XSTK

Xác Suất Thống Kê

Võ Chí Công

Table of contents

Lời nói đầu

Xác suất thống kê là nền tảng giúp ích cho việc phân tích dữ liệu, tổ chức thực hành thí nghiệm, chạy các mô hình giả lập, giải các bài toán tìm nghiệm tối ưu, nghiên cứu và ứng dụng học máy.

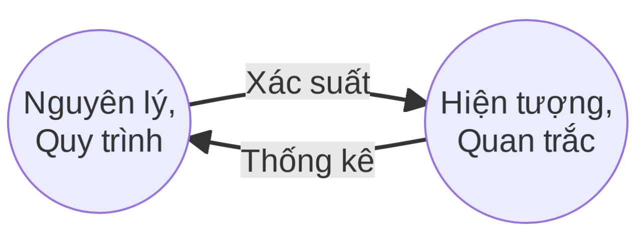
Môn "xác suất" tôi đã học những kiến thức cơ bản vài lần, tính ra là ở cấp 3, trong đại học, và gần đây là học trực tuyến. Môn "thống kê" tôi chưa học được cho ra bài bản lần nào, mấy tháng đầu năm 2022 có thử sức học trực tuyến nhưng đầu tư thời gian không đủ nên thi rớt thảm hại.

Lần này nhất quyết học lại môn thống kê một cách nghiêm túc hơn, tôi tóm tắt lại kiến thức xác suất thống kê bằng tiếng Việt, mặc dù tài liệu học hầu hết là tiếng Anh, tiếng Nhật. Hy vọng tiếng mẹ đẻ sẽ giúp tôi hiểu rõ hơn các vấn đề, và trau dồi vốn từ vựng để chia sẻ kiến thức với các đồng nghiệp và bạn bè người Việt.

Động cơ của việc học xác suất thống kê của tôi là để hiểu rõ hơn các lý thuyết căn bản trong ngành học máy và trí tuệ nhân tạo và áp dụng vào thực tiễn một cách đúng đắn, an toàn và công bằng hơn.

Mục tiêu cụ thể trước mắt tôi đặt ra là học hiểu và lấy được chứng chỉ hoàn thành khoá học Fundamentals of Statistics của Philippe Rigollet (2022), giáo sư đại học MIT dạy trên nền tảng trực tuyến edX. Tài liệu tham khảo là quyển "All of Statistics" của Wasserman (2004), cũng chính là tài liệu tham khảo của khoá học nêu trên.

Môn xác suất nghiên cứu cách suy luận ra các đặc tính của tập dữ liệu sẽ được tạo ra từ một nguyên lý, quy trình sản sinh dữ liệu. Ngược lại, môn thống kê nghiên cứu cách dự đoán đặc tính của một quy trình sản sinh dữ liệu từ tập dữ liệu về hiện tượng đã phát sinh và được quan sát. Hình 0.1 minh hoạ quan hệ giữa "xác suất" và "thống kê".



Hình 0.1: Xác suất và thống kê.

Phân tích, khai thác dữ liệu, học máy và khoa học dữ liệu là những tên gọi khác của thống kê, tùy theo bối cảnh và trào lưu. Một số ứng dụng cụ thể của thống kê là tính toán hồi quy, mật độ, phân loại và giả lập.

Tài liệu này không đi sâu vào các chứng minh chi tiết, nhưng sẽ cố gắng ghi rõ các công thức và định nghĩa. Thuật ngữ chuyên môn trong tài liệu này chắc chắn có nhiều chỗ không chuẩn chỉnh do vốn tiếng Việt và kiến thức hạn chế của tác giả. Xin vui lòng góp ý tại GitHub issues.

Tài liệu này được viết bằng các công cụ là Quarto và VSCode. Truy cập trực tuyến tại xstk .

Phần I

Xác suất

1 Xác suất

1.1 Sự kiện

Không gian Ω là tập hợp chứa tất cả những **hiện tượng** ω có thể xảy ra từ một **thí nghiệm**. Các tập con của Ω là các **sự kiện**.

Ví dụ xem xét thí nghiệm tung một đồng xu đúng hai lần, quan sát đồng xu rớt xuống nằm ngửa (H) hay sấp (T), ta có $\Omega = \{HH, HT, TH, TT\}$ bao gồm 4 kết quả có thể xảy ra. Sự kiện lần tung đầu tiên ra mặt ngửa của đồng xu là tập hợp $\{HH, HT\}$.

Cho một sự kiện $A \subseteq \Omega$, ta nói A **xảy ra**, hoặc A là **đúng**, nếu có một hiện tượng $\omega \in A$ **xảy ra**. Sự kiện **ngược lại** với A là $A^c := \Omega - A := \{\omega \in \Omega : \omega \notin A\}$, tức là "không xảy ra A". Theo định nghĩa, rõ ràng Ω luôn luôn đúng, còn sự kiện rỗng $\emptyset \equiv \Omega^c$ luôn luôn sai. Cho thêm sự kiện B, ta có $A \cup B$ là sự kiện "A **hoặc** B ít nhất một việc xảy ra", còn $AB := A \cap B$ là sự kiện "A **và** B đồng thời xảy ra".

Chuỗi sự kiện $A_1, A_2, \dots$ được gọi là **phân ly** nếu $A_i A_j \equiv \emptyset$ với mọi $i \neq j$. Khi đó nếu $A_1 \cup A_2 \cup \cdots \equiv C$ thì ta nói $A_1, A_2, \dots$ là một cách **phân hoạch** sự kiện C.

