Ôn lại một chút về hai tính chất của hàm logarit: (i) log của một tích bằng tổng của các log, và (ii) vì log là một hàm đồng biến, một biểu thức dương sẽ là lớn nhất nếu log của nó là lớn nhất, và ngược lại.

4.2.3 Ví dụ

Ví dụ 1: phân phối Bernoulli

Bài toán: giả sử tung một đồng xu N lần và nhận được n mặt *head*. Ước lượng xác suất khi tung đồng xu nhận được mặt *head*.

Lời giải:

Một cách trực quan, ta có thể ước lượng được rằng xác suất đó chính là $\lambda = \frac{n}{N}$. Chúng ta cùng ước lượng giá trị này sử dụng MLE.

Giả sử λ là xác suất để nhận được một mặt *head*. Đặt $x_1, x_2, \ldots, x_N$ là các đầu ra nhận được, trong đó có n giá trị bằng 1 tương ứng với mặt *head* và $m = N - n$ giá trị bằng 0 tương ứng với mặt *tail*. Ta có thể suy ra ngay rằng

$$\sum_{i=1}^{N} x_i = n, \quad N - \sum_{i=1}^{N} x_i = N - n = m \tag{4.5}$$

Vì đây là một xác suất của biến ngẫu nhiên nhị phân rời rạc, ta có thể nhận thấy việc nhận được mặt *head* hay *tail* khi tung đồng xu tuân theo phân phối Bernoulli:

$$p(x_i|\lambda) = \lambda^{x_i}(1-\lambda)^{1-x_i} \tag{4.6}$$

Khi đó tham số mô hình λ có thể được ước lượng bằng việc giải bài toán tối ưu sau đây, với giả sử rằng kết quả của các lần tung đồng xu là độc lập với nhau:

$$\lambda = \underset{\lambda}{\text{argmax}} \left[p(x_1, x_2, \ldots, x_N|\lambda)\right] = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} p(x_i|\lambda)\right] \tag{4.7}$$

$$= \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} \lambda^{x_i}(1-\lambda)^{1-x_i}\right] = \underset{\lambda}{\text{argmax}} \left[\lambda^{\sum_{i=1}^{N} x_i}(1-\lambda)^{N-\sum_{i=1}^{N} x_i}\right] \tag{4.8}$$

$$= \underset{\lambda}{\text{argmax}} \left[\lambda^n(1-\lambda)^m\right] = \underset{\lambda}{\text{argmax}} \left[n\log(\lambda) + m\log(1-\lambda)\right] \tag{4.9}$$

trong (4.9), ta đã lấy log của hàm mục tiêu. Tới đây, bài toán tối ưu (4.9) có thể được giải bằng cách lấy đạo hàm của hàm mục tiêu bằng 0. Tức λ là nghiệm của phương trình

$$\frac{n}{\lambda} - \frac{m}{1-\lambda} = 0 \Leftrightarrow \frac{n}{\lambda} = \frac{m}{1-\lambda} \Leftrightarrow \lambda = \frac{n}{n+m} = \frac{n}{N} \tag{4.10}$$

Vậy kết quả ta ước lượng ban đầu là có cơ sở.

Ví dụ 2: Categorical distribution

Một ví dụ khác phức tạp hơn một chút.

Bài toán: giả sử tung một viên xúc xắc sáu mặt có xác suất rơi vào các mặt có thể không đều nhau. Giả sử trong N lần tung, số lượng xuất hiện các mặt thứ nhất, thứ hai,..., thứ sáu lần lượt là $n_1, n_2, \ldots, n_6$ lần với $\sum_{i=1}^{6} n_i = N$. Tính xác suất rơi vào mỗi mặt ở lần tung tiếp theo. Giả sử thêm rằng $n_i > 0,\ \forall i = 1, \ldots, 6$.

Lời giải:

Bài toán này có vẻ phức tạp hơn bài toán trên một chút, nhưng ta cũng có thể dự đoán được ước lượng tốt nhất của xác suất rơi vào mặt thứ i là $\lambda_i = \frac{n_i}{N}$.

