Ước lượng Bayes

1. Một vài khái niệm cơ bản:

- Truy vấn xác suất.
- Truy vấn MAP.

- Các bài toán ước lượng đều là các bài toán tìm max – min. Liên quan đến môn tối ưu hoá (dù chưa nhiều lắm).
- Phần 1 là về dạng ước lượng đơn giản nhất.
- Mục tiêu: đưa ra những khái niệm đơn giản nhất về việc ước lượng giá trị xác suất hậu nghiệm.

Truy vấn:

- **Truy vấn xác suất**: Tính toán xác suất của một biến ngẫu nhiên, biết các điều kiện khác.
- Truy vấn xác suất có hai phần:

  1. Bằng chứng (“Evidence”): Tập E gồm các biến ngẫu nhiên điều kiện.
  2. Biến truy vấn: Tập Y gồm các biến cần truy vấn.

Truy vấn xác suất:

- Vấn đề suy luận: tính $\mathrm{P}(\mathrm{Y}|\mathrm{E} = \mathrm{e})$.
- Tính trên tất cả các giá trị y của Y.
- Phụ thuộc vào điều kiện $\mathrm{E} = \mathrm{e}$.
- Từ đó có thể ước lượng xác suất biên như sau:

$$\mathrm{P}(\mathrm{Y} = \mathrm{y_i}|\mathrm{E} = \mathrm{e}) = \frac{\mathrm{P}(\mathrm{Y} = \mathrm{y_i}, \mathrm{E} = \mathrm{e})}{\mathrm{P}(\mathrm{E} = \mathrm{e})}$$

Truy vấn MAP

- **MAP = Maximum a Posteriori**: Biến hậu nghiệm tối đại.
- Khác với truy vấn xác suất (tính xác suất), truy vấn MAP sẽ tìm giá trị biến hậu nghiệm để xác suất hậu nghiệm tối đại.
- Cho W là tập các biến hậu nghiệm cần truy vấn, E là điều kiện, ($W \subseteq \chi - E$), khi đó:

$$\mathrm{MAP}(W|E = e) = \underset{\{w\}}{\mathrm{argmax}}\, P(w|e)$$

Ví dụ về truy vấn MAP:

- Vấn đề chẩn đoán bệnh:
  Cho A là biến ngẫu nhiên bệnh tật với các giá trị: $a^0 =$ bệnh lao, $a^1 =$ bệnh cúm.
  Cho B là biến ngẫu nhiên triệu chứng với các giá trị: $b^0 =$ đau đầu, $b^1 =$ sốt.

  Tìm MAP $P(A|B)$ và MAP $P(A)$.

P(A)	a^0	a^1
	0,4	0,6

$P(B\|A)$	b^0	b^1
a^0	0,1	0,9
a^1	0,5	0,5

- MAP $\mathrm{P}(\mathrm{A}) = \operatorname{Arg}\max_{a} P(A)$

$$= a^1 = 0{,}6$$

Lời giải cho ví dụ:

- $\text{MAP P(A|B)} = \text{MAP P(A, B)}$

$$= \text{Arg}\max_{a,b} P(A,B)$$

$$= \text{Arg}\max_{a,b} P(A)P(B|A)$$

$$= \text{Arg}\max_{a,b}\{0{,}04;\ 0{,}36;\ 0{,}3;\ 0{,}3\} = a^0, b^1$$

Câu hỏi:

- Có thể tìm ra phân phối của B không ? (Tìm Xác suất biên của $P(b^0)$ và $P(b^1)$)

1*. Nhắc lại về Maximum likelihood:

- Ta nhắc lại ước lượng ML thông qua bài toán đã được giới thiệu từ buổi đầu tiên.

Tiếp cận đơn giản:

- Giả sử ta có một chuỗi tung đồng xu:

$$\mathrm{A} = \mathrm{H} - \mathrm{T} - \mathrm{T} - \mathrm{H} - \mathrm{H}$$

- Vì xác suất mỗi lần thảy độc lập, nên ta có:

$$\mathrm{P}(\mathrm{A}|\theta) = \theta^3(1-\theta)^2$$

Tiếp cận đơn giản:

- Như vậy, ta sẽ tìm $\mathrm{P}(\mathrm{A}|\theta)$ bằng ước lượng hợp lí cực đại maximum likelihood.
- Nói cách khác, ta sẽ tìm:

$$\mathrm{argmax}_\theta\ \mathrm{P}(\mathrm{A}|\theta) = \ \mathrm{argmax}_\theta \theta^3(1-\theta)^2$$

Ở đây $0 < \theta < 1$.

