PATTERN RECOGNITION

Le Dinh Long

Faculty of Information Technology
Industrial University of Ho Chi Minh City, Viet Nam

Ngày 26 tháng 2 năm 2025

CHƯƠNG 3. ƯỚC LƯỢNG DẠNG THAM SỐ

3.1 Giới thiệu

Ở chương hai chúng ta đã tìm hiểu cách xây dựng bộ phân lớp tối ưu dựa trên mô hình thống kê nếu xác suất tiên nghiệm $P(\omega_i)$ và mật độ xác suất có điều kiện $p(\mathbf{x}|\omega_i)$ cho mỗi lớp được xác định. Tuy nhiên, trong phần lớn các ứng dụng nhận dạng mẫu, các đại lượng này không có sẵn. Trong một số trường hợp chúng ta có thông tin rất mơ hồ về nó cùng với tập dữ liệu mẫu hay tập huấn luyện.

Một trong các tiếp cận phổ biến là chúng ta ước lượng các đại lượng $P(\omega_i)$ và $p(\mathbf{x}|\omega_i)$ từ tập dữ liệu huấn luyện được thu thập, sau đó nó được sử dụng để xây dựng bộ nhận dạng. Trong hai đại lượng trên thì xác suất tiên nghiệm $P(\omega_i)$ có thể được ước lượng dễ dàng đối với các bài toán nhận dạng dựa trên mô hình có giám sát. Riêng mật độ xác suất có điều kiện $p(\mathbf{x}|\omega_i)$ tương đối phức tạp hơn, đặc biệt trong trường hợp số mẫu dữ liệu nhỏ và số chiều lớn. Nếu dạng của phân phối được biết (ví dụ như phân phối Gauss) khi đó hàm mật độ xác suất được xác định khi các thông số của phân phối (ví dụ như kỳ vọng $\boldsymbol{\mu}_i$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}_i$) được xác định. Bài toán ước lượng hàm mật độ xác suất đơn giản trở thành bài toán ước lượng các tham số của nó (ví dụ bài toán ước lượng kỳ vọng $\boldsymbol{\mu}_i$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}_i$).

Đối với bài toán ước lượng tham số, có hai phương pháp phổ biến trong thống kê là ước lượng hợp lý cực đại hay viết tắt là MLE (maximum likelihood estimation) và ước lượng Bayes (BE hay Bayesian estimation). Mặc dù kết quả của hai phương pháp ước lượng tham số này thường gần giống nhau trong nhiều ứng dụng, nhưng cách tiếp cận của nó khá khác nhau. Phương pháp MLE giả thiết rằng các tham số mặc dù không biết nhưng nó được cố định ở đâu đó, và mục tiêu của ước lượng là tìm các tham số này sao cho cực đại xác suất đạt được mẫu thu thập. Trong khi đó tiếp cận BE xem các tham số này là các biến ngẫu nhiên với phân phối cho trước. Cùng với các mẫu thu thập sẽ tạo thành mật độ hậu nghiệm để dần hiệu chỉnh độ chính xác về các tham số cần được ước lượng. Chúng ta sẽ thấy rằng, tiếp cận này sẽ làm hàm mật độ xác suất hậu nghiệm của tham số được nhọn hơn (có độ tập trung cao hơn) xung quanh đại lượng chính xác khi số mẫu tăng. Chi tiết về hai phương pháp ước lượng này được trình bài chi tiết ở các phần sau.

3.2 Ước lượng hợp lý cực đại (Maximum likelihood estimation, MLE)

3.2.1 Cơ bản về MLE

MLE là một trong các phương pháp ước lượng được quan tâm rất nhiều do khả năng hội tụ của nó khi số mẫu huấn luyện tăng và đơn giản hơn một số phương pháp ước lượng khác như BE.

MLE yêu cầu đầu tiên có tập dữ liệu huấn luyện. Tập mẫu này có thể được tách thành các tập con tương ứng cho từng lớp. Giả sử có C lớp thì khi đó sẽ có C tập con được ký hiệu là $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_C$ với các mẫu trong mỗi tập $\boldsymbol{D}_i$ được tạo ra một cách độc lập từ mật độ phân phối xác suất $p(\mathbf{x}|\omega_i)$; các mẫu này được xem là các biến ngẫu nhiên phân phối đồng nhất độc lập (independent identically distributed – i.i.d.). Giả sử rằng $p(\mathbf{x}|\omega_i)$ có dạng tham số hóa, có nghĩa là dạng của phân phối này được biết và được xác định chính xác bởi một vector tham số $\boldsymbol{\theta}_i$. Ví dụ nếu như $p(\mathbf{x}|\omega_i)$ có dạng phân phối Gauss, $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, thì tham số $\boldsymbol{\theta}_i$ bao gồm hai thành phần $\boldsymbol{\mu}_i$ và $\boldsymbol{\Sigma}_i$. Chúng ta có thể sử dụng $p(\mathbf{x}|\omega_i; \boldsymbol{\theta}_i)$ để thể hiện sự phụ thuộc của $p(\mathbf{x}|\omega_i)$ với $\boldsymbol{\theta}_i$. Chú ý $\boldsymbol{\theta}_i$ trong trường hợp này không phải là biến ngẫu nhiên. Bài toán ước lượng mật độ xác suất trở thành ước lượng các tham số $\boldsymbol{\theta}_i$ tương ứng cho từng lớp từ tập dữ liệu huấn luyện.

