo, ta sẽ xem xét về vấn đề ước lượng ma trận hiệp phương sai một cách chi tiết hơn. Điều này yêu cầu ước lượng $d(d+1)/2$ tham số - d thành phần đường chéo và $d(d-1)/2$ thành phần độc lập phía bên. Trước tiên ta quan sát ước lượng cực đại likelihood

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m}_n)(\mathbf{x}_k - \mathbf{m}_n)^T \quad (2.59)$$

gồm $O(nd^2)$ phép toán, tổng của $n-1$ ma trận $d \times d$ độc lập *rank* một, và vì vậy chắc chắn sẽ là *singular* nếu $n \leq d$. Do đó ta cần ít nhất $d+1$ mẫu.

Đương nhiên, độ phức tạp tính toán cho việc phân loại sẽ ít hơn. Cho một điểm **x**, ta phải tính $(\mathbf{x} - \hat{\boldsymbol{\mu}})$ với $O(d)$ phép tính. Hơn nữa, với mỗi phân lớp, ta phải nhân nghịch đảo ma trận hiệp phương sai với từng vector riêng rẽ, $O(d^2)$ phép tính. Tìm $max_i g_i(x)$ cần $O(c)$ phép so sánh. Vì c nhỏ, nên ta chỉ còn $O(d^2)$ phép tính. Như vậy, việc phân loại sẽ đơn giản hơn và nhanh hơn là học. Độ phức tạp của trường hợp tương ứng cho việc học Bayes là tương đương với phương pháp cực đại likelihood. Mặc dù vậy, phương pháp học Bayes có độ phức tạp cao hơn vì phải tích hợp những mô hình tham số $\boldsymbol{\theta}$.

Cũng có khi ta nhấn mạnh đến độ phức tạp không gian và thời gian, việc này cũng quan trọng khi cài đặt tính toán song song. Chẳng hạn,

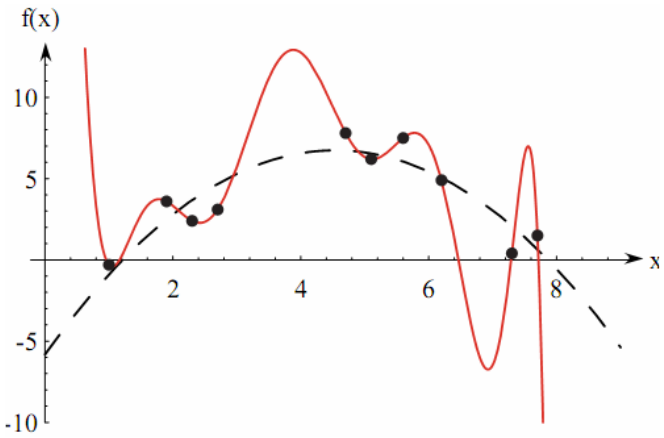
trung bình mẫu của một phân lớp có thể được tính bằng *d* bộ xử lý riêng biệt, mỗi cái cộng *n* giá trị mẫu. Như vậy, ta có thể mô tả phương pháp cài đặt này là *O(d)* trong không gian (tức là lượng bổ nhớ hay số lượng bộ xử lý) và *O(n)* trong thời gian (tức là số lượng các bước thực hiện). Đương nhiên, với một số thuật toán nhất định, có thể cân nhắc để đánh đổi giữa không gian và thời gian. Chẳng hạn sử dụng một bộ xử lý nhiều lần hay sử dụng nhiều bộ xử lý để có thời gian nhanh hơn.

V.3 Quá khớp (overfitting)

Có một vấn đề rất hay gặp phải là số lượng mẫu có sẵn không đủ, và câu hỏi đặt ra là làm sao hệ phân loại vẫn hoạt động hiệu quả. Một cách là ta sẽ giảm số đặc trưng bằng cách thiết kế lại bộ rút trích đặc trưng, lựa chọn lại một tập đặc trưng con của các đặc trưng có sẵn hay gộp một vài đặc trưng lại theo cách nào đó. Một cách khác là ta giả sử tất cả các phân lớp đều chia sẻ chung một ma trận hiệp phương sai và chia sẻ chung dữ liệu có sẵn. Một phương án khác là tìm một ước lượng tốt hơn cho Σ. Nếu một ước lượng hợp lý cho Σ_0 có sẵn, ước lượng Bayes hay giả-Bayes *(pseudo-Bayesian)* của dạng thức $\lambda\Sigma_0 + (1-\lambda)\hat{\Sigma}$ sẽ được chọn. Nếu Σ_0 là ma trận đường chéo thì điều này có thể loại bỏ ảnh hưởng xấu của những sự tương quan tình cờ. Có thể giảm cơ hội của sự tương quan bằng một heuristic sử dụng ngưỡng cho ma trận hiệp phương sai các mẫu. Chẳng hạn như, ta có thể giả thiết rằng tất cả các hiệp phương sai mà ở đó độ lớn của các hệ số tương đồng không gần 1 sẽ là 0. Cách tiếp cận này giả thiết sự độc lập trong thống kế, nhờ đó làm cho tất cả những phần tử không nằm trên đường chéo bằng 0 bất chấp những bằng chứng thực nghiệm ngược lại – tốn *O(nd)* phép tính. Mặc dù giả thiết như vậy hầu như chắc chắn là sai, nhưng kết quả ướng lượng heuristic đôi khi lại cho hiệu quả tốt hơn là ước lượng cực đại likelihood trên toàn bộ không gian tham số.

