Math for Machine Learning

Quy Anh Dang, Hanoi University of Science

Tháng 10, 2022

Mục lục

1 Lời nói đầu **4**
1.1 Vai trò của toán học 4
1.2 Mục đích viết tài liệu 4

2 Ký hiệu **5**

3 Bảng thuật ngữ **6**

4 Bảng thuật ngữ viết tắt **7**

5 Đại số tuyến tính **8**
5.1 Không gian vectơ 8
5.1.1 Không gian Euclid 8
5.1.2 Không gian con 9
5.2 Không gian metric 9
5.3 Không gian định chuẩn 9
5.4 Không gian tích trong 10
5.4.1 Định lý Pytago 11
5.4.2 Bất đẳng thức Cauchy-Schwarz 11
5.5 Chuyển vị 11
5.6 Giá trị riêng và vectơ riêng 12
5.7 Vết 12
5.8 Định thức 13
5.9 Ma trận trực giao 13
5.10 Ma trận đối xứng 13
5.10.1 Thương Rayleigh 14
5.11 Ma trận xác định dương (Bán xác định dương) 15

5.11.1 Hình học của các dạng bậc hai xác định dương . . . 16
5.12 Phân tích giá trị suy biến . . . 16
5.13 Một số nhận dạng hữu ích về ma trận . . . 17
5.13.1 Tích ma trận - vectơ như là một tổ hợp tuyến tính các cột của ma trận . . . 17
5.13.2 Tổng các tích ngoài như là tích của hai ma trận . . . 17
5.13.3 Dạng bậc hai . . . 18

6 Giải tích và Tối ưu hóa **19**
6.1 Cực trị . . . 19
6.2 Gradients . . . 19
6.3 Jacobian . . . 19
6.4 Hessian . . . 20
6.5 Giải tích ma trận . . . 20
6.5.1 Quy tắc chuỗi . . . 20
6.6 Định lý Taylor . . . 21
6.7 Điều kiện cực tiểu địa phương . . . 21
6.8 Tính lồi . . . 22
6.8.1 Tập lồi . . . 23
6.8.2 Khái niệm cơ bản về hàm lồi . . . 23
6.8.3 Một số hệ quả của tính lồi . . . 24
6.8.4 Chứng minh một hàm là hàm lồi . . . 25
6.8.5 Một số ví dụ . . . 27

7 Xác suất **29**
7.1 Khái niệm cơ bản . . . 29
7.1.1 Xác suất có điều kiện . . . 30
7.1.2 Quy tắc chuỗi . . . 30
7.1.3 Quy tắc Bayes . . . 31
7.2 Biến ngẫu nhiên . . . 31
7.2.1 Hàm phân phối tích lũy . . . 31
7.2.2 Các biến ngẫu nhiên rời rạc . . . 32
7.2.3 Biến ngẫu nhiên liên tục . . . 32
7.3 Phân phối đồng thời . . . 33
7.3.1 Tính độc lập của các biến ngẫu nhiên . . . 33
7.3.2 Phân phối biên . . . 33
7.4 Kỳ vọng . . . 33
7.4.1 Tính chất của giá trị kỳ vọng . . . 34

7.5 Phương sai 34
7.5.1 Tính chất của phương sai 34
7.5.2 Độ lệch chuẩn 35
7.6 Hiệp phương sai 35
7.6.1 Tương quan 35
7.7 Vectơ ngẫu nhiên 35
7.8 Ước lượng các tham số 36
7.8.1 Ước lượng hợp lý cực đại 36
7.8.2 Ước lượng hậu nghiệm cực đại 37
7.9 Phân phối Gauss (phân phối chuẩn) 37
7.9.1 Hình học của phân phối Gauss đa biến 38

Tài liệu **39**

1 Lời nói đầu

1.1 Vai trò của toán học

Bạn đọc thân mến, trong những năm gần đây, trí tuệ nhân tạo - AI nói chung và Machine Learning, Deep Learning nói riêng nổi lên như một minh chứng cho sự thành công của cuộc cách mạng khoa học công nghiệp lần thứ 4. Không thể phủ nhận rằng những ứng dụng trí tuệ nhân tạo đang hỗ trợ con người trong nhiều lĩnh vực khác nhau, góp phần nâng cao năng suất lao động và tạo ra những lợi ích trực tiếp cho con người nói chung. Có rất nhiều ứng dụng AI quan trọng mà có thể bạn đang sử dụng hàng ngày nhưng vô tình không nhận ra. Một số chúng đó là: Trợ lý ảo Alexa của Amazon, xe tự hành của Tesla, hệ thống dịch máy của Google, hệ thống nhận diện khuôn mặt của Apple, các mô hình sáng tác âm nhạc, nghệ thuật, thơ ca như MuseNet, Stable Diffusion hay GPT-3.