Mã hoá mỗi quan sát đầu ra thứ i bởi một vector 6 chiều $\mathbf{x}_i \in \{0, 1\}^6$ trong đó các phần tử của nó bằng 0 trừ phần tử tương ứng với mặt quan sát được là bằng 1. Nhận thấy rằng $\sum_{i=1}^{N} x_i^j = n_j,\ \forall j = 1, 2, \ldots, 6$, trong đó x_i^j là thành phần thứ j của vector $\mathbf{x}_i$.

Có thể thấy rằng xác suất rơi vào mỗi mặt tuân theo phân phối categorical với các tham số $\lambda_j > 0, j = 1, 2, \ldots, 6$. Ta dùng $\boldsymbol{\lambda}$ để thể hiện cho cả sáu tham số này.

Với các tham số $\boldsymbol{\lambda}$, xác suất để sự kiện $\mathbf{x}_i$ xảy ra là

$$p(\mathbf{x}_i|\lambda) = \prod_{j=1}^{6} \lambda_j^{x_i^j} \tag{4.11}$$

Khi đó, vẫn với giả sử về sự độc lập giữa các lần tung xúc xắc, ước lượng bộ tham số λ dựa trên việc tối đa log-likelihood ta có:

$$\boldsymbol{\lambda} = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\lambda) \right] = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} \prod_{j=1}^{6} \lambda_j^{x_i^j} \right] \tag{4.12}$$

$$= \underset{\lambda}{\text{argmax}} \left[\prod_{j=1}^{6} \lambda_j^{\sum_{i=1}^{N} x_i^j} \right] = \underset{\lambda}{\text{argmax}} \left[\prod_{j=1}^{6} \lambda_j^{n_j} \right] \tag{4.13}$$

$$= \underset{\lambda}{\text{argmax}} \left[\sum_{j=1}^{6} n_j \log(\lambda_j) \right] \tag{4.14}$$

Khác với bài toán (4.9) một chút, chúng ta không được quên điều kiện $\sum_{j=1}^{6} \lambda_j = 1$. Ta có bài toán tối ưu có ràng buộc sau đây

$$\max_{\lambda} \sum_{j=1}^{6} n_j \log(\lambda_j) \quad \text{thoả mãn:} \sum_{j=1}^{6} \lambda_j = 1 \tag{4.15}$$

Bài toán tối ưu này có thể được giải bằng phương pháp nhân tử Lagrange (xem Phụ lục A).

Lagrangian của bài toán này là

$$\mathcal{L}(\lambda, \mu) = \sum_{j=1}^{6} n_j \log(\lambda_j) + \mu(1 - \sum_{j=1}^{6} \lambda_j) \tag{4.16}$$

Nghiệm của bài toán là nghiệm của hệ đạo hàm của $\mathcal{L}(.)$ theo từng biến bằng 0

$$\frac{\partial \mathcal{L}(\lambda, \mu)}{\partial \lambda_j} = \frac{n_j}{\lambda_j} - \mu \quad = 0, \ \forall j = 1, 2, \ldots, 6 \tag{4.17}$$

$$\frac{\partial \mathcal{L}(\lambda, \mu)}{\partial \mu} = 1 - \sum_{j=1}^{6} \lambda_j = 0 \tag{4.18}$$

Từ (4.17) ta có $\lambda_j = \frac{n_j}{\mu}$. Thay vào (4.18),

$$\sum_{j=1}^{6} \frac{n_j}{\mu} = 1 \Rightarrow \mu = \sum_{j=1}^{6} n_j = N \tag{4.19}$$

Từ đó ta có ước lượng $\lambda_j = \frac{n_j}{N}, \quad \forall j = 1, 2, \ldots, 6.$

Qua hai ví dụ trên ta thấy MLE cho hết quả khá hợp lý.

Ví dụ 3: Univariate normal distribution

Bài toán: Khi thực hiện một phép đo, giả sử rằng rất khó để có thể đo *chính xác* độ dài của một vật. Thay vào đó, người ta thường đo vật đó nhiều lần rồi suy ra kết quả, với giả thiết rằng các phép đo là độc lập với nhau và kết quả mỗi phép đo là một phân phối chuẩn. Ước lượng chiều dài của vật đó dựa trên các kết quả đo được.