Để đơn giản, chúng ta giả sử rằng việc ước lượng các tham số $\boldsymbol{\theta}_i$ cho mỗi lớp ω_i với tập dữ liệu $\boldsymbol{D}_i$ là độc lập nhau; cách thực hiện giống như nhau. Do đó việc ước lượng các tham số này có thể phát biểu thành dạng chung là "ước lượng tham số $\boldsymbol{\theta}$ của hàm mật độ phân phối $p(\mathbf{x}; \boldsymbol{\theta})$ từ tập dữ liệu huấn luyện $\boldsymbol{D}$".

Giả sử tập dữ liệu huấn luyện $\boldsymbol{D}$ có n mẫu được thu nhận từ phân phối đồng nhất độc lập (i.i.d), $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Khi đó ta có

$$\begin{aligned} p(\boldsymbol{D}; \boldsymbol{\theta}) &= p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n; \boldsymbol{\theta}) \\ &= \prod_{j=1}^{n} p(\mathbf{x}_j; \boldsymbol{\theta}) \end{aligned} \tag{3.1}$$

Xác suất này được xem là likelihood của $\boldsymbol{\theta}$ đối với tập dữ liệu $\boldsymbol{D}$ và có thể ký hiệu bằng hàm $L(\cdot)$, ví dụ $L(\boldsymbol{D}; \boldsymbol{\theta})$, gọi là hàm likelihood.

Mục tiêu của MLE là tìm tập tham số $\boldsymbol{\theta}$ sao cho cực đại hàm likelihood này, nghĩa là

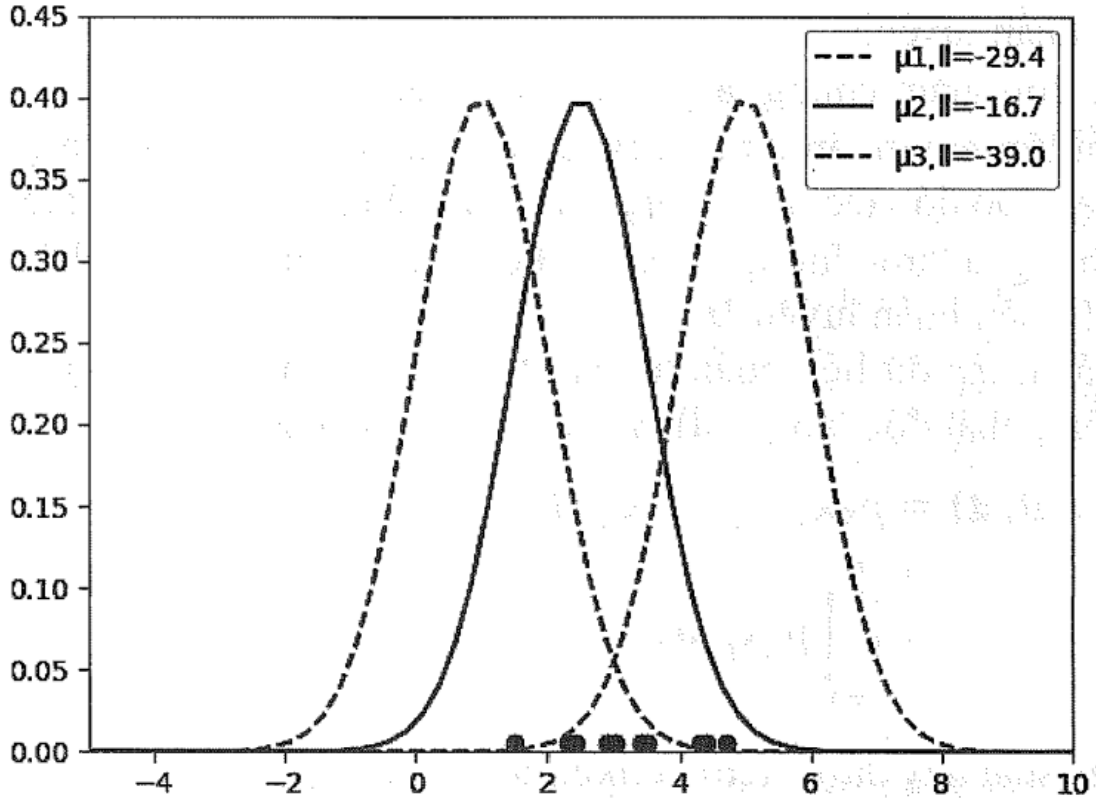
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{D}; \boldsymbol{\theta}), \tag{3.2}$$

trong đó $L(\boldsymbol{D}; \boldsymbol{\theta})$ là tích của các hàm $p(\mathbf{x}_j; \boldsymbol{\theta})$. Tuy nhiên tích các hàm nhỏ có thể dẫn đến tình trạng không ổn định số học (numerical instability). Một trong cách khắc phục tình trạng này là phát biểu lại bài toán. Chúng ta thấy rằng hàm logarit là hàm đơn điệu tăng, nghĩa là nếu $\boldsymbol{\theta}$ làm cho likelihood đạt cực đại thì cũng sẽ làm cho log-likelihood đạt cực đại, do đó sử dụng hàm này để phát biểu lại bài toán là một trong những tiếp cận thường được sử dụng. Chúng ta định nghĩa hàm log-likelihood bởi

$$\begin{aligned} l(\boldsymbol{\theta}) &\triangleq \ln L(\mathbf{D}; \boldsymbol{\theta}) \\ &= \sum_{j=1}^{n} \ln p(\mathbf{x}_j; \boldsymbol{\theta}). \end{aligned} \tag{3.3}$$