Tiếp cận đơn giản:

- Theo phương pháp trên, giá trị θ nào càng khiến cho xác suất $\mathrm{p}(\mathrm{A}|\theta)$ càng lớn thì θ đó càng tiến gần đến giá trị ta mong muốn.
- Như vậy phương pháp likelihood cho ta biết θ sao cho θ phù hợp với xác suất tiên nghiệm nhất.

Trường hợp tổng quát:

- Từ trường hợp cụ thể ở trên, ta sẽ xây dựng một mô hình tổng quát cho phương pháp MLE cho bài toán tung đồng xu.
- Gọi $\mathrm{M}[1]$ là số lần tung đồng xu ra mặt ngửa.
- Gọi $\mathrm{M}[0]$ là số lần tung đồng xu ra mặt sấp.
- Khi đó hàm Likelihood sẽ là:

$$\mathrm{P}(\mathrm{D}|\theta) = \mathrm{L}(\theta : \mathrm{D}) = \theta^{M[1]}(1-\theta)^{\mathrm{M}[0]}$$

Trường hợp tổng quát:

- Lấy log của likelihood, kí hiệu $\mathrm{Log\ L}(\theta\colon \mathrm{D}) = \mathrm{l}(\theta\colon \mathrm{D})$, khi đó ta có:

$$\mathrm{l}(\theta\colon \mathrm{D}) = \mathrm{M}[1]\ln\theta + \mathrm{M}[0]\ln(1-\theta)$$

- Lấy đạo hàm hai vế phải theo θ và cho đạo hàm đó bằng 0, ta có:

$$0 = \mathrm{M}[1].\frac{1}{\theta} + \mathrm{M}[0].\frac{1}{1-\theta}$$

- Lưu ý: ở trên chính là phương pháp tìm cực trị mà ta đã được học từ THPT.
- Như vậy, ta được:

$$\hat{\theta} = \text{argmax}_{\theta}\, \text{l}(\theta : \text{D}) = \frac{\text{M}[1]}{\text{M}[1] + \text{M}[0]} = \frac{M[1]}{n}$$

- Như vậy ta có thể thấy trong bài toán đơn giản này, tham số tốt nhất mà ta nhận được lại chính là phép tính xác suất bằng phép đếm cơ bản.
- Như vậy, nếu ta tung đồng xu ra mặt ngửa 3 lần trong 10 lần tung thì tham số tốt nhất ta được chính là:

$$\hat{\theta} = \frac{3}{10} = 0{,}3$$

2. Ước lượng tham số:

- Phương pháp tiếp cận theo trường phái Bayes cho phép ta có nhiều công cụ xử lí với các bài toán khác.
- Một trong những bài toán đó là bài toán về: lý thuyết chọn lựa (decision theory).
- Một trong những vấn đề của thuyết chọn lựa là tìm cách tính toán xem những lựa chọn nào là tốt hơn.

- Lúc này ta cần khái niệm về hàm risk.
- Hàm risk dùng để đo lường sự rủi ro khi ta áp đặt các tham số cho mô hình.
- Cả hai trường phái truyền thống lẫn Bayes đều có cách tính cho hàm risk, tuy nhiên chúng ta chỉ tìm hiểu về cách tính của trường phái Bayes.

a) Các khái niệm:

- Giả sử ta có một bộ dữ liệu có dạng vector y, và $\delta(y)$ là vector lựa chọn của y.
- Ta cũng có x là môt vector tham số.
- Như đã giới thiệu, ta cần “so sánh” giữa x và $\delta(y)$ để tìm x và delta phù hợp với mô hình.
- Như vậy ta có khái niệm hàm loss $L(x, \delta(y))$.