Sự phát triển mạnh mẽ của AI trong giai đoạn vừa qua là nhờ sự đóng góp rất lớn của các nhà nghiên cứu đã tạo ra các lý thuyết mới về mô hình, thuật toán, phương pháp tối ưu,... trong Machine Learning và Deep Learning. Những công trình này đều dựa trên nền tảng toán học chặt chẽ. Chính vì thế, để đọc hiểu các công trình nghiên cứu cũng như tài liệu trong lĩnh vực AI thì củng cố kiến thức toán là một điều rất cần thiết.

1.2 Mục đích viết tài liệu

Hiện nay, nhu cầu tuyển dụng và học tập AI tại Việt Nam rất lớn. Tuy nhiên, chúng ta thường gặp khó khăn khi học các lý thuyết mô hình do chưa vững kiến thức cơ bản về toán. Tôi viết tài liệu này trước tiên nhằm hệ thống lại kiến thức cho bản thân mình. Sau đó là tạo ra một tài liệu toàn diện, đáng tin cậy cho mọi người. Để thuận tiện cho quá trình học, các kiến thức được sắp xếp theo từng chủ đề một cách logic nhằm tạo ra liên kết kiến thức. Bên cạnh các chương phụ, các chương chính của sách sẽ xoay quanh các chủ đề quan trọng của toán trong AI như:

- Chương 5: Đại số tuyến tính.
- Chương 6: Giải tích và Tối ưu hóa.
- Chương 7: Xác suất

Toán học là một lĩnh vực được đánh giá là khô khan với nhiều bạn đọc. Đặc biệt là các xung đột thuật ngữ dẫn đến sự khó hiểu. Vì vậy tôi tạo ra một bảng thuật ngữ gồm các từ gốc Tiếng Anh và từ dịch thuật Tiếng Việt (chương 3. Bảng thuật ngữ) cũng như các cụm từ viết tắt trong Tiếng Anh kèm theo nghĩa đầy đủ Tiếng Anh và ý nghĩa tham chiếu Tiếng Việt (chương 4. Bảng thuật ngữ viết tắt). Những thuật ngữ đã được nhóm thảo luận và tham khảo từ các tài liệu, giáo trình về Thống Kê, Machine Learning tại các trường Đại Học đầu ngành tại Việt Nam. Đối với bạn đọc chưa quen thuộc với các kí hiệu cơ bản trong toán học, tôi cung cấp chương 2. Ký hiệu.

Để tạo ra một tài liệu chuẩn hóa và bao quát được kiến thức cần thiết, tôi đã dành thời gian khảo cứu từ các tài liệu về toán uy tín được trích dẫn đầy đủ tại mục Tài liệu. Trong đó tài liệu chính được tham khảo từ Garrett Thomas, "Mathematics for Machine Learning", University of California, Berkeley (2018) .

Trong quá trình biên soạn, có thể còn những thiếu sót. Vì vậy để tài liệu chỉn chu hơn nữa, tôi hoan nghênh những góp ý xây dựng từ các bạn gửi về địa chỉ math4machinelearning@gmail.com.

Thân ái!
Quý Anh

2 Ký hiệu

Ký hiệu	Ý nghĩa
$\mathbb{R}$	tập hợp số thực
$\mathbb{R}^n$	tập hợp (không gian vectơ) n chiều trên trường số thực, khép kín với tích vô hướng
$\mathbb{R}^{m\times n}$	tập hợp (không gian vectơ) của các ma trận cỡ $m \times n$ trên trường số thực
δ_{ij}	Kronecker delta, i.e. $\delta_{ij} = 1$ nếu $i = j$, 0 ngược lại
$\nabla f(\mathbf{x})$	gradient của hàm f tại $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$	Hessian của hàm f tại $\mathbf{x}$
$\mathbf{A}^\top$	chuyển vị của ma trận $\mathbf{A}$
Ω	không gian mẫu
$\mathbb{P}(A)$	xác suất của biến cố A
$p(X)$	phân phối của biến ngẫu nhiên X
$p(x)$	hàm mật độ (hàm khối) xác suất tại x
A^c	phần bù của biến cố A
$A \dot\cup B$	hợp của A và B, với điều kiện $A \cap B = \varnothing$
$\mathbb{E}[X]$	kì vọng (trung bình) của biến ngẫu nhiên X
$\mathrm{Var}(X)$	phương sai của biến ngẫu nhiên X
$\mathrm{Cov}(X, Y)$	hiệp phương sai của biến ngẫu nhiên X và Y