Khi đó bài toán trở thành tìm tham số $\boldsymbol{\theta}$ sao cho cực đại hàm log-likelihood này, nghĩa là

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{3.4}$$



Hình 3.1 Minh họa cho log-likehood (ll) với tập dữ liệu được cho như trên. Chúng ta thấy rằng với phân phối có μ_2=2.5 cho giá trị log-likelihood (-16.7) lớn hơn hai trường hợp với phân phối có μ_1=5(-29.4) và μ_3=1(-39.0).

Tổng quát ta phải tìm lời giải toàn cục cho (3.4). Việc tìm lời giải này có thể được thực hiện theo cách tiếp cận truyền thống dựa trên phép tính vi phân. Giả sử thông số $\boldsymbol{\theta}$ là vectơ gồm có K thành phần, $\boldsymbol{\theta}=(\theta_1, \theta_2, \ldots, \theta_K)^T$ và phép tính gradient được định nghĩa bởi

$$\nabla_{\boldsymbol{\theta}} \triangleq \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial}{\partial \theta_K} \end{bmatrix}. \tag{3.5}$$

Khi đó $\hat{\boldsymbol{\theta}}$ là nghiệm của phương trình

$$\nabla_{\boldsymbol{\theta}} l = 0, \tag{3.6}$$

trong đó $\nabla_{\boldsymbol{\theta}} l$ được tính bởi

$$\nabla_{\boldsymbol{\theta}} l = \sum_{j=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_j; \boldsymbol{\theta}). \tag{3.7}$$

Tóm lại để ước lượng hàm mật độ phân phối dựa trên MLE, chúng ta cần thực hiện bốn bước chính sau:

- Thu thập dữ liệu.
- Chọn mô hình hoặc dạng phân phối, với các tham số có thể được cập nhật.
- Xây dựng hàm mục tiêu dựa trên likelihood, ví dụ như log-likelihood.
- Tối ưu hàm mục tiêu để tìm ra các tham số của mô hình.

3.2.2 Trường hợp phân phối chuẩn

Phần này trình bày phương pháp MLE được áp dụng cho trường hợp phân phối cụ thể, đó là phân phối Gauss. Phân phối này có hai tham số là số bình quân (mean) $\boldsymbol{\mu}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}$. Hàm mật độ phân phối có dạng

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad (3.8)$$

trong đó d là số chiều của dữ liệu. Lấy logarit của $(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ ta được

$$\ln p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \quad (3.9)$$

a) Trường hợp không biết $\boldsymbol{\mu}$

Để đơn giản, chúng ta xét trường hợp đầu tiên, chỉ có $\boldsymbol{\mu}$ là không biết và cần được xác định từ tập dữ liệu $\boldsymbol{D}$. Từ (3.7) và (3.9) ta được

$$\nabla_{\boldsymbol{\mu}} l = \sum_{j=1}^{n} \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_j; \boldsymbol{\theta}) = \sum_{j=1}^{n} [\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})]. \qquad (3.10)$$

Thay vào (3.6) ta được

$$\sum_{j=1}^{n} [\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \hat{\boldsymbol{\mu}}] = 0. \qquad (3.11)$$

Nhân $\boldsymbol{\Sigma}$ vào hai vế và thực hiện một số thao tác, ta được kết quả ước lượng của $\boldsymbol{\mu}$ là

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j. \qquad (3.12)$$

Như vậy, MLE của số bình quân của quần thể (population mean) bằng số bình quân mẫu. Nó chính là trung bình cộng của các mẫu huấn luyện.

b) Trường hợp không biết μ và Σ

Trong trường hợp tổng quát hơn, cả mean và ma trận hiệp phương sai đều không biết. Trước tiên chúng ta xét trường hợp một chiều (d=1), khi đó (3.8) và (3.9) trở thành

$$p(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\theta_2}}\exp\left(-\frac{1}{2\theta_2}(x-\theta_1)^2\right) \qquad (3.13)$$

và

$$\ln p(x;\boldsymbol{\theta}) = -\frac{1}{2}\ln[2\pi\theta_2] - \frac{1}{2\theta_2}(x-\theta_1)^2, \qquad (3.14)$$

trong đó $\boldsymbol{\theta}=[\theta_1, \theta_2]^T$, $\theta_1=\mu$, và $\theta_2=\sigma^2$. Đạo hàm của nó là

$$\nabla_{\boldsymbol{\theta}} ln p(x;\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2} \end{bmatrix}. \qquad (3.15)$$