1.2 Xác suất

Nếu một xạ ảnh $\mathbb{P}$ từ không gian các sự kiện $A \subseteq \Omega$ lên tập hợp số thực $\mathbb{R}$ thỏa mãn các điều kiện:

1. $\mathbb{P}(A) \geq 0 \quad \forall A$
2. $\mathbb{P}(\Omega) = 1$
3. Nếu chuỗi $A_1, A_2, \dots$ phân hoạch C thì $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \mathbb{P}(C)$

thì ta gọi $\mathbb{P}$ là một **phân phối xác suất** hoặc **độ đo xác suất**.

Có hai cách cắt nghĩa khái niệm xác suất là tần suất và niềm tin. Theo cách hiểu tần suất thì $\mathbb{P}(A)$ chính là tỷ lệ số lần sự kiện A xảy ra nếu ta thực hiện thí nghiệm vô số lần. Còn theo cách hiểu niềm tin thì $\mathbb{P}(A)$ là thước đo mức độ mà một quan sát viên tin tưởng rằng hiện tượng A sẽ xảy ra.

1.2.1 Biến ngẫu nhiên

Định nghĩa 1.1. ("random variable", *rv*) là một quy tắc ánh xạ $X : \Omega \to \mathbb{R}$ gán cho mỗi hiện tượng ω trong không gian Ω một số thực $X(\omega)$.

1.2.2 Điểm cắt

Định nghĩa 1.2. Điểm cắt tại mức $1 - \alpha$ của biến X là một số q_α mà $\mathbb{P}(X \le q_\alpha) = 1 - \alpha$.

1.2.3 Độc lập

Định nghĩa 1.3. Hai sự kiện A và B là độc lập nếu $\mathbb{P}(AB) \equiv \mathbb{P}(A)\mathbb{P}(B)$.

Hai biến X và Y là độc lập nếu hai sự kiện $X \le x$ và $Y \le y$ là độc lập đối với mọi x, y.

1.2.4 IID

Định nghĩa 1.4. Các biến ngẫu nhiên $X_1, X_2, .$ được gọi là iid ("independent and identically distributed", "độc lập và phân phối giống nhau") nếu chúng cùng tuân theo duy nhất một phân phối xác suất, và từng cặp biến là độc lập với nhau. Dùng biến X để thể hiện phân phối xác suất chung đó, ta viết

$$X_1, \dots, X_n \overset{iid}{\sim} X.$$

1.2.5 Xác suất hợp

Định nghĩa 1.5. Ký hiệu $\mathbb{P}(A, B)$ hoặc $\mathbb{P}(A \cap B)$ chỉ xác suất sự kiện A và sự kiện B đồng thời xảy ra.

1.2.6 Xác suất có điều kiện

Định nghĩa 1.6. Ký hiệu $\mathbb{P}(A|B)$ hoặc $\mathbb{P}_B(A)$ chỉ xác suất của sự kiện A, khi biết sự kiện B đã xảy ra,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}.$$

Định lý 1.1 (Định lý Bayes)**.**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}.$$

1.3 Trung bình

Định nghĩa 1.7. Trung bình, hay **giá trị kỳ vọng** của biến X là

$$\mathbb{E}[X] := \int x dF(x) := \begin{cases} \sum_x x p(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

Nếu giá trị $\int |x| dF(x) < \infty$ ta nói là $\mathbb{E}[X]$ "tồn tại" (well-defined).

Ta có

$$\mathbb{E}[X+Y] \equiv \mathbb{E}[X] + E[Y].$$

1.4 Hiệp phương sai

Định nghĩa 1.8.

$$\begin{aligned} \mathrm{Cov}[X, Y] &:= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] && (1.1) \\ &\equiv \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && (1.2) \end{aligned}$$

Nếu X và Y độc lập thì $\mathrm{Cov}[X, Y] = 0$.

1.5 Phương sai

Định nghĩa 1.9. Phương sai, hay phân tán (variance, ⊠⊠) là

$$\mathbb{V}[X] := \mathrm{Cov}[X, X] \equiv \mathbb{E}[X^2] - \mathbb{E}^2[X].$$

Ta có

$$\mathbb{V}[X + Y] \equiv \mathbb{V}[X] + \mathbb{V}[Y] + 2\mathrm{Cov}[X, Y].$$

1.6 Tích suất

Tích suất (moment, ⊠⊠) thể hiện trọng tâm, độ phân tán, hay độ lệch của phân phối. Tích suất bậc n của biến X với mật độ $f(x)$ là:

$$\mathbb{E}[X^n] := \int x^n f(x) dx$$

1.6.1 Hàm tạo tích suất

Định nghĩa 1.10. "Moment generation function (MGF)" của biến X là

$$\psi_X(t) := \mathbb{E}[e^{tX}], t \in \mathbb{R}.$$

Nếu MGF "tồn tại" xung quanh 0 thì đạo hàm cấp k của ψ tại 0 chính là:

$$\psi_X^{(k)}(0) \equiv \mathbb{E}[X^k].$$

2 Phân phối

Có một số phân phối xác suất thông dụng.