Một số lưu ý khác:

- Vectơ và ma trận là được in đậm (e.g. $\mathbf{x}, \mathbf{A}$). Điều này đúng với các vectơ trong không gian $\mathbb{R}^n$ cũng như với các vectơ trong các không gian vectơ nói chung. Đồng thời, các chữ cái Hy Lạp thường được sử dụng cho các đại lượng vô hướng và các chữ cái La Mã viết hoa cho các ma trận và các biến ngẫu nhiên.
- Ở nhiều chỗ trong tài liệu này, hoàn toàn có thể khái quát cho trường số phức, nhưng chúng ta sẽ chỉ nói về phiên bản áp dụng cho trường số thực.
- Chúng ta giả định rằng vectơ là vectơ cột, tức là một vectơ trong $\mathbb{R}^n$ có thể được xem như là một ma trận cỡ $n \times 1$. Do đó, việc lấy chuyển vị của một vectơ được xác định rõ ràng (và tạo ra một vectơ hàng, là một ma trận cỡ $1 \times n$).

3 Bảng thuật ngữ

Tiếng Anh	Tiếng Việt
complement	phần bù của một biến cố
convexity	tính lồi
convex set	tập lồi
cost function/objective function	hàm mất mát, hàm mục tiêu
countable additivity	cộng tính đếm được
covariance	hiệp phương sai
correlation	tương quan
cumulative distribution function	hàm phân phối tích lũy
eigenvalue	giá trị riêng
eigenvector	vectơ riêng
events	các biến cố trong xác suất
expected value	kì vọng, trung bình
extrema	điểm cực trị
gaussian/normal distribution	phân phối chuẩn
global minimum	điểm cực tiểu toàn cục
inner product space	không gian tích trong (không gian tích vô hướng)
joint distribution	phân phối xác suất đồng thời
level set	tập đồng mức
likelihood	khả năng xảy ra
linearly independent	độc lập tuyến tính
marginal distribution	phân phối xác suất biên
maximum likelihood estimation	ước lượng hợp lý cực đại
maximum a posteriori estimation	ước lượng hậu nghiệm cực đại
normed space	không gian định chuẩn
nullspace	không gian nghiệm (hay nhân) của phép biến đổi tuyến tính
orthogonal matrice	ma trận trực giao
orthonormal	trực chuẩn
posterior	xác suất hậu nghiệm
positive definite matrice	ma trận xác định dương
positive semi-definite matrice	ma trận bán xác định dương
probability mass function	hàm khối xác suất
probability density function	hàm mật độ xác suất
prior	xác suất tiên nghiệm
random variable	biến ngẫu nhiên
random vector	vectơ ngẫu nhiên
range	ảnh của phép biến đổi tuyến tính
Rayleigh quotient	thương số Rayleigh
sample space	không gian mẫu
scalar	số vô hướng
scale invariance	tính bất biến
singular value decomposition	phân tích suy biến (phân tích giá trị kỳ dị)
standard deviation	độ lệch chuẩn
stationary point	điểm dừng
symmetric matrice	ma trận đối xứng
variance	phương sai

4 Bảng thuật ngữ viết tắt

Cụm từ viết tắt	Thuật ngữ gốc	Ý nghĩa
cdf	cumulative distribution function	hàm phần phối tích lũy
pmf	probability mass function	hàm khối xác suất
pdf	probability density function	hàm mật độ xác suất
iid	independent and identically distributed	độc lập và xác định
MLE	maximum likelihood estimation	ước lượng hợp lý cực đại
MAP	maximum a posteriori probability	ước lượng hậu nghiệm tối đa

5 Đại số tuyến tính

Trong phần này, chúng ta sẽ làm quen với các lớp không gian quan trọng, mà trong đó dữ liệu sẽ tồn tại và các toán tử được thực thi, bao gồm: không gian vectơ, không gian metric, không gian định chuẩn và không gian tích trong. Nhìn chung, những khái niệm này vẫn được định nghĩa theo các tính chất của không gian Euclid nhưng theo một cách khái quát hơn.