Lời giải:Vì biết rằng kết quả phép đo tuân theo phân phối chuẩn, ta sẽ cố gắng đi xây dựng phân phối chuẩn đó. Chiều dài của vật có thể được coi là giá trị mà hàm mật độ xác suất đạt giá trị cao nhất, tức khả năng rơi vào khoảng giá trị xung quanh nó là lớn nhất. Trong phân phối chuẩn, ta biết rằng hàm mật độ xác suất đạt giá trị lớn nhất tại chính kỳ vọng của phân phối đó. Chú ý rằng kỳ vọng của phân phối và kỳ vọng của dữ liệu quan sát được có thể không chính xác bằng nhau, nhưng rất gần nhau. Nếu ước lượng kỳ vọng của phân phối như cách làm dưới đây sử dụng MLE, ta sẽ thấy rằng kỳ vọng của dữ liệu chính là đánh giá tốt nhất cho kỳ vọng của phân phối.

Thật vậy, giả sử các kích thước quan sát được là $x_1, x_2, \ldots, x_N$. Ta cần đi tìm một phân phối chuẩn, được mô tả bởi một giá trị kỳ vọng μ và phương sai σ^2, sao cho các giá trị $x_1, x_2, \ldots, x_N$ là *likely nhất*. Ta đã biết rằng, hàm mật độ xác suất tại x_i của môt phân phối chuẩn có kỳ vọng μ và phương sai σ^2 là

$$p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{4.20}$$

Vậy, để đánh giá μ và σ, ta sử dụng MLE với giả thiết rằng kết quả các phép đo là độc lập:

$$\mu, \sigma = \underset{\mu,\sigma}{\text{argmax}} \left[\prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \right] \tag{4.21}$$

$$= \underset{\mu,\sigma}{\text{argmax}} \left[\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2} \right) \right] \tag{4.22}$$

$$= \underset{\mu,\sigma}{\text{argmax}} \left[-N\log(\sigma) - \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2} \triangleq J(\mu, \sigma) \right] \tag{4.23}$$

Ta đã lấy log của hàm bên trong dấu ngoặc vuông của (4.22) để được (4.23), phần hằng số có chứa 2π cũng đã được bỏ đi vì nó không ảnh hưởng tới kết quả.

Để tìm μ và σ, ta giải hệ phương trình đạo hàm của $J(\mu, \sigma)$ theo mỗi biến bằng không:

$$\frac{\partial J}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{N}(x_i - \mu) = 0 \tag{4.24}$$

$$\frac{\partial J}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{N}(x_i - \mu)^2 = 0 \tag{4.25}$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{4.26}$$

Kết quả thu được không có gì bất ngờ.

Ví dụ 4: Multivariate normal distribution

Bài toán: Giả sử tập dữ liệu ta thu được là các giá trị nhiều chiều $\mathbf{x}_1, \ldots, \mathbf{x}_N$ tuân theo phân phối chuẩn. Hãy đánh giá các tham số, vector kỳ vọng $\boldsymbol{\mu}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}$ của phân phối này dựa trên MLE, giả sử rằng các $\mathbf{x}_1, \ldots, \mathbf{x}_N$ là độc lập.

Lời giải: Việc chứng minh các công thức

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N} \tag{4.27}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{i=1}^{N}(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \tag{4.28}$$

xin được dành lại cho bạn đọc như một bài tập nhỏ. Dưới đây là một vài gợi ý:

- Hàm mật độ xác suất của phân phối chuẩn nhiều chiều là

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right) \tag{4.29}$$

  Chú ý rằng ma trận hiệp phương sai $\boldsymbol{\Sigma}$ là xác định dương nên có nghịch đảo.

- Một vài đạo hàm theo ma trận:

$$\nabla_{\mathbf{\Sigma}} \log |\mathbf{\Sigma}| = (\mathbf{\Sigma}^{-1})^T \triangleq \mathbf{\Sigma}^{-T} \quad \text{(chuyển vị của nghịch đảo)} \tag{4.30}$$

$$\nabla_{\mathbf{\Sigma}} (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = -\mathbf{\Sigma}^{-T} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-T} \tag{4.31}$$

(Xem thêm Matrix Calculus, mục D.2.1 và D.2.4 tại *https://goo.gl/JKg631*.)