Thay đổi một vài kí hiệu:

- Ta sẽ viết:

❖ $p(x) = p(x|C)$ hoặc $\pi(x)$ là phân phối xác suất tiên nghiệm, trong đó C là thông tin của bộ dữ liệu.

❖ $p(y) = p(y|C)$ là phân phối xác suất của của bộ dữ liệu.

❖ $p(x|y) = p(x|y, C)$ là phân phối xác suất hậu nghiệm.

❖Lưu ý hàm p ở trên là hàm mật độ xác suất.

Hàm Loss:

- Khi làm việc với phương pháp suy luận của trường phái Bayes, cái khái niệm và kí hiệu cũng sẽ thay đổi theo.
- Khi này, ta có khái niệm “hàm mất mát kì vọng với hàm mật độ xác suất hậu nghiệm” (Expected loss function with posterior density):

$$E\big[L\big(x,\delta(y)\big)\big] = \int_{\chi} L\big(x,\delta(y)\big).\,p(x|y,C)dx$$

- Ở đây, ta có χ là không gian xác định của giá trị các tham số trong vector x.
- Ta có thể thấy rằng ta sẽ chọn delta sao cho nó khiến hàm loss hậu nghiệm đạt giá trị nhỏ nhất.
- Trong trường hợp này, ta gọi quy tắc chọn delta là quy tắc Bayes (Bayes rule).

Hàm risk:

- Trường phái cổ điển có một khái niệm gần như tương đương với hàm loss, đó là hàm risk.
- Về hàm risk, ta có công thức như sau:

$$R(\delta, x) = \int_{\gamma} L\big(x, \delta(y)\big).\, p(x|y, C)dy$$

- Ở đây không gian γ la không gian xác định của vector data y.
- Ta có thể thấy phương pháp tính toán trên đi theo quan điểm của trường phái cổ điển.
- Trường phái cổ điển dựa trên giá trị của vector x để tính toán hàm risk.
- Trong trường hợp này x cố định.

Cổ điển và Bayes:

- Tuy nhiên hai công thức về hàm risk và loss không mâu thuẫn với nhau.
- Xét hàm mật độ xác suất tiên nghiệm $p(x|C)$, khi đó ta có hàm Bayes risk:

$$r(\delta) = \int_{\chi} R(x, \delta) p(x|C) dx$$

- Trong trường hợp này việc tìm giá trị nhỏ nhất cho hàm Bayes risk cũng sẽ dẫn đến giá trị nhỏ nhất của hàm loss:

$$r(\delta) = \int_{\chi} \int_{\gamma} L\big(x, \delta(y)\big).\, p(x|y, C)\, p(x|C) dy dx$$

$$= \int_{\gamma} E\Big[L\big(x, \delta(y)\big)\Big].\, p(y|C) dy$$

- Trong trường hợp trên

b. Ước lượng điểm:

- Theo như các công thức ở trên, việc ước lượng vector tham số x là cần thiết.
- Có hai cách ước lượng dành cho x:

❖ Ước lượng giá trị nhỏ nhất của vector x.

❖ Ước lượng khoảng tin cậy của x.

Hàm loss bậc hai:

- Trong phần này ta sẽ tìm hiểu về ước lượng điểm.
- Cụ thể, ta sẽ tìm $\hat{x} = \delta(y)$.
- Ta gọi đặt:

$$x - \hat{x}$$

Là vector khoảng cách giữa x và $\hat{x}$.

- Lúc này ta có hàm loss đơn giản được xác định bởi:

$$L(x, \hat{x}) = (x - \hat{x}).(x - \hat{x})^T$$

- Đây là một đa thức bậc hai. Vì thế ta gọi nó là hàm loss bậc hai (quadratic loss function).

Trường hợp đơn giản:

- Ta coi $x = \theta$ và $\hat{x} = \hat{\theta}$ trong đó:

❖ θ là một biến thực.

❖ $\widehat{\theta}$ là một số thực dương.