Thay (3.15) vào (3.6) và (3.7) ta có

$$\sum_{j=1}^{n} \begin{bmatrix} \frac{1}{\theta_2}(x_j-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_j-\theta_1)^2}{2\theta_2^2} \end{bmatrix} = 0. \qquad (3.16)$$

Như vậy MLE của θ_1 và θ_2 thỏa các phương trình

$$n\hat{\theta}_1 - \sum_{j=1}^{n} x_j = 0 \qquad (3.17)$$

và

$$n\hat{\theta}_2 - \sum_{j=1}^{n} (x_j - \hat{\theta}_1)^2 = 0 \qquad (3.18)$$

Thay $\hat{\theta}_1 = \hat{\mu}$ và $\hat{\theta}_2 = \hat{\sigma}^2$ chúng ta được MLE cho μ và σ^2 lần lượt là

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad (3.19)$$

và

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^2. \qquad (3.20)$$

Trong trường hợp dữ liệu nhiều chiều (d>1), quá trình cũng được thực hiện tương tự và chúng ta cũng đạt được MLE $\hat{\boldsymbol{\mu}}$ và $\hat{\boldsymbol{\Sigma}}$ lần lượt cho $\boldsymbol{\mu}$ và $\boldsymbol{\Sigma}$ là

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j \qquad (3.21)$$

và

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \hat{\boldsymbol{\mu}})(\mathbf{x}_j - \hat{\boldsymbol{\mu}})^T. \qquad (3.22)$$

Như vậy, trong trường hợp này ước lượng của vectơ số bình quân (mean) cũng là số bình quân của tập mẫu (sample mean) và ước lượng của ma trận hiệp phương sai cũng là trung bình cộng của n ma trận $(\mathbf{x}_j - \hat{\boldsymbol{\mu}})(\mathbf{x}_j - \hat{\boldsymbol{\mu}})^T$.

3.3 Ước lượng không chệch (unbiased estimation)

Một trong các tiêu chí đánh giá bộ ước lượng là tính không chệch (unbias) của nó. Một bộ ước lượng được gọi là không chệch (unbiased estimator) nếu kỳ vọng của nó bằng giá trị đúng của đại lượng cần ước lượng, nghĩa là nếu $\hat{\boldsymbol{\theta}}$ là ước lượng của $\boldsymbol{\theta}$ thì

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}. \tag{3.23}$$

Ước lượng MLE $\hat{\boldsymbol{\mu}}$ và $\hat{\boldsymbol{\Sigma}}$ có những đặc điểm về độ không chênh lệch như sau:

a) Ước lượng MLE $\hat{\boldsymbol{\mu}}$ của $\boldsymbol{\mu}$ theo (3.12) và (3.21) là không chệch.

b) Ước lượng MLE $\hat{\boldsymbol{\Sigma}}$ của $\boldsymbol{\Sigma}$ là chệnh và ước lượng không chệch của $\boldsymbol{\Sigma}$ là

$$\hat{\boldsymbol{\Sigma}}_C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_j - \hat{\boldsymbol{\mu}})(\mathbf{x}_j - \hat{\boldsymbol{\mu}})^T. \tag{3.24}$$

3.4 Phương pháp ước lượng Bayes

Phần này trình bày phương pháp ước lượng phổ biến thứ hai, đó là ước lượng Bayes (BE). Mặc dù kết quả của phương pháp này gần giống như kết quả của MLE. Tuy nhiên tiếp cận của hai phương pháp này khác nhau. Phương pháp MLE xem các đại lượng cần ước lượng là cố định, trong khi phương pháp này xem các đại lượng cần ước lượng là các biến ngẫu nhiên theo một phân phối nào đó. Phương pháp BE sẽ chuyển từ phân phối đó thành phân phối xác suất hậu nghiệm.

Từ công thức Bayes, để tính xác suất hậu nghiệm tương ứng mỗi lớp, ta cần tính các đại lượng xác suất tiên nghiệm $P(\omega_i)$ và mật độ xác suất có điều kiện $p(\mathbf{x}|\omega_i)$. Các đại lượng này được xác định dựa vào thông tin từ tập mẫu $\boldsymbol{D}$. Do đó, để nhấn mạnh vai trò của tập mẫu, công thức Bayes có thể được viết thành