4.3 Maximum a Posteriori

4.3.1 Ý tưởng

Quay lại với ví dụ 1 về tung đồng xu. Nếu tung đồng xu 5000 lần và nhận được 1000 lần *head*, ta có thể đánh giá xác suất của *head* là 1/5 và việc đánh giá này là đáng tin vì số mẫu là lớn. Nếu tung 5 lần và chỉ nhận được 1 mặt *head*, theo MLE, xác suất để có một mặt *head* được đánh giá là 1/5. Tuy nhiên với chỉ 5 kết quả, ước lượng này là không đáng tin, nhiều khả năng việc đánh giá đã bị overfitting. Khi tập huấn luyện quá nhỏ (*low-training*) chúng ta cần phải quan tâm tới một vài giả thiết của các tham số. Trong ví dụ này, giả thiết của chúng ta là xác suất nhận được mặt *head* phải gần 1/2.

Maximum A Posteriori (MAP) ra đời nhằm giải quyết vấn đề này. Trong MAP, chúng ta giới thiệu một giả thiết biết trước, được gọi là *prior*, của tham số θ. Từ giả thiết này, chúng ta có thể suy ra các khoảng giá trị và phân bố của tham số.

Ngược với MLE, trong MAP, chúng ta sẽ đánh giá tham số như là một xác suất có điều kiện của dữ liệu:

$$\theta = \arg\max_{\theta} \underbrace{p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N)}_{\text{posterior}} \tag{4.32}$$

Biểu thức $p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N)$ còn được gọi là *xác suất posterior* của θ. Chính vì vậy mà việc ước lượng θ theo (4.32) được gọi là *Maximum A Posteriori*.

Thông thường, hàm tối ưu trong (4.32) khó xác định dạng một cách trực tiếp. Chúng ta thường biết điều ngược lại, tức nếu biết tham số, ta có thể tính được hàm mật độ xác suất của dữ liệu. Vì vậy, để giải bài toán MAP, ta thường sử dụng quy tắc Bayes. Bài toán MAP thường được biến đổi thành

$$\theta = \arg\max_{\theta} p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N) = \arg\max_{\theta} \left[\frac{\overbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_N|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_N)}_{\text{evidence}}} \right] \tag{4.33}$$

$$= \arg\max_{\theta} \left[p(\mathbf{x}_1, \dots, \mathbf{x}_N|\theta) p(\theta) \right] \tag{4.34}$$

$$= \arg\max_{\theta} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\theta) p(\theta) \right] \tag{4.35}$$

Đẳng thức (4.33) xảy ra theo quy tắc Bayes. Đẳng thức (4.34) xảy ra vì mẫu số của (4.33) không phụ thuộc vào tham số θ. Đẳng thức (4.35) xảy ra nếu chúng ta giả thiết về sự độc lập giữa các $\mathbf{x}_i$. Chú ý rằng giả thiết độc lập thường xuyên được sử dụng.

Như vậy, điểm khác biệt lớn nhất giữa hai bài toán tối ưu MLE và MAP là việc hàm mục tiêu của MAP có thêm $p(\theta)$, tức phân phối của θ. Phân phối này chính là những thông tin ta biết trước về θ và được gọi là *prior*. Ta kết luận rằng ***posterior* tỉ lệ thuận với tích của *likelihood* và *prior***.

Vậy chọn *prior* thế nào? chúng ta cùng làm quen với một khái niệm mới: *conjugate prior*.

4.3.2 Conjugate prior

Nếu phân phối xác suất posterior $p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *có cùng dạng* (*same family*) với phân phối xác suất $p(\theta)$, prior và posterior được gọi là *conjugate distributions*, và $p(\theta)$ được gọi là *conjugate prior* cho hàm likelihood $p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)$. Nghiệm của bài toán MAP và MLE có cấu trúc giống nhau.