Ở đây ta lại vấp phải một nghịch lý khác. Hệ phân lớp có kết quả từ giả thiết độc lập lại gần như chắc chắn không tối ưu. Cũng dễ hiểu khi nó có hiệu quả hơn nếu như những đặc trưng này thật sự độc lập, nhưng làm sao nó có thể cho kết quả tốt hơn nếu như giả thiết của nó

đã không đúng? Câu trả lời có liên quan đến sự thiếu dữ liệu, và bản chất của nó tương tự như vấn đề về ước lượng đường cong *(curve fitting)*. Hình 24 cho ta thấy một bộ 10 điểm và hai đường cong được xem xét để khớp với chúng. Những điểm dữ liệu có được từ cộng nhiễu độc lập chuẩn trung bình 0 vào một parabol. Như vậy, trong tất cả những đa thức, parabol có thể cho kết quả tốt nhất, giả sử rằng ta cũng muốn những điểm trong tương lai cũng khớp với đường cong chứ không chỉ những điểm đang có. Mặc dù đường thẳng cũng hợp với nó khá tốt, parabol cho kết quả tốt hơn, nhưng cũng có thể đặt câu hỏi là liệu dữ liệu có đủ để ước lượng đường cong hay chưa. Một parabol tốt nhất cho một bộ dữ liệu lớn hơn có thể khá khác biệt, và trên hết một đường thẳng có vẻ như thắng thế. Một đa thức bậc 10 sẽ hoàn toàn vừa vặn với những điểm đã cho. Mặc dù vậy, chúng ta không mong đợi một đa thức bậc 10 ở đây. Nhìn chung, một nội suy hay ngoại suy đáng tin cậy không thể có được trừ khi lời giải là quá chắc chắn, tức là có nhiều điểm hơn là số lượng tham số trong hàm.



Hình 24 Những điểm huấn luyện (điểm đen) được chọn từ hàm bậc 2 có thêm nhiễu Gauss, nghĩa là $f(x)=ax^2+bx+c+\varepsilon$ với $\varepsilon \sim N(0,\sigma^2)$. Một đa thức bậc 10 khớp với tập điểm một cách hoàn hảo nhưng thực tế là ta cần một hàm bậc 2 vì nó sẽ khớp hơn với các điểm huấn luyện mới.

Để tìm được một đường cong khớp với các điểm trong Hình 24, ta có thể bắt đầu xem xét với một đa thức bậc cao (chẳng hạn, bậc 10), và từ từ đơn giản mô hình lại bằng cách bỏ những thành phần bậc cao nhất.

Mặc dù điều này có thể dẫn đến sai số trong "tập huấn luyện", nhưng ta mong đợi một sự tổng quát hoá cao hơn.

Một cách tương tự, có một vài phương pháp heuristic có thể sử dụng trong trường hợp hệ phân loại Gauss. Chẳng hạn, giả sử ta muốn thiết kế một hệ phân loại cho phân bố $N(\boldsymbol{\mu}_1, \Sigma_1)$ và $N(\boldsymbol{\mu}_2, \Sigma_2)$ và ta nghĩ rằng ta không có đủ dữ liệu để ướng lượng chính xác các tham số. Ta có thể giả thiết rằng hai phân bố này có cùng hiệp phương sai, tức là $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ và $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ và ta ước lượng $\boldsymbol{\Sigma}$ theo đó. Những ước lượng như vậy luôn cần chuẩn hoá dữ liệu.

Một phương pháp trung gian là giả thiết một sự kết hợp có trọng số giữa các hiệp phương sai bằng nhau, kỹ thuật này được gọi là *shrinkage*, vì từng hiệp phương sai được co lại thành một cái chung. Nếu *i* là số hiệu của *c* lớp đang nghi vấn, ta có

$$\boldsymbol{\Sigma}_i(\alpha) = \frac{(1-\alpha)n_i\boldsymbol{\Sigma}_i + \alpha n\boldsymbol{\Sigma}}{(1-\alpha)n_i + \alpha n} \tag{2.60}$$

với $0 < \alpha < 1$. Ta cũng có thể gom ước lượng của ma trận hiệp phương sai chung (theo giả thiết) tiến đến ma trận đơn vị như sau

$$\boldsymbol{\Sigma}(\beta) = (1-\beta)\boldsymbol{\Sigma} + \beta\mathbf{I} \tag{2.61}$$

với $0 < \beta < 1$.

VI. KẾT LUẬN

Nếu ta đã biết dạng tham số của các mật độ khi biết phân lớp thì ta có thể suy giảm từ bài toán xác định phân bố xuống bài toán dễ hơn nhiều là xác định các tham số (thường được ký hiệu là $\boldsymbol{\theta}_i$ cho mỗi lớp ω_i). Trong chương này, với điều giả thiết như vậy, ta đã khảo sát hai phương pháp phổ biến là

- ***Cực đại likelihood:*** phương pháp này tìm kiếm giá trị các tham số phù hợp nhất cho tập huận luyện – như tên gọi của nó là làm cực đại hóa xác suất để có được tập huấn luyện.