2.1 Phân phối Bernoulli

Định nghĩa 2.1. $X \sim \text{Ber}(p)$ có

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

và $\mathbb{E}[X] = p, \mathbb{V}[X] = p(1 - p)$.

2.2 Phân phối Binomial

Định nghĩa 2.2. $X \sim \text{Bin}(n, p)$ với $n \in \mathbb{Z}_+, p \in (0, 1)$ mô tả tổng số lần thành công của n thí nghiệm độc lập $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$. Ta có

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

và $\mathbb{E}[X] = np, \mathbb{V}[X] = np(1 - p)$.

2.3 Phân phối Poisson

Định nghĩa 2.3. $X \sim \text{Poi}(\lambda)$ thường dùng để mô tả số lần k mà sự kiện phát sinh trong một giới hạn cố định, với giả định tần suất phát sinh sự kiện là $\lambda > 0$ cố định, và các sự kiện phát sinh độc lập.

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, 2, \ldots$$

và $\mathbb{E}[X] = \mathbb{V}[X] = \lambda$.

Khi n đủ lớn và p đủ nhỏ thì $\text{Poi}(np)$ gần với $\text{Bin}(n, p)$.

2.4 Phân phối Geometric

Định nghĩa 2.4. $X \sim \text{Geom}(p)$, $p \in (0, 1)$ có

$$\mathbb{P}(X = k) = p(1 - p)^{k-1}, k = 1, 2, \ldots$$

và $\mathbb{E}[X] = 1/p, \mathbb{V}[X] = (1 - p)/p^2$.

2.5 Phân phối Exponential

Định nghĩa 2.5. $X \sim \text{Exp}(\lambda)$ dùng để mô tả khoảng cách x giữa hai lần phát sinh của một chuỗi sự kiện kiểu Poisson (hai lần phát sinh sự kiện liên tiếp là độc lập với nhau, và tần suất phát sinh $\lambda > 0$ là cố định). Có

$$f(x) = \lambda e^{-\lambda x}, x \in \mathbb{R}_+$$

và $\mathbb{E}[X] = 1/\lambda, \mathbb{V}[X] = 1/\lambda^2$.

2.6 Phân phối chuẩn

Gaussian Distribution

Định nghĩa 2.6 (Phân phối chuẩn 1 biến)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$ có mật độ

$$f(x) = \frac{1}{\sqrt{\pi 2\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), x \in \mathbb{R}$$

và $\mathbb{E}[X] = \mu, \mathbb{V}[X] = \sigma^2$.

Định lý 2.1 (Tổng của các phân phối chuẩn)**.** *Nếu* $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)(i = 1, \dots, n)$ *thì*

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Định nghĩa 2.7 (Vector phân phối chuẩn)**.** Nếu mọi tổ hợp tuyến tính các yếu tố của vector $\boldsymbol{v}$ đều thuộc phân phối chuẩn 1 biến thì $\boldsymbol{v}$ được gọi là một vector phân phối chuẩn.

2.7 Phân phối χ^2

Nếu $X_1, \dots, X_k \overset{iid}{\sim} \mathcal{N}(0, 1)$ thì tổng bình phương của chúng

$$X := \sum_{i=0}^{k} X_i^2.$$

tuân theo phân phối χ_k^2 và có $\mathbb{E}[X] = k, \mathbb{V}[X] = 2k$.

2.8 Phân phối Student's T

Nếu có $Z \sim \mathcal{N}(0,1), V \sim \chi_k^2$ độc lập với nhau thì

$$X := \frac{Z}{\sqrt{V/k}}$$

tuân theo phân phối Student's t_k.

3 Hội tụ

Có một số kiểu hội tụ của biến ngẫu nhiên.

3.1 Hội tụ xác suất

Định nghĩa 3.1. Cho một chuỗi biến ngẫu nhiên $X_1, X_2, .$ và một biến ngẫu nhiên X.

1. Hội tụ gần tuyệt đối:

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \iff \mathbb{P}(\{\omega \in \Omega \, : \, X_n(\omega) \to X(\omega)\}) = 1.$$

2. Hội tụ theo xác suất:

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \iff \mathbb{P}(|X_n - X| > \epsilon) \xrightarrow[n\to\infty]{} 0, \quad \forall \epsilon > 0.$$

3. Hội tụ theo phân phối:

$$X_n \xrightarrow[n\to\infty]{(d)} X \iff \mathbb{P}[X_n(x) \le x] \xrightarrow[n\to\infty]{} \mathbb{P}[X \le x]$$

   tại mọi điểm x mà cdf của X liên tục.

4. Hội tụ về chuẩn: (asymptotically normal) với phương sai σ^2

$$\sqrt{n}\left(\frac{X_n - X}{|\sigma|}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1).$$

3.2 Độ mạnh

Định lý 3.1. *Xếp theo thứ tự từ mạnh đến yếu:*

$$X_n \xrightarrow[n\to\infty]{a.s.} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{(d)} X.$$

Nếu $X_n \xrightarrow[n\to\infty]{(d)} X$*, và* X *có mật độ xác suất, thì* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Nếu chuỗi X_n *có* $\mathbb{E}[X_n] \to \mu$ *và* $\mathbb{V}[X_n] \to 0$ *thì* $X_n \xrightarrow{\mathbb{P}} \mu$.

Nếu $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *thì* $\mathbb{P}(a \leq X_n \leq b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \leq X \leq b)$ *với mọi khoảng* $[a, b]$.