Một vài cặp các *conjugate distributions*[1]:

- Nếu likelihood function là một Gaussian (phân phối chuẩn), và prior cho vector kỳ vọng cũng là một Gaussian, thế thì phân phối posterior cũng là một Gaussian. Ta nói rằng Gaussian conjugate với chính nó (hay còn gọi là *self-conjugate*).
- Nếu likelihood function là một Gaussian và prior cho phương sai là một phân phối gamma[2], phân phối posterior cũng là một Gaussian. Ta nói rằng phân phối gamma là conjugate prior cho phương sai của Gassian. Chú ý rằng phương sai có thể được coi là một biến giúp đo *độ chính xác* của mô hình. Phương sai càng nhỏ thì độ chính xác càng cao.
- Phân phối Beta là conjuate của phân phối Bernoulli.
- Phân phối Dirichlet là conjugate của phân phối categorical.

4.3.3 Hyperparameters

Xét một ví dụ nhỏ với phân phối Bernoulli với hàm mật độ xác suất:

$$p(x|\lambda) = \lambda^x(1-\lambda)^{1-x} \tag{4.36}$$

và conjugate của nó, phân phối Beta, có hàm phân mật độ xác suất:

$$p(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \tag{4.37}$$

Bỏ qua thừa số hằng số chỉ mang mục đích chuẩn hoá cho tích phân của hàm mật độ xác suất bằng một, ta có thể nhận thấy rằng phần còn lại của phân phối Beta có cùng *họ* (*family*)

[1] Đọc thêm: *Conjugate prior–Wikipedia* (*https://goo.gl/E2SHbD*).
[2] *Gamma distribution–Wikipedia*, (*https://goo.gl/kdWd2R*.)

với phân phối Bernoulli. Cụ thể, nếu sử dụng phân phối Beta làm *prior* cho tham số λ, và bỏ qua phần thừa số hằng số, posterior sẽ có dạng

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)p(\lambda) \\ &\propto \lambda^{x+\alpha-1}(1-\lambda)^{1-x+\beta-1} \end{aligned} \tag{4.38}$$

trong đó, $\propto$ là ký hiệu của *tỉ lệ với*.

Nhận thấy rằng (4.38) *vẫn có dạng của một phân phối Bernoulli*. Chính vì vậy mà phân phối Beta được gọi là một *conjugate prior* cho phân phối Bernoulli.

Trong ví dụ này, tham số λ phụ thuộc vào hai tham số khác là α và β. Để tránh nhầm lẫn, hai tham số (α, β) được gọi là *siêu tham số* (*hyperparameters*).

Quay trở lại ví dụ về bài toán tung đồng xu N lần có n lần nhận được mặt *head* và $m = N-n$ lần nhận được mặt *tail*. Nếu sử dụng MLE, ta nhận được ước lượng $\lambda = n/M$. Nếu sử dụng MAP với prior là một Beta$[\alpha, \beta]$ thì kết quả sẽ thay đổi thế nào?

Bài toán tối ưu MAP:

$$\begin{aligned} \lambda &= \underset{\lambda}{\text{argmax}} \left[p(x_1, \ldots, x_N|\lambda)p(\lambda)\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\left(\prod_{i=1}^{N} \lambda^{x_i}(1-\lambda)^{1-x_i}\right) \lambda^{\alpha-1}(1-\lambda)^{\beta-1}\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\lambda^{\sum_{i=1}^{N} x_i+\alpha-1}(1-\lambda)^{N-\sum_{i=1}^{N} x_i+\beta-1}\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\lambda^{n+\alpha-1}(1-\lambda)^{m+\beta-1}\right] \end{aligned} \tag{4.39}$$

Bài toán tối ưu (4.39) chính là bài toán tối ưu (4.38) với tham số thay đổi một chút. Tương tự như (4.38), nghiệm của (4.39) có thể được suy ra là