- ***Ước lượng Bayes:*** trong phương pháp này, các tham số chưa biết được thể hiện như là các biến ngẫu nhiên với các mật độ tiền định cho trước. Các mẫu huấn luyện có vai trò là chuyển đổi mật độ tiền định này sang mật độ hậu định. Hồi quy Bayes, khi có một mẫu mới thêm vào, giúp cập nhật ước lượng tham số.

Trong 2 phương pháp trên, ước lượng Bayes được ưa thích hơn về mặt lý thuyết. Tuy nhiên, về mặt thực tiễn, cực đại likelihood, nhìn chung, dễ thực thi hơn và nếu có được một tập huấn luyện lớn thì sẽ có được hệ phân lớp gần như chính xác.

VII. BÀI TẬP

ƯỚC LƯỢNG CỰC ĐẠI LIKELIHOOD

Bài 1

Giả sử x có phân phối mật độ hàm mũ như sau

$$p(x \mid \theta) = \begin{cases} \theta e^{-\theta x} & if\ x \geq 0 \\ 0 & otherwise \end{cases}$$

1. Vẽ p(x|θ) theo x với θ=1, theo θ (0≤θ≤5) với x=2.
2. Giả sử có n mẫu x_i được lấy độc lập từ p(x|θ). Chứng minh là ước lượng cực đại likelihood cho θ được cho bởi

$$\theta = \frac{1}{1/n \sum_{i=1}^{n} x_i}$$

Giải:

1. (sử dụng R để vẽ)

2. Đặt D = $\{x_1, x_2, \ldots, x_n\}$

$$\ln p(D \mid \theta) = \sum_{i=1}^{n} \ln p(x_i \mid \theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

$$\nabla_\theta \ln p(D \mid \theta) = n.\frac{1}{\theta} - \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \theta = \frac{1}{1/n \sum_{i=1}^{n} x_i}$$

Bài 2

Cho x có phân bố đồng nhất như sau

$$p(x \mid \theta) = \begin{cases} 1/\theta & if\ 0 \le x \le \theta \\ 0 & otherwise \end{cases}$$

1. Giả sử có tập n mẫu D = $\{x_1, x_2, \ldots, x_n\}$ được lấy độc lập từ p(x|θ). Chứng minh là ước lượng cực đại likelihood cho θ chính bằng max(D).
2. Cho 5 điểm được lấy độc lập từ p(x|θ) và điểm có giá trị lớn nhất là 0.6. Hãy vẽ likelihood p(D|θ) với $0 \le \theta \le 1$ và giải thích tại sao ta chỉ cần biết điểm có giá trị lớn nhất mà không quan tâm đến 4 điểm còn lại.

Giải:

1. Nếu tồn tại $x_k > \theta$ thì $p(x_k|\theta) = 0$ dẫn tới

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = 0$$

Do đó $\max(D) \le \theta$. Khi đó

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \frac{1}{\theta^n}$$

lớn nhất khi θ=max(D).

2. (sử dụng R để vẽ).

Bài 3

Cực đại likelihood còn có thể được áp dụng để tìm xác suất tiền định. Giả sử ta cần tìm xác suất tiền định của lớp ω_i là $P(\omega_i)$. Ta lấy mẫu tuần tự, chọn một cách độc lập. Ký hiệu z_{ik} là biến quan sát với $z_{ik}=1$ nếu mẫu thứ k thuộc về lớp ω_i và $z_{ik}=0$ nếu ngược lại.

1. Chứng minh

$$P(z_{i1}, z_{i2}, ..., z_{in} \left| P(\omega_i)) = \prod_{k=1}^{n} P(\omega_i)^{z_{ik}} (1 - P(\omega_i))^{1-z_{ik}}\right.$$

2. Chứng minh là ước lượng cực đại likelihood cho $P(\omega_i)$ là

$$\hat{P}(\omega_i) = \frac{1}{n} \sum_{k=1}^{n} z_{ik}$$

Hãy diễn dịch kết quả bằng lời.

Giải:

1. Với mỗi mẫu được chọn thì do $z_{ik}=1$ nếu mẫu thứ k thuộc về lớp ω_i và $z_{ik}=0$ nếu ngược lại nên

$$P(z_{ik} \left| P(\omega_i)) = P(\omega_i)^{z_{i1}} (1 - P(\omega_i))^{1-z_{i1}}\right.$$

Do lấy mẫu độc lập nên

$$P(z_{i1}, z_{i2}, ..., z_{in} \left| P(\omega_i)) = \prod_{k=1}^{n} P(z_{ik} \right| P(\omega_i))$$
$$= \prod_{k=1}^{n} P(\omega_i)^{z_{ik}} (1 - P(\omega_i))^{1-z_{ik}}$$