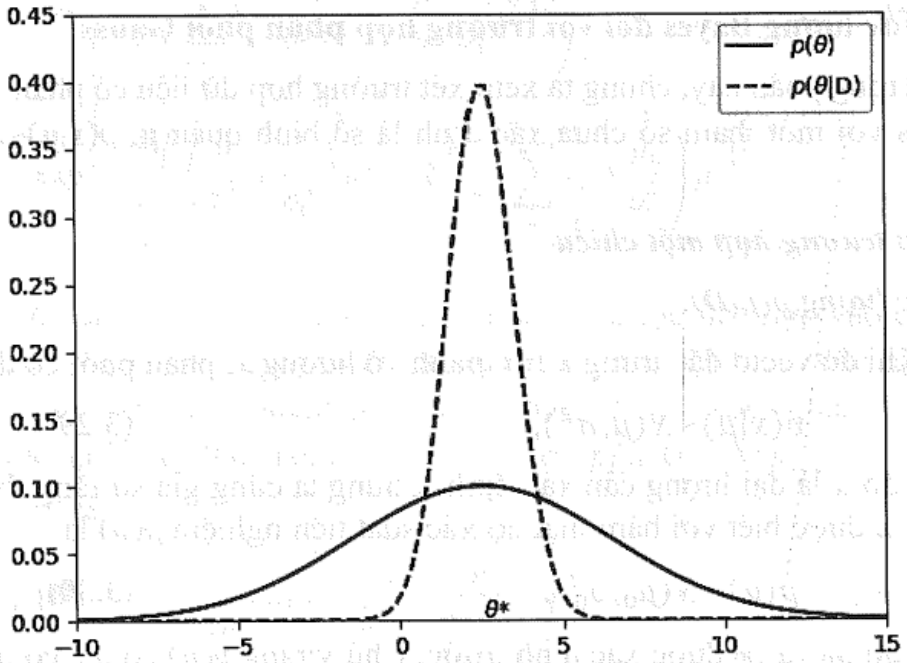
$$P(\omega_i|\mathbf{x}, \boldsymbol{D}) = \frac{p(\mathbf{x}|\omega_i, \boldsymbol{D})P(\omega_i|\boldsymbol{D})}{\sum_{j=1}^{C} p(\mathbf{x}|\omega_j, \boldsymbol{D})P(\omega_j|\boldsymbol{D})}. \quad (3.25)$$

Để đơn giản, chúng ta có thể xem đại lượng $P(\omega_i|\mathbf{D})$ được xác định hoặc tính dễ dàng từ tập mẫu. Nó có thể được thay bằng $P(\omega_i)$ và không quan tâm nhiều ở đây. Hơn nữa, tương tự như phương pháp MLE, tập dữ liệu dùng cho huấn luyện có thể được phân thành C tập con tương ứng với C lớp. Việc xác định hàm mật độ phân phối cho mỗi lớp không chịu sự ảnh hưởng từ tập dữ liệu của lớp khác, nghĩa là tập con dữ liệu $\boldsymbol{D}_j$ sẽ không ảnh hưởng đến việc xác định $p(\mathbf{x}|\omega_i; \boldsymbol{D}_i)$ nếu $i \neq j$. Do đó công thức (3.25) trở thành

$$P(\omega_i|\mathbf{x}, \boldsymbol{D}) = \frac{p(\mathbf{x}|\omega_i, \boldsymbol{D}_i)P(\omega_i)}{\sum_{j=1}^{C} p(\mathbf{x}|\omega_j, \boldsymbol{D}_j)P(\omega_j)}. \qquad (3.26)$$

Việc ước lượng tham số của các phân phối cho các lớp là tương tự nhau và có thể được thực hiện độc lập nhau. Bài toán trở thành "sử dụng tập dữ liệu huấn luyện $\boldsymbol{D}$ từ một phân phối chưa biết (nhưng cố định) $p(\mathbf{x})$ để ước lượng mật độ phân phối $p(\mathbf{x}|\boldsymbol{D})$".

Mặc dù $p(\mathbf{x})$ là không biết, nhưng dạng của nó thể hiện qua $p(\mathbf{x}|\boldsymbol{\theta})$ được xác định và được tham số hóa thông qua $\boldsymbol{\theta}$. Bài toán tìm mật độ phân phối trở thành bài toán tìm tham số $\boldsymbol{\theta}$. Khác với phương pháp MLE, ở đó tham số của phân phối $\boldsymbol{\theta}$ là cố định, tham số của phân phối trong trường hợp này được xem là biến ngẫu nhiên và thông tin về nó là xác suất tiên nghiệm với hàm mật độ $p(\boldsymbol{\theta})$ được xác định trước. Từ tập dữ liệu huấn luyện thu thập được $\boldsymbol{D}$ chúng ta xác định xác suất hậu nghiệm $p(\boldsymbol{\theta}|\boldsymbol{D})$ được kỳ vọng có đỉnh xấp xỉ và nhọn hơn (sharp) tại giá trị chính xác của $\boldsymbol{\theta}$ như trong Hình 3.2.



Hình 3.2 Phân phối xác suất tiên nghiệm của tham số $p(\boldsymbol{\theta})$. Khi có tập dữ liệu huấn luyện, nó sẽ làm cho hàm mật độ xác suất được nhọn hơn (với phương sai nhỏ hơn) tại giá trị chính xác của tham số, $\boldsymbol{\theta}^*$.