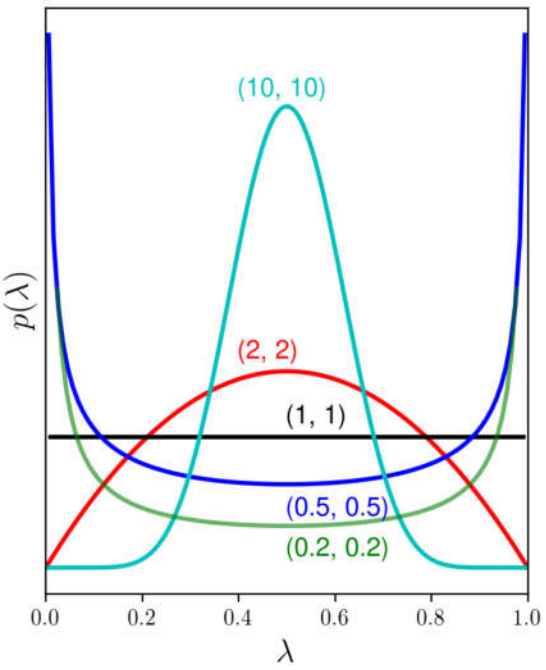
$$\lambda = \frac{n+\alpha-1}{N+\alpha+\beta-2} \tag{4.40}$$

Nhờ việc chọn prior phù hợp, ở đây là conjugate prior, posterior và likelihood có dạng giống nhau, khiến cho việc tối ưu bài toán MAP được thuận lợi.

Việc còn lại là chọn cặp *hyperparameters* α và β.

Chúng ta cùng xem lại hình dạng của phân phối Beta và nhận thấy rằng khi $\alpha = \beta > 1$, hàm mật độ xác suất của phân phối Beta đối xứng qua điểm 0.5 và đạt giá trị cao nhất tại 0.5. Xét Hình 4.1, ta nhận thấy rằng khi $\alpha = \beta > 1$, mật độ xác suất xung quanh điểm 0.5 nhận giá trị cao, điều này chứng tỏ λ có xu hướng gần với 0.5.

Nếu ta chọn $\alpha = \beta = 1$, ta nhận được phân phối đều vì đồ thị hàm mật độ xác suất là một đường thẳng. Lúc này, xác suất của λ tại mọi vị trí trong khoảng $[0, 1]$ là như nhau. Thực



Hình 4.1: Đồ thị hàm mật độ xác suất của phân phối Beta khi $\alpha = \beta$ và nhận các giá trị khác nhau. Khi cả hai giá trị này lớn, xác suất để λ gần 0.5 sẽ cao hơn.

chất, nếu ta thay $\alpha = \beta = 1$ vào (4.40) ta sẽ thu được $\lambda = n/N$, đây chính là ước lượng thu được bằng MLE. MLE là một trường hợp đặc biệt của MAP khi prior là một phân phối đều.

Nếu ta chọn $\alpha = \beta = 2$, ta sẽ thu được: $\lambda = \dfrac{n+1}{N+2}$. Chẳng hạn khi $N = 5, n = 1$ như trong ví dụ. MLE cho kết quả $\lambda = 1/5$, MAP sẽ cho kết quả $\lambda = 2/7$, gần với $1/2$ hơn.

Nếu chọn $\alpha = \beta = 10$ ta sẽ có $\lambda = (1+9)/(5+18) = 10/23$. Ta thấy rằng khi $\alpha = \beta$ và càng lớn thì ta sẽ thu được λ càng gần $1/2$. Điều này có thể dễ nhận thấy vì prior nhận giá trị rất cao tại 0.5 khi các siêu tham số $\alpha = \beta$ lớn.

4.3.4 MAP giúp tránh overfitting

Việc chọn các hyperparameter thường được dựa trên thực nghiệm, chẳng hạn bằng cross-validation. Việc thử nhiều bộ tham số rồi chọn ra bộ tốt nhất là việc mà các kỹ sư machine learning thường xuyên phải đối mặt. Cũng giống như việc chọn regularization parameter để tránh overfitting vậy.

Nếu viết lại bài toán MAP dưới dạng:

$$\theta = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{X}|\theta)p(\theta) \tag{4.41}$$

$$= \underset{\lambda}{\operatorname{argmax}} \left[\log \underbrace{p(\mathbf{X}|\theta)}_{\text{likelihood}} + \log \underbrace{p(\theta)}_{\text{prior}} \right] \tag{4.42}$$