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. Fig 0.1 minh hoạ quan hệ giữa "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 xstk và có thể tải PDF.

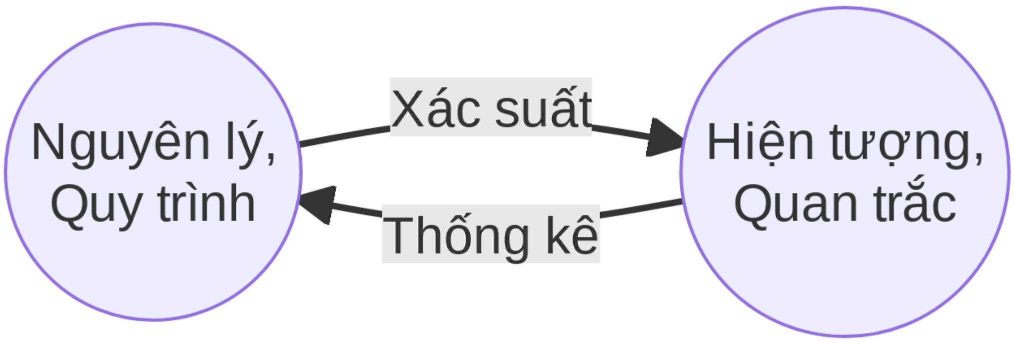

Fig 0.1: Xác suất và thống kê.

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 $A_1, A_2, \ldots$ 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 $A_1, A_2, \ldots$ 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 $A_1, A_2, \ldots$ 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.

Def 1.1 (Biến ngẫu nhiên)**.** ("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 $X(\omega)$.

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

Def 1.3 (Độc lập)**.** 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 $X \leq x$ và $Y \leq y$ là độc lập đối với mọi x, y.

Def 1.4 (Xác suất hợp)**.** 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.

Def 1.5 (Xác suất có điều kiện)**.** 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,

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

Thm 1.1 (Định lý Baye)**.**

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

1.3 Tích suất

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

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

Def 1.6 (Trung bình)**.** là tích suất bậc 1, tức là $\mathbb{E}[X]$.

Ta có

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

Def 1.7 (Hiệp phương sai)**.**

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

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

Def 1.8 (Phương sai)**.** Phương sai, hay phân tán (variance,) là

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

Ta có

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

Def 1.9 (Hàm tạo tích suất)**.** "Moment geenration function (MGF)" hoặc "Laplace transform" của biến X là

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

với $t \in \mathbb{R}$.

Nếu MGF là "well defined" trên một khoảng mở xung quanh 0 thì đạo hàm cấp k của ψ tại 0 bằng đúng $\mathbb{E}[X^k]$:

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

1.4 Phân phối

Có một số phân phối xác suất thông dụng.

Def 1.10 (IID)**.** 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

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

Def 1.11 (Phân phối Bernoulli)**.** $X \sim \mathrm{Ber}(p)$ có

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

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

Def 1.12 (Phân phối Binomial)**.** $X \sim \mathrm{Bin}(n, p)$ với $n \in \mathbb{Z}_+, p \in (0, 1)$ mô tả tổng số lần thành công của n thí nghiệm độc lập $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Ber}(p)$. Ta có

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

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

Def 1.13 (Phân phối Poisson)**.** $X \sim \mathrm{Poi}(\lambda)$ 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.

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

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

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

Def 1.14 (Phân phối Geometric)**.** $X \sim \text{Geom}(p), p \in (0,1)$ có

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

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

Def 1.15 (Phân phối Exponential)**.** $X \sim \text{Exp}(\lambda)$ 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 $\mathbb{E}[X] = 1/\lambda, \mathbb{V}[X] = 1/\lambda^2$.

Def 1.16 (Phân phối Gaussian)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$ 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 $\mathbb{E}[X] = \mu, \mathbb{V}[X] = \sigma^2$.

1.5 Hội tụ

Có một số kiểu hội tụ của biến ngẫu nhiên.

Def 1.17 (Hội tụ xác suất)**.** 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

$$\sqrt{n}(X_n - X) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2).$$

Thm 1.2 (Độ mạnh)**.** *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* $\mathbb{P}(a \le X_n \le b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \le X \le b)$ *với mọi khoảng* $[a, b]$*.*

Thm 1.3 (Tổng và tích)**.** *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 và tích của chúng cũng hội tụ theo cùng kiểu (gần tuyệt đối, hoặc theo xác suất) về tổng* $X + Y$ *hoặc tích* XY *tương ứng.*

Thm 1.4 (Slutsky)**.** *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.*

Thm 1.5 (Ánh xạ liên tục)**.** *(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.*

Thm 1.6 (Đại đa số (Law of Large Numbers))**.** *Cho n biến iid $X_1, X_2, ., X_n$ có chung $\mathbb{E}[X_i] = \mu < \infty$. Khi đó:*

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

Thm 1.7 (Hội tụ trung tâm (Central Limit))**.** *Giả sử thêm là $\mathbb{V}[X_i] = \sigma^2 < \infty$. Khi đó*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2).$$

Thm 1.8 (Bất đẳng thức Hoefding)**.** *Nếu có một khoảng cố định $[a, b]$ gần như tuyệt đối (almost surely) chứa các biến $X_i (i = 1, 2, \ldots, n)$ thì*

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

Thm 1.9 (Phương pháp Delta)**.** *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,*

$$\sqrt{n}(g(\theta_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, (g'(\theta))^2\sigma^2)$$

2 Thống kê

2.1 Mô hình

Trên không gian đo được $E \subseteq \mathbb{R}$ ta quan sát các mẫu $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}$. 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 cpecified).

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

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

2.2 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 (Def 1.17).

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

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

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

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

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

2.3 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.

2.4 Kiểm định

3 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 $X_1, \dots, X_n$ trên tập mẫu E và bộ tham số Θ . Ngầm định tồn tại *tham số thật* $\theta^* \in \Theta$ để $X_1 \sim \mathbb{P}_{\theta^*}$.

3.1 Tổng biến động

Def 3.1 (Khoảng cách)**.** tổng biến động (total variation distance) giữa hai độ đo xác suất $\mathbb{P}_\theta$ và $\mathbb{P}_\eta$ là

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

Thm 3.1 (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

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

3.2 Phân kỳ KL

Def 3.2 (Phân kỳ KL)**.** (KL divergence) Ký hiệu f là mật độ hoặc hàm khối xác suất:

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

Proposition 3.1 (Đặc điểm)**.** *Phân kỳ KL thỏa mãn 2/4 đặc điểm của "khoảng cách":*

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

3.3 MLE

Maximum Likelihood Estimation

Def 3.3 (Likelihood)**.**

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

Def 3.4 (Log likelihood)**.**

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

Def 3.5 (Maximum Likelihood Estimator)**.**

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

Def 3.6 (Mix model)**.** 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

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

Thm 3.2 (Giải thuật EM)**.** *(Estimation Maximization) để tìm các tham số của mô hình hỗn hợp. Nhập liệu là các quan sát $X_1, \dots, X_n$. Gọi các trọng số tiềm ẩn là $Z_1, \dots, Z_n$. Sau khi khởi tạo các tham số bằng giá trị ngẫu nhiên, ta lặp lại 2 bước E, M như sau để cập nhật các tham số cho đến khi chúng hội tụ.*

- *Bước E: Ước lượng $Z_i \approx \omega_i := \mathbb{E}[Z|X_i], i = 1, \dots, n$.*
- *Bước M: Thay Z_i bởi ω_i vào log likelihood để tìm MLE cho X.*

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 khoảng tự tin (Parametric Estimation and Confidence Intervals)