Mục đích cuối cùng của chúng ta trong phương pháp này là xác định $p(\mathbf{x}|\boldsymbol{D})$ để xấp xỉ về phân phối đúng $p(\mathbf{x})$. $p(\mathbf{x}|\boldsymbol{D})$ có thể được tính dựa trên xác suất kết hợp (joint probability) $p(\mathbf{x},\boldsymbol{\theta}|\boldsymbol{D})$ như sau

$$p(\mathbf{x}|\boldsymbol{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{D})\, d\boldsymbol{\theta}. \qquad (3.27)$$

Trong đó tích phân được thực hiện trên toàn bộ không gian thông số $\boldsymbol{\theta}$. Chúng ta có $p(\mathbf{x},\boldsymbol{\theta}|\boldsymbol{D}) = p(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{D})p(\boldsymbol{\theta}|\boldsymbol{D})$, hơn nữa $\mathbf{x}$ và $\boldsymbol{D}$ được chọn một cách độc lập nhau, do đó $p(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{D}) = p(\mathbf{x}|\boldsymbol{\theta})$. Như vậy (3.27) sẽ trở thành

$$p(\mathbf{x}|\boldsymbol{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\boldsymbol{D}) d\boldsymbol{\theta}. \qquad (3.28)$$

Nếu $p(\boldsymbol{\theta}|\boldsymbol{D})$ có phương sai rất nhỏ và số bình quân (mean) là $\widehat{\boldsymbol{\theta}}$ thì $p(\mathbf{x}|\boldsymbol{D})$ từ (3.28) có thể được xấp xỉ là $p(\mathbf{x}|\boldsymbol{D}) \approx p(\mathbf{x}|\widehat{\boldsymbol{\theta}})$. Trong trường hợp tổng quát, giá trị của $\boldsymbol{\theta}$ không chắc chắn, chúng ta có thể ước lượng $p(\mathbf{x}|\boldsymbol{D})$ dựa trên trung bình $p(\mathbf{x}|\boldsymbol{\theta})$ trên tất cả các giá trị có thể của $\boldsymbol{\theta}$. Tích phân (3.28) cũng có thể được tính thông qua một số mô hình, ví dụ như Monte-Carlo.

3.5 Ước lượng Bayes đối với trường hợp phân phối Gauss

Trong phần này, chúng ta xem xét trường hợp dữ liệu có phân phối Gauss với một tham số chưa xác định là số bình quân $\boldsymbol{\mu}$, $p(\mathbf{x}|\boldsymbol{\mu})\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

a) Xét trường hợp một chiều

** Ước lượng $p(\mu|\boldsymbol{D})$:*

Khi đó vectơ đặc trưng $\mathbf{x}$ trở thành vô hướng x, phân phối có dạng

$$p(x|\mu)\sim N(\mu,\sigma^2), \tag{3.29}$$

trong đó μ là đại lượng cần xác định. Chúng ta cũng giả sử rằng thông tin về μ được biết với hàm mật độ xác suất tiên nghiệm $p(\mu)$ là

$$p(\mu)\sim N(\mu_0,\sigma_0^2), \tag{3.30}$$

trong đó μ_0 và σ_0 được xác định trước. Chú ý rằng $p(\mu)$ có thể có dạng hàm mật độ phân phối khác thay vì Gauss. Với dạng Gauss trên, nó thể hiện giá trị dự đoán tốt nhất của μ là μ_0 và độ lệch chuẩn σ_0 thể hiện sự không chắc chắn của dự đoán trên.

Với giá trị μ được chọn từ hàm mật độ phân phối $p(\mu)$, tập mẫu $\boldsymbol{D}=\{x_1, x_2,\ldots,x_n\}$ được chọn ngẫu nhiên từ hàm mật độ phân phối $p(x|\mu)$. Theo công thức Bayes ta có

$$\begin{aligned} p(\mu|\boldsymbol{D}) &= \frac{p(\boldsymbol{D}|\mu)p(\mu)}{\int p(\boldsymbol{D}|\mu)p(\mu)d\mu} \\ &= a\prod_{j=1}^{n} p(x_j|\mu)p(\mu), \end{aligned} \tag{3.31}$$

trong đó a là hệ số phụ thuộc tập mẫu $\boldsymbol{D}$ nhưng độc lập với μ, nó đóng vai trò chuẩn hóa. Thay $p(x|\mu)\sim N(\mu,\sigma^2)$ và $p(\mu)\sim N(\mu_0,\sigma_0^2)$ vào (3.31) ta được

$$
\begin{aligned}
&p(\mu|\boldsymbol{D}) \\
&= a\prod_{j=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma}\right)^2\right]\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \\
&= a_1\exp\left[-\frac{1}{2}\left(\sum_{j=1}^{n}\left(\frac{x_j-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \qquad (3.32)\\
&= a_2\exp\left[-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{j=1}^{n}x_j+\frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right],
\end{aligned}
$$

trong đó các thành phần không phụ thuộc μ được tích hợp vào các hệ số a, a_1, và a_2. Rõ ràng $p(\mu|\boldsymbol{D})$ có dạng hàm mũ của biểu thức bậc hai theo μ, có nghĩa nó cũng là phân phối chuẩn. Nếu ta biểu diễn hàm mật độ phân phối này theo hai tham số μ_n và σ_n, $p(\mu|\boldsymbol{D})\sim N(\mu_n,\sigma_n^2)$, khi đó nó có dạng