2. Ta có

$$\ln P(z_{i1}, z_{i2}, ..., z_{in} \left| P(\omega_i)\right.) = \sum_{k=1}^{n} (z_{ik} \ln P(\omega_i) + (1 - z_{ik}) \ln(1 - P(\omega_i))$$

$$= \ln P(\omega_i) \sum_{k=1}^{n} z_{ik} + \ln(1 - P(\omega_i)) \sum_{k=1}^{n} (1 - z_{ik})$$

$$\nabla_{P(\omega_i)} \ln P(z_{i1}, z_{i2}, ..., z_{in} \left| P(\omega_i)\right.) = \frac{1}{P(\omega_i)} \sum_{k=1}^{n} z_{ik} - \frac{1}{1 - P(\omega_i)} \sum_{k=1}^{n} (1 - z_{ik})$$

$$\nabla_{P(\omega_i)} \ln P(z_{i1}, z_{i2}, ..., z_{in} \left| P(\omega_i)\right.) = 0 \Leftrightarrow P(\omega_i) = \frac{1}{n} \sum_{k=1}^{n} z_{ik}$$

Để ước lượng xác suất tiền định của lớp ω_i, ta lấy độc lập n mẫu và đếm xem có bao nhiêu mẫu thuộc lớp ω_i. Khi đó, xác suất tiền định $P(\omega_i)$ được ước lượng bằng tỉ số của số mẫu thuộc lớp ω_i so với n.

Bài 4

Cho **x** là vector nhị phân (0 hoặc 1) ngẫu nhiên d-chiều có phân bố Bernoulli

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i} (1 - \theta_i)^{1 - x_i}$$

trong đó $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_d)$ là vector tham số chưa biết với θ_i là xác suất để $x_i=1$. Chứng minh rằng ước lượng cực đại likelihood của **θ** là

$$\hat{\boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_n$$

Giải:

Xét tập n mẫu $D = \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ được lấy độc lập từ $p(\mathbf{x}|\boldsymbol{\theta})$ ta có

$$p(D \left| \boldsymbol{\theta}\right.) = \prod_{k=1}^{n} p(\mathbf{x}_k \left| \boldsymbol{\theta}\right.) = \prod_{k=1}^{n} \prod_{i=1}^{d} \theta_i^{x_{k,i}} (1 - \theta_i)^{1 - x_{k,i}}$$

$$= \prod_{i=1}^{d} \underbrace{\prod_{k=1}^{n} \theta_i^{x_{k,i}} (1 - \theta_i)^{1 - x_{k,i}}}_{p(D_i \mid \theta_i)}$$

Áp dụng Bài 3 cho từng $p(D_i|\theta_i)$ ta được điều cần tìm.

Bài 5

Cho từng thành phần x_i của **x** có giá trị nhị phân (0 hoặc 1). Cho rằng xác suất để có giá trị 1 tại mỗi thành phần là

$$p_{i1} = p,\ p_{i2} = 1 - p$$

Xác suất lỗi được biết là tiến tới 0 khi số chiều tăng đến vô hạn. Bài này sẽ khảo sát vấn đề khi tăng số đặc trưng trong khi số lượng mẫu vẫn là 1.

1. Cho một mẫu $\mathbf{x}=(x_1,x_2,\ldots,x_d)^T$ được lấy từ lớp ω_1. Chứng minh rằng ước lượng cực đại likelihood cho p được tính bởi

$$\hat{p} = \frac{1}{d}\sum_{i=1}^{d} x_i$$

2. Hãy cho biết $\hat{p}$ sẽ như thế nào nếu d tiến đến vô hạn. Giải thích tại sao việc đó có nghĩa là nếu số lượng đặc trưng tiến đến vô hạn thì ta có được hệ phân lớp không có lỗi sai mặc dù rằng ta chỉ có một mẫu duy nhất từ mỗi phân lớp.

Giải:

1. Áp dụng Bài 3.

2. Khi d tiến đến vô hạn thì ước lượng cực đại likelihood sẽ cho ra kết quả chính xác, tức là $\hat{p}$ =p. Khi đó, từ điều đã biết là "Xác suất lỗi được biết là tiến tới 0 khi số chiều tăng đến vô hạn" kết hợp với $\hat{p}$ =p ta được hệ phân lớp không có lỗi sai mặc dù chỉ có một mẫu duy nhất cho mỗi phân lớp.

Bài 6

Bài tập này là một ví dụ để cho thấy nếu mô hình tệ thì ước lượng cực đại likelihood sẽ cho hệ phân lớp không tốt – thậm chí không phải là tốt nhất trong tập mô hình tệ.

Giả sử có hệ phân 2 lớp với $P(\omega_1) = P(\omega_2) = 0.5$. Ta biết là $p(x|\omega_1) \sim N(0,1)$ nhưng giả định rằng $p(x|\omega_2) \sim N(\mu,1)$. Nhưng trên thực tế thì $p(x|\omega_2) \sim N(1,10^6)$.

1. Tìm ước lượng cực đại likelihood $\hat{\mu}$ nếu ta có một tập rất lớn dữ liệu.
2. Tìm biên ra quyết định từ mô hình ước lượng này.
3. Tim biên ra quyết định với mô hình đúng thực tế của nó: $p(x|\omega_2) \sim N(1,10^6)$.
4. Từ mô hình tệ $p(x|\omega_2) \sim N(\mu,1)$ và kết quả đạt được ở (3) hãy tìm ước lượng của μ cho lỗi sai thấp hơn so với mô hình ước lượng ở (2).