- Trong trường hợp này ta tìm hiểu công thức hàm loss bậc hai đơn giản:

$$L(x, \hat{x}) = \left(\theta - \hat{\theta}\right)^2$$

- Như vậy ta có Bayes risk:

$$r(\delta) = E_{\theta|y}[L(\delta, \theta)]$$

$$= \int_{\Theta} \left(\hat{\theta} - \theta\right)^2 p(\theta|y) d\theta$$

Một bổ đề quan trọng:

- Trong trường hợp một biến, nếu $L(x, \hat{x}) = \left(\theta - \hat{\theta}\right)^2$, ước lượng Bayes cho tham số θ là $E(\theta|y)$.
- Khi đó, hàm Bayes risk sẽ là:

$$E[E(\theta|y) - \theta]^2 = Var(\theta|y)$$

Bài tập 1:

- Chứng minh bổ đề trên.
- Gợi ý: sử dụng định nghĩa của ước lượng Bayes và định nghĩa về kì vọng và phương sai.

Ví dụ:

- Cho $X_1, \dots, X_n$ là các biến ngẫu nhiên từ phân phối Bernouli với tham số θ. Xét phân phối $Beta(\alpha, \beta)$ là phân phối liên hợp của tham số theta.
- Khi đó ta có phân phối xác suất hậu nghiệm của bài toán này :

$$\theta|y \sim Beta(\alpha + t, \beta + n - t)$$

Với $t = \sum x_i$

Ví dụ:

- Chứng minh điều trên. (tham khảo slide đầu tiên)
- Bằng bổ đề đơn giản đã đưa ra, ta có:

$$E(\theta|y) = \frac{\alpha + \sum x_i}{\alpha + \beta + n}$$

- Có được nhờ công thức trung bình của phân phối beta (hoặc là phân phối Bernoulli).

Quay lại với công thức tổng quát

- Ta sẽ quay lại với bài toán tổng quát cho trường hợp x là một vector ngẫu nhiên và y là vector dữ liệu.
- Giả sử ta có $x = [X_1, \dots, X_n]$.
- Đặt $E(X_i) = \mu_i$. Khi đó ta có covariance σ_{ij} được định nghĩa như sau:

$$\sigma_{ij} = E\left((X_i - \mu_i)(X_j - \mu_j)\right)$$

- Theo ngôn ngữ của tích phân, ta được:

$$\sigma_{ij} = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} (x_i - \mu_i)\left(x_j - \mu_j\right) p(x_1 \dots x_n | C) dx_1 \dots dx_n$$

Chú thích: $p(x_1 \dots x_n)$ là hàm mât độ của phân phối hợp $P(X_1 \dots X_n)$

- Đặt $E(X_i) = \mu_i$. Khi đó ta có variance $\sigma_i^2 = \sigma_{ii}$ được định nghĩa như sau:

$$\sigma_{ii} = E\left((X_i - \mu_i)^2\right)$$

$$= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} (x_i - \mu_i)^2 p(x_1 \dots x_n | C) dx_1 \dots dx_n$$

Đây cũng chính là phương sai của X_i

Ma trận variance – covariance:

- Cho một vector ngẫu nhiên $x = [X_1, \dots, X_n]$, ta gọi ma trận $D(x)$ là ma trân variance và covariance nếu:

$$D(x) = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{n1} \\ \vdots & \ddots & \vdots \\ \sigma_{1n} & \cdots & \sigma_n^2 \end{bmatrix}$$

- Như vậy ta có thể định nghĩa hàm loss tổng quát như sau:

$$L(x, \hat{x}) = (x - \hat{x})^T . D(x)^{-1} . (x - \hat{x})$$

- Lưu ý: $D(x)$ luôn là ma trận vuông và là một ma trận xác định dương trên trường số thực.

Hàm risk và luật Bayes:

- Sau khi có hàm loss, ta sẽ tìm được công thức của hàm risk như sau:

$$r(\delta) = \int_{\gamma} E[L(x, \hat{x})].p(y|C)dy$$

$$= \int_{\gamma} E[x^T D(x)^{-1} x\ ].p(y|C)dy$$

Tính toán $E[x^T D(x)^{-1} x\ ]$:

- Theo như công thức ở trên, ta chỉ còn bước tính $E[(x-\hat{x})^T D(x)^{-1}(x-\hat{x})\ \ ]$.
- $E[x^T D(x)^{-1} x\ ]$ có thể được tính như sau:

$$E[(x-\hat{x})^T D(x)^{-1}(x-\hat{x})\ \ ]$$

$$= E\left[\left(x - E(x) - \left(\hat{x} - E(x)\right)\right)^T D(x)^{-1}\left(x - E(x) - \left(\hat{x} - E(x)\right)\right)\right]$$

Ta được tiếp:

$$E[(x-\hat{x})^T D(x)^{-1}(x-\hat{x})\ ]$$

$$= E\left[\left(x-E(x)\right)^T D(x)^{-1}\left(x-E(x)\right)\right] + E\left[\left(\hat{x}-E(x)\right)^T D(x)^{-1}\left(\hat{x}-E(x)\right)\right] (*)$$

(Hãy thử viết ra biểu thức từ slide trước ra biểu thức *)

- Trong trường hợp biểu thức (*) có hai biểu thức nhỏ.
- Biểu thức đầu tiên là
  $E\left[\left(x-E(x)\right)^{T}D(x)^{-1}\left(x-E(x)\right)\right]$ không dựa vào $\hat{x}$.
- Biểu thức thứ hai:

$\left(\hat{x}-E(x)\right)^{T}D(x)^{-1}\left(\hat{x}-E(x)\right)$ đạt giá trị nhỏ nhất khi và chỉ khi $E[(x-\hat{x})^{T}D(x)^{-1}(x-\hat{x})\ ]$ đạt giá trị nhỏ nhất.

- Ta có:

$$Min\ E[(x-\hat{x})^T D(x)^{-1}(x-\hat{x})]$$

$$= Min \int_{\chi} [(x-\hat{x})^T D(x)^{-1}(x-\hat{x})].p(x|y,C)dx \quad (**)$$

- Thông qua (**), ta được $E[(x-\hat{x})^T D(x)^{-1}(x-\hat{x})]$ đạt giá trị nhỏ nhất khi và chỉ khi $\hat{x} = E(x|y)$.
- Điều này có được vì khi thực hiện phép đổi biến trên ma trận, ta có:

$$Min \int_{\chi} [(x-\hat{x})^T D(x)^{-1}(x-\hat{x})].p(x|y,C)dx$$

$$= Min \int_{\chi} [(x|y-\hat{x})^T D(x|y)^{-1}(x|y-\hat{x})].p(x|y,C)dx$$

- Khi $\hat{x} = E(x|y)$, ta được:

$$\int_{\chi} [(x|y - \hat{x})^T D(x|y)^{-1}(x|y - \hat{x})].p(x|y, C)dx = E(x|y)$$

Khi này biểu thức đạt tới kì vọng và đạt giá trị nhỏ nhất.

Lưu ý: $\int_{\chi} [(x|y - \hat{x})^T D(x|y)^{-1}(x|y - \hat{x})].p(x|y, C)dx$ có thể coi là một đa thức bậc hai nên giá trị nhỏ nhất của nó chính là đáy, cũng như là giá trị kì vọng.

- Khi $\hat{x}$ đạt cực tiểu thì nó chính là "ước lượng Bayes" (Bayes estimator), ta gọi nó là $\hat{x}_B$.
- Khi có $\hat{x}_B$, việc tính toán còn lại sẽ đơn giản theo các công thức đã có.

$$E[L(x, \hat{x}_B)] = E\{tr[D(x)^{-1}.(x - \hat{x}_B)(x - \hat{x}_B)^T]\} \ (***)$$

Lưu ý: $(x - \hat{x}_B)D(x)^{-1}(x - \hat{x}_B)^T = tr[D(x)^{-1}.(x - \hat{x}_B)(x - \hat{x}_B)^T]$