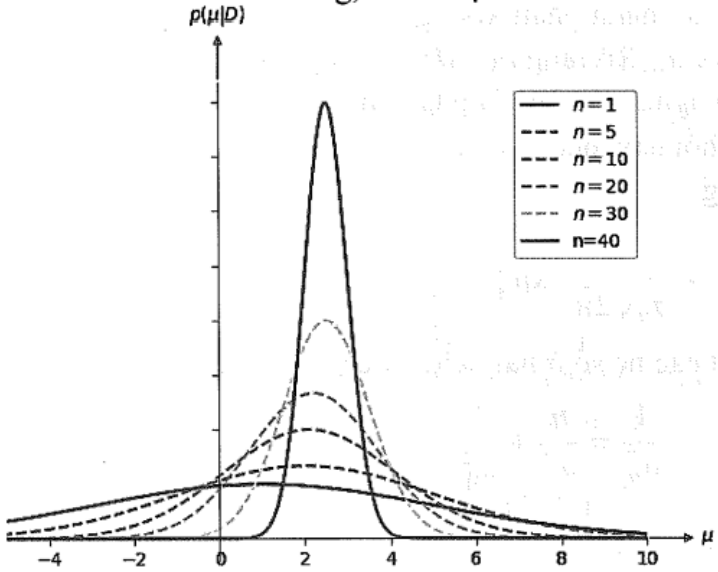
Giải hệ hai phương trình (3.34) và (3.35) ta được

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 \qquad (3.37)$$

và

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}. \qquad (3.38)$$

Hai phương trình trên cung cấp mối quan hệ giữa xác suất hậu nghiệm của thông số $p(\mu|\boldsymbol{D})$ với thông tin xác suất tiên nghiệm $p(\mu)$ và tập dữ liệu mẫu $\boldsymbol{D}$. Cụ thể nếu không có tập mẫu $\boldsymbol{D}$, n=0, thì dự đoán tốt nhất của μ sẽ hướng về μ_0, với tập dữ liệu $\boldsymbol{D}$ có n mẫu thì ước lượng của μ là μ_n với độ không chắc chắn được xác định dựa vào σ_n^2. Từ (3.38) dễ dàng thấy rằng σ_n^2 đơn điệu giảm khi n tăng và tiệm cận σ^2/n khi n tiến về vô cùng. Khi n càng tăng hàm $p(\mu|\boldsymbol{D})$ sẽ càng nhọn hơn và tiệm cận về hàm Dirac khi n tiến về vô cùng, minh họa như Hình 3.3.



từ xác suất tiên nghiệm là rất cao, nó không làm thay đổi quan điểm giá trị của μ_n cho dù chúng ta có thêm tập mẫu. Ngược lại, nếu σ_0 rất lớn so với σ ($\sigma_0 >> \sigma$), có nghĩa độ chắc chắc của ước lượng $\mu_n = \mu_0$ rất thấp, khi đó ước lượng của μ_n hướng về tập dữ liệu thu thập, nghĩa là $\mu_n \approx \bar{x}_n$. Trong các trường hợp còn lại, khi số mẫu đủ lớn thì đóng góp của μ_0 và σ_0 sẽ giảm dần và μ_n sẽ hội tụ về số bình quân mẫu, $\bar{x}_n$.

* *Ước lượng* $p(x|\boldsymbol{D})$:

Sau khi ước lượng $p(\mu|\boldsymbol{D})$, công việc còn lại là ước lượng hàm mật độ xác suất có điều kiện $p(x|\boldsymbol{D})$. Từ (3.28) ta có

$$p(x|\boldsymbol{D}) = \int p(x|\mu)\, p(\mu|\boldsymbol{D}) d\mu. \tag{3.39}$$

Thay (3.29) và (3.33) vào (3.39) ta được

$$\begin{aligned} & p(x|\boldsymbol{D}) \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] A, \end{aligned} \tag{3.40}$$

trong đó hệ số A độc lập với x và được định nghĩa bởi

$$A = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu. \qquad (3.41)$$

Từ (3.40), chúng ta có thể thấy rằng $p(x|\boldsymbol{D})$ là một phân phối chuẩn với mean là μ_n và phương sai là $\sigma^2+\sigma_n^2$; $p(x|\boldsymbol{D})\sim N(\mu_n, \sigma^2+\sigma_n^2)$. Hơn nữa, chúng ta cũng dễ dàng thấy rằng x có thể được xem là tổng của hai biến ngẫu nhiên với mật độ phân phối $p(\mu|\mathbf{D})\sim N(\mu_n, \sigma_n^2)$ và $p(w)\sim N(0, \sigma^2)$, do đó hàm mật độ phân phối của nó có dạng $p(x|\boldsymbol{D})\sim N(\mu_n, \sigma^2+\sigma_n^2)$.