Giải:

1. Với mô hình đúng là $p(x|\omega_2) \sim N(1,10^6)$ thì tập dữ liệu rất lớn bất kỳ $D = \{x_1,x_2,\ldots,x_n\} \sim N(1,10^6)$ có ước lượng cực đại likelihood trung bình là $\hat{\mu} = 1$.

2. Với mô hình ước lượng cực đại likelihood $p(x|\omega_2) \sim N(1,1)$ và do các xác suất tiền định bằng nhau nên biên ra quyết định là

$$x_0 = \frac{1}{2}(\mu_1 + \mu_2) = 0.5$$

3. Với mô hình đúng thực tế $p(x|\omega_2) \sim N(1,10^6)$, biệt hàm của 2 lớp lần lượt là

$$g_1(x) = -\frac{1}{2}x^2 + \ln P(\omega_1)$$

$$g_2(x) = -\frac{1}{2.10^6}(x-1)^2 - \frac{1}{2}\ln 10^6 + \ln P(\omega_2)$$

do xác suất tiền định bằng nhau nên suy ra biên ra quyết định là

$$g_1(x) = g_2(x)$$
$$\Leftrightarrow -\frac{1}{2}x^2 = -\frac{1}{2.10^6}(x-1)^2 - \frac{1}{2}\ln 10^6$$
$$\Leftrightarrow x_1 \approx 3.7169,\ x_2 \approx -3.7169$$

4. Như vậy, từ kết quả của (3), với mô hình $p(x|\omega_2) \sim N(\mu,1)$ thì ước lượng μ sao cho biên ra quyết định ở x_1, tức là $\hat{\mu} = 2x_1 - \mu_1 \approx 7.4338$, sẽ cho lỗi sai thấp hơn so với ước lượng cực đại likelihood.

Bài 7

Xét bài toàn phân 2 lớp với $P(\omega_1) = P(\omega_2) = 0.5$. Giả sử dữ liệu có phân bố thực sự là

$$p(x|\omega_1) \sim [(1-k)\delta(x-1) + k\delta(x+X)]$$
$$p(x|\omega_2) \sim [(1-k)\delta(x+1) + k\delta(x-X)]$$

trong đó X là một số dương, $0 \le k < 0.5$, δ(.) là hàm Dirac delta. Giả sử mô hình của ta là $p(x|\omega_1,\mu_1) \sim N(\mu_1,\sigma_1^2)$ và $p(x|\omega_2,\mu_2) \sim N(\mu_2,\sigma_2^2)$ và ta xây dựng hệ phân lớp với ước lượng cực đại likelihood.

1. Xem xét tính đối xứng của bài toán và chứng minh rằng với tập dữ liệu lớn vô hạn, biên ra quyết định luôn luôn là x=0, bất chấp giá trị của k và X.
2. Tìm sự phụ thuộc của X(k) vào k để đảm bảo rằng ước lượng trung bình $\hat{\mu}_1$ của $p(x|\omega_1,\mu_1)$ nhỏ hơn 0 (từ tính đối xứng suy ra $\hat{\mu}_2 > 0$)
3. Từ X(k) tìm được ở (2), hãy tìm lỗi sai phân lớp theo biến k.

Giải:

1. Rõ ràng là mật độ thực sự của $p(x|\omega_1)$ và $p(x|\omega_2)$ là đối xứng qua x=0.

Nhớ lại, nếu ta có tập dữ liệu lớn vô hạn thì ước lượng cực đại likelihood đạt được đúng bằng với giá trị thực của nó. Do đó, với mô

hình tệ $p(x|\omega_1,\mu_1) \sim N(\mu_1,\sigma_1^2)$ và $p(x|\omega_2,\mu_2) \sim N(\mu_2,\sigma_2^2)$ ta được ước lượng cực đại likelihood thỏa $\hat{\sigma}_1^2 = \hat{\sigma}_2^2$ và $\hat{\mu}_1, \hat{\mu}_2$ đối xứng qua x = 0. Do đó biên ra quyết định là x = 0 bất chấp giá trị k, X.

2. Với tập dữ liệu lớn vô hạn thì ước lượng cực đại likelihood bằng với giá trị thực của nó, do đó

$$\hat{\mu}_1 = \int xp(x|\omega_1)dx = \int x[(1-k)\delta(x-1) + k\delta(x+X)]dx$$
$$= (1-k) - kX$$

Để $\hat{\mu}_1$ nhỏ hơn 0 thì

- nếu k = 0 thì không thể có được,
- ngược lại thì X > (1-k)/k.

3…

Bài 8

Cho $p(\mathbf{x}|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}|\boldsymbol{\Sigma})$ trong đó $\boldsymbol{\mu}$ đã biết còn $\boldsymbol{\Sigma}$ chưa biết. Chứng minh rằng ước lượng cực đại likelihood của Σ là

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T$$

theo các bước hướng dẫn sau

1. Chứng minh là hàm likelihood có thể được viết dưới dạng sau

$$p(\mathbf{x}_1,...,\mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}}\left|\boldsymbol{\Sigma}^{-1}\right|^{n/2}\exp\left[-\frac{1}{2}tr\left[\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T\right]\right]$$

trong đó tr[A] (trace) là tổng đường chéo của ma trận A bằng cách sử dụng tính chất $a^TAa = tr[Aaa^T]$.