3.3 Tổng và tích

Định lý 3.2. *Nếu có hai chuỗi biến ngẫu nhiên* X_n, Y_n *hội tụ gần tuyệt đối hoặc hội tụ theo xác suất về* X, Y*, thì tổng* $X_n + Y_n$ *và tích* X_nY_n *cũng hội tụ tương tự (gần tuyệt đối, hoặc theo xác suất) về tổng* $X + Y$ *và tích* XY.

3.4 Slutsky

Định lý 3.3. *Hơn nữa, ở Định lý 3.2 nếu* $Y_n \xrightarrow{\mathbb{P}} y$*,* y *là một số thực cố định thì có thể nới lỏng điều kiện đối với* X_n *thành hội tụ theo phân phối.*

3.5 Ánh xạ liên tục

Định lý 3.4. *(Continuous mapping) Nếu* $X_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} X$ *thì đối với mọi hàm* f *liên tục:*

- $f(X_n) \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} f(X)$.
- $\mathbb{E}[f(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(X)]$ *nếu* f *còn bị chặn.*

3.6 Đại đa số (Law of Large Numbers)

Định lý 3.5. *Cho* n *biến iid* $X_1, X_2, ., X_n$ *có chung* $\mathbb{E}[X_i] = \mu < \infty$*. Khi đó:*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P},\, a.s.} \mu.$$

3.7 Hội tụ trung tâm (Central Limit)

Định lý 3.6. *Giả sử thêm là* $\mathbb{V}[X_i] = \sigma^2 < \infty$*. Khi đó*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{|\sigma|}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

3.8 Bất đẳng thức Hoefding

Định lý 3.7. *Nếu có một khoảng cố định* $[a, b]$ *gần như tuyệt đối (almost surely) chứa các biến* $X_i (i = 1, 2, \dots, n)$ *thì*

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \epsilon] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}, \quad \forall \epsilon > 0.$$

3.9 Phương pháp Delta

Định lý 3.8. *Giả sử chuỗi $(\theta_n)_{n\geq 1}$ chuẩn tiến với phương sai σ^2 về một điểm $\theta \in \mathbb{R}$. Giả sử $g : \mathbb{R} \to \mathbb{R}$ có vi phân g' liên tục tại θ. Khi đó,*

$$\sqrt{n}\left(\frac{g(\theta_n) - g(\theta)}{|\sigma g'(\theta)|}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

Phần II

Thống kê

4 Thống kê

4.1 Mô hình

Trên không gian đo được $E \subseteq \mathbb{R}$ ta quan sát các mẫu $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$. Với tập tham số Θ ta xây dựng bộ độ đo xác suất $(\mathbb{P}_\theta)_{\theta \in \Theta}$ để mô phỏng $\mathbb{P}$.

Nếu tồn tại $\theta \in \Theta$ để $\mathbb{P}_\theta \equiv \mathbb{P}$ thì ta nói mô hình "hợp lệ" (well specified).

Nếu từ tập Θ, ánh xạ $\theta \mapsto \mathbb{P}_\theta$ là đơn ánh thì ta nói θ (hoặc $\mathbb{P}_\theta$) là "có thể xác định" (identifiable).

Nếu $\Theta \subseteq \mathbb{R}^d$ thì mô hình được nói có số biến hữu hạn (parametric model).

4.2 Ước lượng

Thống kê lượng (statistic) θ là bất cứ hàm số nào có thể đo được (measurable function) trên tập dữ liệu đối tượng X_i.

Đánh giá (estimator) θ_n đối với mục tiêu thống kê θ là một thống kê khác không phụ thuộc vào θ.

Đánh giá được gọi là nhất quán (consistent) nếu nó hội tụ về mục tiêu (Định nghĩa 3.1).

Độ lệch (bias) giữa đánh giá θ_n với mục tiêu θ là

$$\mathrm{bias}_\theta(\theta_n) := \mathbb{E}[\theta_n] - \theta.$$

Sai số bậc 2 (quadratic risk, mean squared error) giữa đánh giá θ_n với mục tiêu θ là

$$\mathrm{MSE}(\theta_n) := \mathbb{E}[(\theta_n - \theta)^2] \equiv \mathrm{bias}_\theta(\theta_n)^2 + \mathbb{V}[\theta_n]$$

bao hàm cả độ lệch và độ nhiễu của đánh giá.

4.3 Khoảng tin cậy

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ xây dựng dựa trên quan sát $X_1, \dots, X_n$, giả sử số $\alpha \in (0, 1)$. Khoảng tin cậy (confidence interval) cấp $1-\alpha$ đối với θ là một khoảng ngẫu nhiên $\mathcal{I}$ (phụ thuộc vào $X_1, \dots, X_n$, không phụ thuộc vào θ), mà xác suất $\mathcal{I}$ có chứa θ là đủ cao:

$$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

Nếu

$$\lim_{n\to\infty} \mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall \theta \in \Theta$$

thì ta gọi $\mathcal{I}$ là khoảng tin cậy tiệm cận cấp $1 - \alpha$ đối với θ.

5 Ước lượng

Chiến lược ước lượng phân phối *thật* (true distribution)

- Định nghĩa khoảng cách giữa các phân phối (TV distance)
- Ước lượng khoảng cách nêu trên (KL divergence)
- Tìm điểm cực tiểu của ước lượng nêu trên (minimization).