Sau khi xác định $p(x|\boldsymbol{D})$ cho từng lớp ta được $p(x|\omega_i, \boldsymbol{D})$ cùng với xác suất tiên nghiệm $P(\omega_i)$ sẽ có đủ thông tin để xây dựng bộ phân lớp.

b) Xét trường hợp nhiều chiều

Phần này tổng quát hóa trường hợp một chiều ở trên, nghĩa là ma trận hiệp phương sai $\boldsymbol{\Sigma}$ được biết nhưng vectơ bình quân (mean) $\boldsymbol{\mu}$ không biết. Như phần trước, các phân phối dạng Gauss như sau

$$p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ và } p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \tag{3.42}$$

trong đó $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, và $\boldsymbol{\Sigma}$ giả sử là đã được xác định. Với tập mẫu $\boldsymbol{D}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, tương tự như (3.31) ta có

$$p(\boldsymbol{\mu}|\boldsymbol{D}) = a \prod_{j=1}^{n} p(\mathbf{x}_j|\boldsymbol{\mu}) p(\boldsymbol{\mu})$$

$$= a_1 \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^T(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{j=1}^{n}\mathbf{x}_j + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right)\right], \tag{3.43}$$

trong đó các thành phần độc lập $\boldsymbol{\mu}$ được tích hợp trong hệ số a_1. Rõ ràng (3.43) có dạng hàm mũ của biểu thức bậc hai

$$p(\boldsymbol{\mu}|\mathbf{D}) = a_2 \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right]. \tag{3.44}$$

Như vậy $p(\boldsymbol{\mu}|\boldsymbol{D})$ có dạng phân phối Gauss dạng $p(\boldsymbol{\mu}|\boldsymbol{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Đồng nhất (3.43) và (3.44) ta được

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \tag{3.45}$$

và

$$\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0, \tag{3.46}$$

trong đó $\bar{\mathbf{x}}_n$ là bình quân mẫu (sample mean)

$$\bar{\mathbf{x}}_n = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j. \tag{3.47}$$

Áp dụng tính chất

$$(\mathbf{X}^{-1} + \mathbf{Y}^{-1})^{-1} = \mathbf{X}(\mathbf{X}+\mathbf{Y})^{-1}\mathbf{Y} = \mathbf{Y}(\mathbf{X}+\mathbf{Y})^{-1}\mathbf{X} \tag{3.48}$$

vào (3.45) ta được

$$\boldsymbol{\Sigma}_n = \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0\right)^{-1}\boldsymbol{\Sigma}_0 \tag{3.49}$$

và

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0\right)^{-1}\bar{\mathbf{x}}_n + \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0\right)^{-1}\boldsymbol{\mu}_0. \tag{3.50}$$

Nhận thấy rằng $\mathbf{x}$ được xem như là tổng của hai biến ngẫu nhiên với mật độ phân phối $p(\boldsymbol{\mu}|\boldsymbol{D})\sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ và $p(\mathbf{w})\sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Do đó hàm mật độ phân phối $p(\mathbf{x}|\boldsymbol{D})$ sẽ là

$$p(\mathbf{x}|\boldsymbol{D})\sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}). \qquad (3.51)$$

Ngoài ra, chúng ta cũng có thể xác định $p(\mathbf{x}|\boldsymbol{D})$ dựa vào $p(\mathbf{x}|\boldsymbol{D}) = \int p(\mathbf{x}|\boldsymbol{\mu})\, p(\boldsymbol{\mu}|\boldsymbol{D}) d\boldsymbol{\mu}$.

3.6 Ước lượng tham số Bayes: Tổng quát

Phần trên giới thiệu phương pháp ước lượng Bayes trong trường hợp phân phối Gauss. Tiếp cận này có thể được tổng quát cho các trường hợp phân phối khác đã được biết dạng của nó. Quá trình được tóm lược như sau:

- Dạng của hàm mật độ xác suất có điều kiện (hay likelihood) $p(\mathbf{x}|\boldsymbol{\theta})$ được biết, nhưng vectơ tham số của nó, $\boldsymbol{\theta}$, chưa biết và ta cần phải xác định.
- Thông tin về tham số $\boldsymbol{\theta}$ chỉ được biết qua hàm mật độ xác suất tiên nghiệm $p(\boldsymbol{\theta})$.
- Thông tin còn lại liên quan đến $\boldsymbol{\theta}$ là tập dữ liệu $\boldsymbol{D}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ bao gồm các mẫu được rút trích độc lập từ hàm mật độ phân phối chưa biết $p(\mathbf{x})$.

Bài toán trở thành tính mật độ xác suất hậu nghiệm $p(\boldsymbol{\theta}|\boldsymbol{D})$, từ đó chúng ta tính $p(\mathbf{x}|\boldsymbol{D})$ theo công thức (3.28). Sử dụng công thức Bayes ta có

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{D})}, \tag{3.52}$$

trong đó

$$p(\boldsymbol{D}) = \int p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3.53}$$

đóng vai trò là hệ số chuẩn hóa, nó là tích phân đối với tham số $\boldsymbol{\theta}$. Trong trường hợp phân phối được giả thiết là phân phối Gauss như trên hoặc dữ liệu với số chiều nhỏ thì việc xác định có thể được thực hiện. Tuy nhiên khi phân phối là bất kỳ và/hoặc dữ liệu với số chiều lớn, việc tính tích phân này trở nên phức tạp hơn. Một số tiếp cận để giải quyết bài toán này bao gồm:

- Sử dụng tích phân số.
- Xấp xỉ các hàm được sử dụng để tính hậu nghiệm bởi các hàm đơn giản hơn (variational Bayes).

- Sử dụng các phương pháp Monte Carlo, cụ thể là Markov Chain Monte Carlo (MCMC).