2. Đặt $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$ và $\lambda_1,\ldots,\lambda_d$ là các eigenvalues của A; từ kết quả (2) chứng minh

$$p(\mathbf{x}_1,...,\mathbf{x}_n \mid \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}\left|\hat{\boldsymbol{\Sigma}}\right|^{n/2}}(\lambda_1...\lambda_d)^{n/2}\exp\left[-\frac{n}{2}(\lambda_1+...+\lambda_d)\right]$$

Sử dụng tính chất: |AB| = |A||B|, nếu $\lambda_1,\ldots,\lambda_d$ là các eigenvalue của ma trận A thì $|A| = \lambda_1\ldots\lambda_d$ và $tr[A] = \lambda_1+\ldots+\lambda_d$

3. Chứng minh rằng cực đại likelihood đạt được bằng cách chọn $\lambda_1 = \ldots = \lambda_d = 1$. Từ đó suy ra ước lượng cực đại likelihood của $\boldsymbol{\Sigma}$.

Giải:

1. Ta có

$$p(\mathbf{x}_1,...,\mathbf{x}_n \mid \boldsymbol{\Sigma}) = \prod_{k=1}^{n} p(\mathbf{x}_1 \mid \boldsymbol{\Sigma})$$

$$= \frac{1}{(2\pi)^{nd/2}}\left|\boldsymbol{\Sigma}^{-1}\right|^{n/2}\exp[-\frac{n}{2}\sum_{k=1}^{n}\underbrace{(\mathbf{x}_k-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})^T}_{tr[\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^T]}]$$

$$= \frac{1}{(2\pi)^{nd/2}}\left|\boldsymbol{\Sigma}^{-1}\right|^{n/2}\exp\left[-\frac{n}{2}tr\left[\underbrace{\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^T}_{\mathbf{A}}\right]\right]$$

2. Do $\left|\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right| = |\mathbf{A}| = \lambda_1...\lambda_n \Rightarrow \left|\boldsymbol{\Sigma}^{-1}\right|^{n/2} = (\lambda_1...\lambda_n)^{n/2} / \left|\hat{\boldsymbol{\Sigma}}\right|^{n/2}$ và do tr[A] = $\lambda_1+\ldots+\lambda_d$ nên suy ra

$$p(\mathbf{x}_1,...,\mathbf{x}_n \mid \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}\left|\hat{\boldsymbol{\Sigma}}\right|^{n/2}}(\lambda_1...\lambda_d)^{n/2}\exp\left[-\frac{n}{2}(\lambda_1+...+\lambda_d)\right]$$

3. Hàm likelihood đạt cực đại khi $(\lambda_1...\lambda_d)^{n/2}\exp\left[-\frac{n}{2}(\lambda_1+...+\lambda_d)\right]$ đạt giá trị lớn nhất. Ta có

$$(\lambda_1...\lambda_d)^{n/2}\exp\left[-\frac{n}{2}(\lambda_1+...+\lambda_d)\right]=\left(\prod_{i=1}^{d}(\lambda_i e^{-\lambda_i})\right)^{n/2}$$

Khảo sát hàm $f(x)=xe^{-x}$ đạt giá trị lớn nhất tại x=1. Do đó vế trái đạt được giá trị lớn nhất khi $\lambda_1 = \ldots = \lambda_d = 1$.

Khi này, $\boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}$ là một lời giải.

ƯỚC LƯỢNG BAYES

Bài 9

Xét bài toán học trung bình của phân phối chuẩn đơn biến. Giả sử $n_0 = \sigma^2/\sigma_0^2$ và tưởng tượng là μ_0 là trung bình của n_0 mẫu giả định x_k, k = -n_0+1,..., 0.

1. Chứng minh là từ (2.32) và (2.33) ta suy ra được

$$\mu_n=\frac{1}{n+n_0}\sum_{k=-n_0+1}^{n}x_k,\quad \sigma_n^2=\frac{\sigma^2}{n+n_0}$$

2. Hãy diễn giải kết quả trên theo $p(\mu) \sim N(\mu_0,\sigma_0^2)$.

Giải:

1.

$$\frac{1}{\sigma_n^2}=\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\Rightarrow\sigma_n^2=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}=\frac{\sigma^2}{n+n_0}$$

$$\frac{\mu_n}{\sigma_n^2}=\frac{n}{\sigma_2}\bar{x}_n+\frac{\mu_0}{\sigma_0^2}\Rightarrow\mu_n=\left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0=\frac{1}{n+n_0}\sum_{k=-n_0+1}^{n}x_k$$

2. Với σ^2 cố định. Quan sát từ công thức ta thấy khi σ_0^2 tăng thì n_0 giảm dẫn tới số mẫu để ước lượng cho μ_0 giảm, do đó μ_0 giảm ảnh hưởng trên μ_n. Về mặt cảm tính thì do độ không chắc chắn của μ_0 tăng nên rõ ràng là μ_n sẽ ít bị chi phối bởi μ_0. Điều ngược lại cũng lý giải tương tự: khi σ_0^2 giảm thì μ_n sẽ phụ thuộc nhiều hơn vào μ_0.