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ xây dựng dựa trên quan sát iid rv $X_1, \ldots, X_n$ trên tập mẫu E và bộ tham số Θ . Ngầm định tồn tại *tham số thật* $\theta^* \in \Theta$ để $X_1 \sim \mathbb{P}_{\theta^*}$.

5.1 Pivotal statistic

Một phân phối hay một ước lượng được gọi là pivotal nếu nó không phụ thuộc vào giá trị cụ thể của tham số thật.

5.2 Tổng biến động

5.2.1 Khoảng cách

Định nghĩa 5.1. tổng biến động (total variation distance) giữa hai độ đo xác suất $\mathbb{P}_\theta$ và $\mathbb{P}_\eta$ là

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_\eta) = \max_{A\subset E} |\, \mathbb{P}_\theta(A) - \mathbb{P}_\eta(A) \,|.$$

Định lý 5.1 (Công thức tính). *Nếu tập mẫu E là rời rạc (discrete: countable or finite), xác suất $\mathbb{P}_\theta, \mathbb{P}_\eta$ có hàm khối lần lượt là p_θ, p_η thì*

$$TV(\mathbb{P}_\theta, \mathbb{P}_\eta) = \frac{1}{2} \sum_{x \in E} | \, p_\theta(x) - p_\eta(x) \, | \, .$$

Nếu tập mẫu E là liên tục (continuous), xác suất $\mathbb{P}_\theta, \mathbb{P}_\eta$ có mật độ lần lượt là f_θ, f_η thì

$$TV(\mathbb{P}_\theta, \mathbb{P}_\eta) = \frac{1}{2} \int_E | \, f_\theta(x) - f_\eta(x) \, | \, dx.$$

5.3 Phân kỳ KL

5.3.1 Phân kỳ KL

Định nghĩa 5.2. (KL divergence) Ký hiệu f là mật độ hoặc hàm khối xác suất:

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_\eta) := \mathbb{E}_\theta \left[\ln \frac{f_\theta(x)}{f_\eta(x)} \right] = \mathbb{E}_\theta \left[\ln f_\theta(x) \right] - \mathbb{E}_\theta \left[\ln f_\eta(x) \right] .$$

5.3.2 Đặc điểm

Proposition 5.1. *Phân kỳ KL thỏa mãn 2/4 đặc điểm của "khoảng cách":*

1. $KL(\mathbb{P}_\theta, \mathbb{P}_\eta) \geq 0$
2. $KL(\mathbb{P}_\theta, \mathbb{P}_\eta) \equiv 0 \iff \mathbb{P}_\theta \equiv \mathbb{P}_\eta$.

5.4 MLE

Ước lượng hợp lý cực đại (Maximum Likelihood Estimation)

5.4.1 Likelihood

Định nghĩa 5.3.

$$L_\theta(x_1, \dots, x_n) := \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} p_\theta(X_i = x_i).$$

5.4.2 Log likelihood

Định nghĩa 5.4.

$$\ell_\theta(x_1, \dots, x_n) := \ln L_\theta(x_1, \dots, x_n) = \sum_{i=1}^{n} \ln p_\theta(X_i = x_i).$$

5.4.3 Maximum Likelihood Estimator

Định nghĩa 5.5.

$$\hat{\theta}_n := \underset{\theta}{\operatorname{argmax}}\, L_\theta(x_1, \dots, x_n) \equiv \underset{\theta}{\operatorname{argmax}}\, \ell_\theta(x_1, \dots, x_n).$$

5.5 Mix model

Định nghĩa 5.6. Cho các mô hình gốc $X^{(k)}, k = 1, \dots, K$, lấy biến tiềm ẩn Z trên $\{1, \dots, K\}$ làm trọng số, ta có mô hình hỗn hợp

$$X = \sum_{k=1}^{K} \mathbb{P}(Z = k) X^{(k)}.$$

5.5.1 Giải thuật EM

Định nghĩa 5.7. (**E**stimation **M**aximization) *có thể* tìm được tham số θ của mô hình hỗn hợp Định nghĩa 5.6.

Giả sử ta quan sát được $X_1 = x_1, \ldots, X_n = x_n$. Gọi các trọng số tiềm ẩn tương ứng là $Z_1 = z_1, \ldots, Z_n = z_n$.

Sau khi khởi tạo $\theta = \theta_0$ ngẫu nhiên, ta lặp lại 2 bước **E, M** như sau để cập nhật $\theta_k, k = 1, 2, \ldots$ cho đến khi hội tụ.

- **E**stimate: Ước lượng $Z_i \approx \omega_i := \mathbb{E}[Z|X_i = x_i, \theta = \theta_{k-1}], i = 1, \ldots, n$.
- **M**aximize: Thay Z_i bởi ω_i vào công thức likelihood để tối ưu MLE $\theta = \theta_k$

5.6 Chuẩn tính của MLE

Định nghĩa 5.8. Giả sử log likelihood đối với một quan sát X theo mô hình θ là $\ell(\theta) = \ln L_1(X, \theta), \theta \in \Theta \subset \mathbb{R}$. Giả sử $\ell(\theta)$ có đạo hàm bậc hai. Dưới một số điều kiện chuẩn, **thông tin Fisher** của mô hình được định nghĩa là

$$I(\theta) = \mathbb{V}[\ell'(\theta)] = \mathbb{E}[(\ell'(\theta))^2] = -\mathbb{E}[\ell''(\theta)].$$

