Xác Suất Thống Kê

Võ Chí Công

2022/8/25

Table of contents

Lời nói đầu

"Xác suất" và "thống kê" là hai môn học gây cho tôi nhiều khó khăn nhất. 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 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.

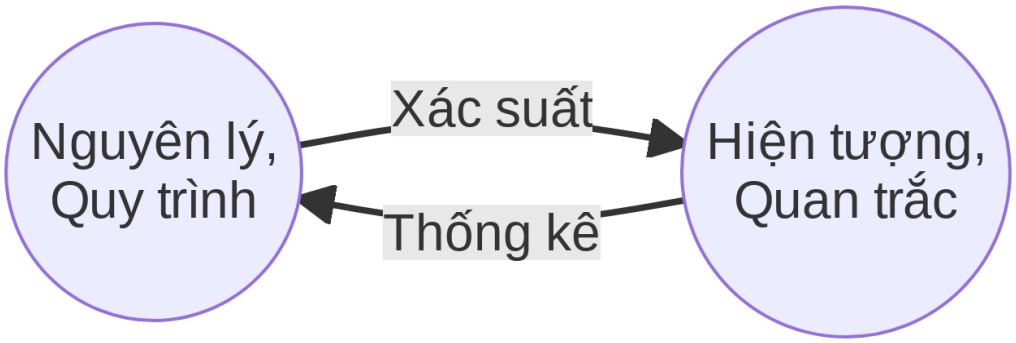

Figure 1: Xác suất và thống kê.

1 MITx 18.6501x

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

- Important dates
- Calendar (private)

1.1 Unit 0

1.1.1 Kỳ hạn nộp bài

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

1.1.2 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

1.1.3 Cách 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)

2 Xác suất

2.1 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 $A_1, A_2, ...$ 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 $A_1, A_2, ...$ là một cách **phân hoạch** sự kiện C.

2.2 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 $A_1, A_2, ...$ 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.

Định nghĩa 2.1 (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 $X(\omega)$.

Định nghĩa 2.2 (Điểm cắt)**.** Điểm cắt tại mức $1 - \alpha$ của biến X là một số q_α mà $\mathbb{P}(X \leq q_\alpha) = 1 - \alpha$.

2.3 Tích suất

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

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

Định nghĩa 2.3 (Trung bình)**.** là tích suất hạng 1, tức là $\mathbb{E}[X]$.

Định nghĩa 2.4 (Phương sai)**.**

$$\mathbb{V}[X] := \mathbb{E}[X^2] - (E[X])^2.$$

2.4 Phân phối

Có một số phân phối xác suất thông dụng.

Định nghĩa 2.5 (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.

Định nghĩa 2.6 (Phân phối Bernoulli)**.** $X \sim \text{Ber}(p)$ có

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

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

Định nghĩa 2.7 (Phân phối Binomial)**.** $X \sim \text{Bin}(n, p)$ với $n \in \mathbb{Z}_+, p \in (0, 1)$ mô tả tổng số lần thành công của n thí nghiệm $\text{Ber}(p)$ độc lập.

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

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

Định nghĩa 2.8 (Phân phối Poisson)**.** $X \sim \text{Poi}(\lambda)$ 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. Có mật độ

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

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

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

Định nghĩa 2.9 (Phân phối Geometric)**.** $X \sim \text{Geom}(p), p \in (0, 1)$ có mật độ

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

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

Định nghĩa 2.10 (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 mật độ

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

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

Định nghĩa 2.11 (Phân phối Gaussian)**.** $X \sim \mathcal{N}(\mu, \sigma^2)$ có mật độ

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

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

2.5 Hội tụ

Có một số kiểu hội tụ của biến ngẫu nhiên.

Định nghĩa 2.12 (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:

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

   tại mọi điểm x mà cdf của X liên tục.

2.5.1 Tính chất

Xếp theo thứ tự từ mạnh đến yếu:

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

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 $[a, b]$.

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.

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 (định lý Slutsky).

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

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

Cho n biến iid $X_1, X_2, .., X_n$ có chung $\mu = \mathbb{E}[X_i], \sigma^2 = \mathbb{V}[X_i]$. Lấy giá trị trung bình:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Định lý 2.1 (Định lý số lớn (LLN))**.**

$$\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P},\, a.s.} \mu$$

Định lý 2.2 (Định lý trung tâm (CLT))**.**

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

Định lý 2.3 (Bất đẳng thức Hoefding)**.** *Nếu có một khoảng cố định $[a, b]$ gần như tuyệt đối (almost surely) chứa các biến $X_i (i = 1, 2, \dots, n)$ thì*

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

3 Tóm tắt

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.

Trích dẫn

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.