Bài 10

Giả sử ta có tập dữ liệu huấn luyện gồm n mẫu D = $\{x_1,\ldots,x_n\}$ có phân phối chuẩn trung bình $\boldsymbol{\mu}$ chưa biết và hiệp phương sai $\boldsymbol{\Sigma}$ đã biết. Một ước lượng cực đại xác suất hậu định *(maximum a porteriori - MAP)* của $\boldsymbol{\mu}$ là $\hat{\boldsymbol{\mu}} = \arg\max_{\mu} p(\boldsymbol{\mu} \mid D)$.

Hãy tìm MAP của $\boldsymbol{\mu}$ trong 2 trường hợp sau và so sánh với nếu ta áp dụng ước lượng cực đại likelihood và ước lượng Bayes.

1. Khi không biết bất kỳ thông tin tiền định nào của $\boldsymbol{\mu}$, người ta thường cho p($\boldsymbol{\mu}$) là phân phối đồng nhất, hoặc phẳng.

2. Giả sử biết $\boldsymbol{\mu}$ đặc trưng bởi mật độ Gauss N($\mathbf{m}_0$,$\boldsymbol{\Sigma}_0$).

Giải:

Nhận thấy

$$\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} p(\boldsymbol{\mu} \mid D) = \arg\max_{\boldsymbol{\mu}} \frac{p(D \mid \boldsymbol{\mu}) p(\boldsymbol{\mu})}{\int p(D \mid \boldsymbol{\mu}) p(\boldsymbol{\mu}) d\boldsymbol{\mu}}$$

$$= \arg\max_{\boldsymbol{\mu}} p(D \mid \boldsymbol{\mu}) p(\boldsymbol{\mu})$$

Đặt $l(\boldsymbol{\mu}) = \ln p(D \mid \boldsymbol{\mu}) p(\boldsymbol{\mu}) = \ln p(\boldsymbol{\mu}) + \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu})$

$\Rightarrow \hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} l(\boldsymbol{\mu})$

1. Do p($\boldsymbol{\mu}$) là phẳng nên suy ra $\hat{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} \ln p(D \mid \boldsymbol{\mu})$, và đây chính là ước lượng cực đại likelihood. Do đó $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$

2. p($\boldsymbol{\mu}$) = N($\mathbf{m}_0$,$\boldsymbol{\Sigma}_0$) và p(x|$\boldsymbol{\mu}$) = N($\boldsymbol{\mu}$,$\boldsymbol{\Sigma}$) nên

$$\begin{aligned} l(\boldsymbol{\mu}) &= \ln p(\boldsymbol{\mu}) + \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\mu}) \\ &= -\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}_0\right|\right] - \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m}_0) + \\ &\sum_{k=1}^{n}\left(-\frac{1}{2}\ln\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right] - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right) \\ &= \alpha - \frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m}_0) - \sum_{k=1}^{n}\left(-\frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right) \end{aligned}$$

trong đó α là một hằng số. Lấy đạo hàm $l(\boldsymbol{\mu})$ theo $\boldsymbol{\mu}$ ta được

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} l(\boldsymbol{\mu}) &= \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) - \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m}_0) \\ &= -\left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{\mu} + \left(\boldsymbol{\Sigma}_0^{-1}\mathbf{m}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right) \end{aligned}$$

trong đó $\bar{\mathbf{x}}$ là trung bình mẫu. Để đạt được giá trị lớn nhất thì

$$\nabla_{\boldsymbol{\mu}} l(\hat{\boldsymbol{\mu}}) = 0 \Leftrightarrow \hat{\boldsymbol{\mu}} = \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\mathbf{m}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right)$$

Để ý thấy rằng đây chính là trung bình của $\boldsymbol{\mu}$ có được theo phương pháp ước lượng Bayes.

Bài 11

Bài tập này liên quan đến hệ phân lớp Bayes cho trường hợp phân lớp đa biến d-chiều Bernoulli (xem thêm Bài 4). Ta thực hiện trên từng

phân lớp riêng biệt và ngầm định rằng P(**x**|D) là P(**x**|D_i,ω_i). Xác suất có điều kiện cho một phân lớp biết trước được cho bởi

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$$

Cho tập mẫu D = {x_1,…,x_n} được lấy độc lập theo mật độ phân phối trên.

1. Giả sử **s** = (s_1,…,s_d)^T là tổng của n mẫu. Chứng minh

$$P(D|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i}$$

2. Giả sử θ có phân phối đồng nhất, sử dụng

$$\int_0^1 \theta^m (1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!}$$

để chứng minh

$$p(\boldsymbol{\theta} \mid D) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \theta_i^{s_i} (1-\theta_i)^{n-s_i}$$

3. Lấy tích phân P(**x**|**θ**)p(**θ**|D) theo **θ** để có được xác suất có điều kiện

$$P(\mathbf{x}|D) = \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i} \left(1-\frac{s_i+1}{n+2}\right)^{1-x_i}$$

4. Nếu ta nghĩ đến việc có được P(x|D) bằng cách thay thế ước lượng $\hat{\boldsymbol{\theta}}$ cho **θ** trong P(x|**θ**) thì ước lượng Bayes hiệu quả cho **θ** là gì?