Định lý 5.2. *Gọi $\theta^* \in \Theta$ là tham số* thật *cần tìm. Giả sử*

1. *Các tham số là indentifiable*
2. *Support của $\mathbb{P}_\theta$ không phụ thuộc vào θ với mọi $\theta \in \Theta$*
3. *θ^* không nằm trên biên giới của Θ*
4. *Thông tin Fisher $I(\theta) \neq 0$ xung quanh θ^**
5. *Một số điều kiện kỹ thuật khác*

Khi đó, chuỗi $\hat{\theta}_n^{MLE}$ thỏa mãn:

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta^*}} \theta^*$$

$$\sqrt{nI(\theta^*)}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)\ w.r.t.\mathbb{P}_{\theta^*}} \mathcal{N}(0,1).$$

Chú ý là điều kiện số 2 dễ bị vi phạm, ví dụ $X_i \overset{iid}{\sim} Unif[0,\theta]$ mà lại cần tìm tham số θ.

5.7 M-estimatior

Định nghĩa 5.9. Với mục tiêu ước lượng thuộc tính μ^* của xác suất $\mathbb{P}(X)$, ta tìm một "hàm tổn thất" $\rho(X,\mu)$ có giá trị kỳ vọng đạt cực tiểu tại $\mu = \mu^*$:

$$\mathcal{Q}(\mu) := \mathbb{E}_{\mathbb{P}}[\rho(X,\mu)].$$

Nếu quan sát được $X_1,\dots,X_n \overset{iid}{\sim} \mathbb{P}(X)$, ta ước lượng

$$\mathcal{Q}(\mu) \approx \mathcal{Q}_n(\mu) := \frac{1}{n}\sum_{i=1}^{n}\rho(X_i,\mu).$$

Khi đó $\mu^* \approx \hat{\mu}$ với "M-estimator" $\hat{\mu}$ là

$$\hat{\mu} := \underset{\mu}{\operatorname{argmin}}\, \mathcal{Q}_n(\mu).$$

Ví dụ,

- với $\rho(x,\theta) = -\ln p_\theta(x)$ ta có MLE để ước lượng tham số θ^* của mô hình $\mathbb{P}$
- với $\boldsymbol{x},\boldsymbol{\mu} \in \mathbb{R}^d$, dùng $\rho(\boldsymbol{x},\boldsymbol{\mu}) = \|\boldsymbol{x}-\boldsymbol{\mu}\|^2$ ta ước lượng được $\boldsymbol{\mu}^* = \mathbb{E}[\boldsymbol{x}]$.

6 Kiểm định

Hypothesis testing

- Định nghĩa giả thuyết không và giả thuyết đối.
- Thiết kế kiểm định thống kê
- Phân loại lỗi Loại 1 và Loại 2
- Tính hàm công suất của kiểm định
- Định mức đối với kiểm định
- Tính toán và giải thích giá trị p

6.1 Giả thuyết không và đối

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$, sử dụng bộ mẫu dữ liệu iid $X_1, \dots, X_n$, ta xem xét hai giả thuyết về tham số θ như sau:

$$\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$$

với Θ_0, Θ_1 là phân mảnh (không giao nhau) của Θ, Θ_0 là "thường thức" (status quo), còn Θ_1 là "phát hiện" (discovery) mới. Ta gọi H_0 là giả thuyết không, còn H_1 là giả thuyết đối (thay thế).

6.2 Kiểm định

Ta sẽ **kiểm định** H_0 đối với H_1 bằng cách chọn và sử dụng một định lượng thống kê $\psi(X_1, \dots, X_n) \in \{0, 1\}$.

	$\psi = 0$: chấp nhận H_0	$\psi = 1$: phủ nhận H_0
$\theta \in \Theta_0$	Kiểm định đúng	Lỗi loại 1
$\theta \in \Theta_1$	Lỗi loại 2	Kiểm định đúng

Có thể viết $\psi = \mathbb{1}\{R_\psi\}$ với sự kiện R_ψ là **vùng phủ nhận**, còn R_ψ^c là **vùng chấp nhận** .

Ta thiết kế kiểm định sao cho hàm **công suất** sau đây có giá trị nhỏ khi $\theta \in \Theta_0$ và lớn khi $\theta \in \Theta_1$:

$$\beta_\psi(\theta) := \mathbb{P}_\theta(\psi = 1) \equiv \mathbb{P}_\theta(R_\psi) \in [0, 1]$$

Vùng phủ nhận thường có dạng

$$R_\psi = \{X_i : T(X_i) \geq c\}$$

với T là một lượng thống kê còn c là một giá trị biên.

6.3 Mức độ lỗi

Kiểm định ψ là ở mức (significance level) $\alpha \in (0, 1)$ nếu có xác suất lỗi loại 1 không vượt quá α :

$$\sup_{\theta \in \Theta_0} \beta_\psi(\theta) \leq \alpha.$$

Chuỗi kiểm định $(\psi_n)_{n=1,2,\ldots}$ được gọi là tiệm cận về mức α nếu

$$\lim_{n \to \infty} \sup_{\theta \in \Theta_0} \beta_{\psi_n}(\theta) \leq \alpha.$$

Phương thức Neyman-Pearson chọn một mức α, đảm bảo xác suất lỗi loại 1 không vượt quá α rồi tối thiểu hóa xác suất lỗi loại 2. Nói cách khác là giữ cho công suất $\beta_\psi(\theta)$ đủ nhỏ khi $\psi \in \Theta_0$, rồi tối đại hóa công suất khi $\psi \in \Theta_1$.