- Mà ta cũng có:

$$D(x|y) = E\left(\left(x - E(x|y)\right).(x - E(x|y)^T)\right) \text{ (định nghĩa)}$$

$$= \int_{\chi} (x - \hat{x}_B)(x - \hat{x}_B)p(x|y, C)dx$$

- Như vậy ta có (***) là:

$$E\{tr[D(x)^{-1}.(x-\hat{x}_B)(x-\hat{x}_B)^T]\}$$

$$=\int_{\chi} tr[D(x)^{-1}.(x-\hat{x}_B)(x-\hat{x}_B)^T]p(x|y,C)dx$$

- Ta có công thức: $tr(AB) = tr(A).tr(B)$ (thử chứng minh).
- Khi dó ta được:

$$\int_{\chi} tr[D(x)^{-1}.(x-\hat{x}_B)(x-\hat{x}_B)^T]p(x|y,C)dx$$

$$= \int_{\chi} tr[D(x)^{-1}].tr[(x-\hat{x}_B)(x-\hat{x}_B)^T]p(x|y,C)dx$$

$$= tr[D(x)^{-1}] \int_{\chi} tr[(x-\hat{x}_B)(x-\hat{x}_B)^T]p(x|y,C)dx$$

- Ta lại có:

$$\int_{\chi} tr[(x - \hat{x}_B)(x - \hat{x}_B)^T]p(x|y, C)dx$$

$$= tr \int_{\chi} [(x - \hat{x}_B)(x - \hat{x}_B)^T]p(x|y, C)dx = tr\ D(x|y)$$

- Như vậy ta được:

$$E[L(x, \hat{x}_B)] = tr[D(x)^{-1}].\, tr\ D(x|y)$$

$$= tr[D(x)^{-1}.\, D(x|y)]$$

- Hàm bayes risk sẽ là:

$$E(E(x|y) - x)]^2$$

$$= Var(x|y)$$

Ví dụ:

- Ta xét s là một đại lượng chưa biết, chẳng hạn như là độ đo góc của một cung đường, $y = [y_1, \dots, y_n]$ là một data vector.
- Giả sử ta có phân phối xác suất:

$$y_i | s \sim N(s, \sigma^2)$$

Với σ cố định.

- Như vậy ta sẽ có:

$$E(y_1|s) = s, V(y_1) = \sigma^2$$

$$\ldots$$

$$E(y_n|s) = s, V(y_n) = \sigma^2$$

- Theo phương pháp của suy luận Bayes, ta có giả sử :

$$s \sim N(\mu_s, \sigma^2\sigma_s^2)$$

- Bằng phương pháp đã được học, ta tính ra được:

$$s|y \sim N(\mu_{0s}, \sigma^2\sigma_{0s}^2)$$

- Trong đó:

$$\mu_{0s} = \frac{\sum_{i=1}^{n} y_i + \frac{\mu}{\sigma_s^2}}{n + \frac{1}{\sigma_s^2}}$$

$$\sigma_{0s}^2 = \frac{1}{n + (\sigma_s^2)^{-1}}$$

- Ước lượng Bayes của s là:

$$\hat{s}_B = E(x|y) = E(s|y) = \mu_{0s}$$

- Thông qua ước lượng Bayes, ta được hàm bayes risk:

$$\mathrm{r}(\delta) = V(s|y) = \int_{\chi} (s - \hat{s}_B)^2 p(s|y, C) ds = D(s|y) = \sigma^2 \sigma_s^2$$

Bài tập 2:

- Chứng minh với n = 1, ta có được công thức hàm loss được nhắc ở trên.
- Viết công thức với n = 2 và bài toán có phân phối Bernoulli.

Bài tập 3:

- Tìm ước lượng bayes bình phương cho trường hợp $X_1, \dots, X_n$ là bộ dữ liệu có phân phối Poisson và tham số θ.
- Tính hàm bayes risk.

Ước lượng điểm:

- Trong phần này ta sẽ tìm hiểu về ước lượng điểm.
- Cụ thể, ta sẽ tìm $\hat{x} = \delta(y)$.
- Trong trường hợp này ta có hàm error, môt phiên bản đặc biệt của hàm loss, ta gọi là hàm lỗi toàn phần:

$$L(x, \hat{x}) = |x - \hat{x}|$$

Lưu ý:

- Điểm chung thú vị: trên thực tế các bài toán vừa làm đều thông qua việc ước lượng giống như ước lượng MLE ban đầu.
- Vì hàm loss bình phương là một hàm lồi và có dạng đa thức, cho nên cách tính giá trị nhỏ nhất cũng tương đồng với cách tìm giá trị nhỏ nhất như bài tung đồng xu.