Giải:

1. Ta có

$$P(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} P(\mathbf{x}_k|\boldsymbol{\theta}) = \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}} (1-\theta_i)^{1-x_{ki}}$$
$$= \prod_{i=1}^{d}\prod_{k=1}^{n} \theta_i^{x_{ki}} (1-\theta_i)^{1-x_{ki}} = \prod_{i=1}^{d} \theta_i^{\sum_{k=1}^{n} x_{ki}} (1-\theta_i)^{1-\sum_{k=1}^{n} x_{ki}}$$
$$= \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i}$$

2. Ta có

$$p(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int P(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i}}{\int_0^1 \ldots \int_0^1 \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta_1 \ldots d\theta_d}$$

$$\int_0^1 \ldots \int_0^1 \prod_{i=1}^{d} \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta_1 \ldots d\theta_d = \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta_i = \prod_{i=1}^{d} \frac{s_i!(n-s_i)!}{(n+1)!}$$

$$\Rightarrow p(\boldsymbol{\theta}|D) = \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \theta_i^{s_i} (1-\theta_i)^{n-s_i}$$

3.

$$P(\mathbf{x}|D) = \int P(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$$
$$= \int_0^1 \ldots \int_0^1 \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i} \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \theta_i^{s_i} (1-\theta_i)^{n-s_i} d\theta_1 \ldots d\theta_d$$
$$= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \prod_{i=1}^{d} \int_0^1 \theta_i^{x_i+s_i} (1-\theta_i)^{1+n-(x_i+s_i)} d\theta_i$$
$$= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!} \prod_{i=1}^{d} \frac{(x_i+s_i)!(1+n-x_i-s_i)!}{(n+2)!}$$
$$= \prod_{i=1}^{d} \frac{(x_i+s_i)!(1+n-x_i-s_i)!}{s_i!(n-s_i)!(n+2)}$$

Do mỗi x_i chỉ nhận giá trị là 0 hoặc 1

- nếu $x_i=0$ thì

$$\frac{(x_i+s_i)!(1+n-x_i-s_i)!}{s_i!(n-s_i)!(n+2)} = \frac{s_i!(1+n-s_i)!}{s_i!(n-s_i)!(n+2)}$$
$$= \frac{1+n-s_i}{n+2} = \left(\frac{s_i+1}{n+2}\right)^0 \left(1-\frac{s_i+1}{n+2}\right)^{1-0}$$

- nếu $x_i=1$ thì

$$\frac{(x_i+s_i)!(1+n-x_i-s_i)!}{s_i!(n-s_i)!(n+2)} = \frac{(s_i+1)!(n-s_i)!}{s_i!(n-s_i)!(n+2)}$$
$$= \frac{s_i+1}{n+2} = \left(\frac{s_i+1}{n+2}\right)^1 \left(1-\frac{s_i+1}{n+2}\right)^{1-1}$$

Do đó $P(\mathbf{x}|D) = \prod_{i=1}^{d} \left(\frac{s_i+1}{n+2}\right)^{x_i} \left(1-\frac{s_i+1}{n+2}\right)^{1-x_i}$.

4. Thay thế ước lượng $\hat{\boldsymbol{\theta}}$ cho $\boldsymbol{\theta}$ trong $P(x|\boldsymbol{\theta})$ để được $P(x|D)$ thì từ kết quả của câu (3) và công thức $P(x|\boldsymbol{\theta})$ ta suy ra được ước lượng Bayes của $\boldsymbol{\theta}$ là

$$\hat{\boldsymbol{\theta}} = \frac{s+1}{n+2}$$

ĐỘ LỆCH

Bài 12

Có người đề xuất một phương pháp ước lượng trung bình của tập dữ liệu $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ như sau: ta chỉ cần chỉ định giá trị trung bình cho điểm thứ nhất, tức là $\mathbf{x}_1$.

1. Chứng minh rằng ước lượng này là không lệch.
2. Tại sao phương pháp này lại không thích hợp.

Giải:

1. Gọi trung bình thực sự là $\boldsymbol{\mu}$. Ước lượng được đề xuất không lệch là do

$$E\left[(\mathbf{x}_1 - \boldsymbol{\mu})\right] = E[\mathbf{x}_1] - \boldsymbol{\mu} = 0$$

2. Gọi ma trận hiệp phương sai là $\boldsymbol{\Sigma}$. Người ta không dùng phương pháp này là do ước lượng này có phương sai bằng với $\boldsymbol{\Sigma}$

$$E\left[(\mathbf{x}_1 - \boldsymbol{\mu})(\mathbf{x}_1 - \boldsymbol{\mu})^T\right] = \boldsymbol{\Sigma}$$

Trong thực tế thì ước lượng có phương sai như vậy là khó chấp nhận được.