6.3.1 p-value

Từ quan sát $X_1, \ldots, X_n$ ta tính giá trị mức α (tiệm cận) nhỏ nhất tại đó kiểm định ψ phủ nhận H_0, gọi nó là p-value (tiệm cận) của ψ. Nếu p-value càng nhỏ thì ta càng tự tin phủ nhận H_0.

$$\text{p-value} := \inf_{X_1, \ldots, X_n; \theta \in H_0} \beta_\psi(\theta)$$

p-value	chứng cứ phủ nhận H_0
$< 0.1\%$	vô cùng mạnh
0.1%—1%	rất mạnh

p-value	chứng cứ phủ nhận H_0
1%—5%	mạnh
5%—10%	yếu
> 10%	không có

6.4 Khoảng tin cậy

Thông thường ta có thể xây dựng được kiểm định từ khoảng tin cậy. Ví dụ, ta muốn kiếm định tham số θ, $H_0 : \theta = \theta_0$, đối $H_1 : \theta \neq \theta_0$. Giả sử ta có khoảng tin cậy $\mathcal{I}$ ở mức $1 - \alpha$, tức là

$$\mathbb{P}_\theta(I \ni \theta) \geq 1 - \alpha.$$

Khi đó, $\psi = \mathbb{1}\{\theta_0 \notin \mathcal{I}\}$ là kiểm định mức α

$$\beta_\psi(\theta_0) = \mathbb{P}_{\theta_0}(\theta_0 \notin I) \leq \alpha.$$

6.5 Wald Test

Giả sử $\hat{\theta}$ là ước lượng của tham số θ, và $\hat{\mathbb{V}}[\hat{\theta}]$ là ước lượng phương sai của $\hat{\theta}$, sao cho

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\mathbb{V}}[\hat{\theta}]}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1).$$

Đặt

$$W := \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\mathbb{V}}[\hat{\theta}]}},$$

ta có thể xây dựng các kiểm định Wald có mức tiệm cận là α, tức là $\mathbb{P}_{H_0}(\psi_\alpha) \xrightarrow[n\to\infty]{} \alpha$.

Giả thuyết	Kiểm định Wald	asymp. p-value
$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$	$\psi_\alpha = \mathbb{1}\{\lvert W\rvert > q_{\alpha/2}\}$	$\mathbb{P}(\lvert Z\rvert > \lvert W^{obs}\rvert)$
$H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$	$\psi_\alpha = \mathbb{1}\{W > q_\alpha\}$	$\mathbb{P}(Z > W^{obs})$

Giả thuyết	Kiểm định Wald	asymp. p-value
$H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$	$\psi_\alpha = \mathbb{1}\{W < -q_\alpha\}$	$\mathbb{P}(Z < W^{obs})$

Trong bảng trên, p-value được tính từ W^{obs} là một quan sát đối với W.

6.6 Định lý Cochran

Định lý 6.1. *Giả sử $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Đặt*

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Khi đó $\mathbb{E}[S_n^2] \equiv \sigma^2$,

$$\frac{S_n^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1},$$

và $\bar{X}_n, S_n^2$ độc lập với nhau.

6.7 Student's T test

Giả sử $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, μ, σ chưa biết, và ta muốn kiểm định μ. Theo Định lý 6.1,

$$T := \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \equiv \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{\sqrt{S_n^2/\sigma^2}} \sim t_{n-1}$$

tuân theo phân phối Student's T. Ta có thể xây dựng các kiểm định Student có mức α, tức là $\mathbb{P}_{H_0}(\psi_\alpha) \equiv \alpha$.

Giả thuyết	Kiểm định Student	p-value
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	$\psi_\alpha = \mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(\lvert Z\rvert > \lvert T^{obs}\rvert)$
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$\psi_\alpha = \mathbb{1}\{T > q_\alpha^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(Z > T^{obs})$
$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	$\psi_\alpha = \mathbb{1}\{T < -q_\alpha^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(Z < T^{obs})$

Trong bảng trên, p-value được tính từ T^{obs} là một quan sát đối với T.

6.8 Two-sample T-test

Giả sử $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_1, \dots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y^2)$, với $\mu_x, \sigma_x, \mu_y, \sigma_y$ chưa biết, và ta muốn kiểm định $\mu_x - \mu_y$. Đặt

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2,$$

$$S_m^2 := \frac{1}{m-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_m)^2,$$

Ta có gần đúng

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{S_n^2/n + S_m^2/m}} \sim t_N$$

là phân phối Student's T với độ tự do tuân theo công thức WS (Welch-Satterthwaite):

$$N = \frac{\sqrt{S_n^2/n + S_m^2/m}}{\sqrt{\frac{S_n^4}{n^2(n-1)} + \frac{S_m^4}{m^2(m-1)}}} \geq \min(n, m)$$

Đặt

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_n^2/n + S_m^2/m}}$$

Giả thuyết	Kiểm định 2 mẫu	p-value
$H_0 : \mu_x = \mu_y, H_1 : \mu_x \neq \mu_y$	$\psi_\alpha = \mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_N}\}$	$\mathbb{P}_{t_N}(\lvert Z\rvert > \lvert T^{obs}\rvert)$
$H_0 : \mu_x \leq \mu_y, H_1 : \mu_x > \mu_y$	$\psi_\alpha = \mathbb{1}\{T > q_\alpha^{t_N}\}$	$\mathbb{P}_{t_N}(Z > T^{obs})$
$H_0 : \mu_x \geq \mu_y, H_1 : \mu_x < \mu_y$	$\psi_\alpha = \mathbb{1}\{T < -q_\alpha^{t_N}\}$	$\mathbb{P}_{t_N}(Z < T^{obs})$

6.9 Kiểm định tỷ lệ hợp lý

Likelihood ratio test

Với dữ liệu $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$, để kiểm định $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$, định lượng tỷ lệ hợp lý là

$$T_n = 2\ln\frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)}$$

với $L(\hat{\theta})$ là hợp lý cực đại tổng quát, còn $L(\hat{\theta}_0)$ là hợp lý cực đại khi H_0 đúng.

6.9.1 Định lý Wilks

Giả sử $\theta = \left(\theta_1, \ldots, \theta_{q+r}\right) \in \Theta \subset \mathbb{R}^{q+r}$,

$$\Theta_0 = \left\{\theta \in \Theta \,:\, \left(\theta_{q+1}, \ldots, \theta_{q+r}\right) = \left(\theta_{q+1}^{(0)}, \ldots, \theta_{q+r}^{(0)}\right)\right\}$$

với $\left(\theta_{q+1}^{(0)}, \ldots, \theta_r^{(0)}\right) \in \mathbb{R}^r$ là cố định. Nếu H_0 đúng và các điều kiện hội tụ của MLE (Định lý 5.2) được thỏa mãn thì:

$$T_n \xrightarrow[n\to\infty]{(d)} \chi_r^2.$$

6.10 Kiểm định nhiều lần

Gọi số lần thực hiện và quan sát kết quả kiểm định là t, mức độ lỗi loại 1 cho phép là α. Nếu H_0 là đúng thì lỗi loại 1 sẽ xuất hiện khoảng αt lần, số lần này sẽ càng lớn nếu t càng lớn.

Gọi tỷ lệ p-value không vượt quá ngưỡng α là $F(\alpha) := \mathbb{P}(\text{p-value} \leq \alpha)$. Ta có $F(\alpha) \leq \alpha$ với mọi loại kiểm định. Hơn nữa, $F(\alpha) = \alpha$ với kiểm định Student's T.

Giáo trình

Tham khảo

John Tsitsiklis, Dimitri Bertsekas, Partick Jaillet. 2022. "Probability - The Science of Uncertainty and Data". MITx. https://www.edx.org/course/probability-the-science-of-uncertainty-and-data.

Philippe Rigollet, Tyler Maunu, Jan Christian Huetter. 2022. "Fundamentals of Statistics". MITx. https://www.edx.org/course/fundamentals-of-statistics.

Wasserman, Larry. 2004. *All of statistics : a concise course in statistical inference*. New York: Springer. https://archive.org/details/springer_10.1007-978-0-387-21736-9.

MITx 18.6501x

"Fundamentals of Statistics" (MITx 18.6501x) là khóa học của Philippe Rigollet (2022) đại học MIT dạy trên edX.

Quy định

Kỳ hạn

- Exercises and homework: Wednesdays 11:59AM UTC (Wed. 20:59 JST)
- Exams (48 hours): Tuesdays 11:59AM UTC(Tue. 20:59 JST)

Thời gian cần thiết

Mỗi tuần khoảng hơn 12 tiếng

- 5-7 hours on exercises, including 3 hours of lecture clips
- 1-2 hours watching recitations
- 5-7 hours for weekly problem sets

Tính điểm

Điểm thi đậu chứng chỉ là 60% tổng số điểm tối đa.

- 20% for the lecture exercises (divided equally among the 20 out of 23 lectures)
- 20% for the homeworks (divided equally among 10 (out of 12) homeworks)
- 18% for the first midterm exam (timed)
- 18% for the second midterm exam (timed)
- 24% for the final exam (timed)

Unit 1

Ôn lại kiến thức về xác suất "Probability - The Science of Uncertainty and Data" by John Tsitsiklis (2022).

Unit 2

Học về nền tảng của suy luận (Foundations of Inference).

- Lecture 3: Mô hình xác suất có số biến hữu hạn (Parametric Statistical Models)
- Lecture 4: Dự đoán với số biến hữu hạn và vùng tự tin (Parametric Estimation and Confidence Intervals)
- Lecture 5: Vùng tự tin và phương pháp Delta (Confidence Intervals and Delta Method)

Unit 3

Phương pháp ước lượng tham số của mô hình xác suất

- Lecture 6: Khoảng cách giữa các phân phối xác suất
- Lecture 7: Maximum Likelihood Estimator
- Lecture 8: Ví dụ cho Maximum Likelihood Estimator
- Lecture 9: Các tính chất thống kê của Maximum Likelihood Estimator
- Lecture 10: Phương pháp tích suất và M-Estimation

Unit 4

Kiểm định thống kê

- Lecture 11: Introduction to Parametric Hypothesis Testing
- Lecture 12: The Wald Test and Likelihood Ratio Test
- Lecture 13: The T-test
- Lecture 14: Multiple Hypothesis Testing