5.1 Không gian vectơ

Không gian vectơ V là một tập hợp mà các phần tử trong không gian là các vectơ, được trang bị bởi hai phép toán là: phép cộng các vectơ và phép nhân vectơ với một số vô hướng. Các phép toàn này thỏa mãn những điều kiện sau

(i) $\mathbf{x} + \mathbf{0} = \mathbf{x}, \quad \forall \mathbf{x} \in V$

(ii) Tồn tại đối của của $\mathbf{x}$ là $-\mathbf{x}$ sao cho $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$

(iii) $1\mathbf{x} = \mathbf{x}, \quad \forall \mathbf{x} \in V$

(iv) Tính giao hoán: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in V$

(v) Tính kết hợp: $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ và $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ và $\alpha, \beta \in \mathbb{R}$

(vi) Tính phân phối: $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ và $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ với mọi $\mathbf{x}, \mathbf{y} \in V$ và $\alpha, \beta \in \mathbb{R}$

5.1.1 Không gian Euclid

Không gian **Euclid** là không gian vectơ thuần túy, được ký hiệu là $\mathbb{R}^n$. Mỗi vectơ trong không gian này là một nhóm gồm n số thực:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

Để thuận tiện về sau, ta có thể viết lại vectơ này dưới dạng ma trận cỡ $n \times 1$, hay **vectơ cột**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Phép cộng hai vectơ và nhân vectơ với một vô hướng $\mathbb{R}^n$:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}, \qquad \alpha\mathbf{x} = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$$

Không gian Euclid thường được sử dụng để biểu diễn một cách toán học các đại lượng trong không gian vật lý, với các khái niệm như khoảng cách, chiều dài và góc. Mặc dù rất khó để biểu diễn không gian trong trường hợp $n > 3$ chiều, nhưng những khái niệm này khái quát toán học theo nhiều cách hiển nhiên. Thậm chí khi bạn đang làm việc với một không gian tổng quát hơn $\mathbb{R}^n$ thì nó vẫn hữu ích để biểu diễn phép cộng vectơ và phép nhân với số vô hướng đối với vectơ 2D trong mặt phẳng hoặc vectơ 3D trong không gian.

5.1.2 Không gian con

Không gian vectơ có thể chứa các không gian vectơ khác. Nếu V là một không gian vectơ , thì $S \subseteq V$ được gọi là **không gian con** của V nếu:

(i) $\mathbf{0} \in S$

(ii) S là khép kín với phép cộng vectơ : $\mathbf{x}, \mathbf{y} \in S$ thì $\mathbf{x} + \mathbf{y} \in S$

(iii) S là khép kín phép nhân vectơ với một số vô hướng : $\mathbf{x} \in S, \alpha \in \mathbb{R}$ thì $\alpha\mathbf{x} \in S$

Chú ý rằng V luôn là không không gian con của chính V vì luôn thỏa mãn ba tính chất đã nêu trên. Một ví dụ củng cố vững chắc về không gian vectơ con, một đường thẳng đi qua gốc tọa độ là một không gian con của không gian Euclid.

Ngoài ra, còn có một số không gian con quan trọng khác cảm sinh từ phép biến đổi tuyến tính. Cụ thể, nếu $T : V \to W$ là một phép biến đổi tuyến tính, ta định nghĩa **nullspace** (không gian nghiệm) của T

$$\text{null}(T) = \{\mathbf{x} \in V \mid T\mathbf{x} = \mathbf{0}\}$$

và **range** (hay ảnh) của T là

$$\text{range}(T) = \{\mathbf{y} \in W \mid \exists \mathbf{x} \in V \text{ sao cho } T\mathbf{x} = \mathbf{y}\}$$

5.2 Không gian metric

Metrics được dùng để định nghĩa về khoảng cách trong không gian Euclid (Tuy nhiên, không gian metric thì không cần là không gian vectơ).

Một **metric** trên S là một hàm $d : S \times S \to \mathbb{R}$ thỏa mãn

(i) Tính xác định dương: $d(x, y) \geq 0$, và $d(x, y) = 0 \Leftrightarrow x = y$

(ii) Tính đối xứng: $d(x, y) = d(y, x)$

(iii) Bất đẳng thức tam giác: $d(x, z) \leq d(x, y) + d(y, z)$

với mọi $x, y, z \in S$.

Một động lực chính cho các metric đó là chúng cho phép các giới hạn được định nghĩa như các đối tượng toán học hơn là các số thực. Chúng ta nói rằng một dãy $\{x_n\} \subseteq S$ hội tụ về giới hạn x nếu với mọi $\epsilon > 0$, tồn tại $n \in \mathbb{N}$ sao cho $d(x_n, x) < \epsilon$ với mọi $n \geq N$. Chú ý rằng định nghĩa về giới hạn của dãy các số thực mà bạn tìm thấy trong một lớp giải tích là một trường hợp đặc biệt của định nghĩa này khi sử dụng metric $d(x, y) = |x - y|$.

5.3 Không gian định chuẩn

Chuẩn là một khái niệm về độ dài trong không gian Euclid.

Chuẩn trên không gian vectơ trường số thực V là một hàm $\| \cdot \| : V \to \mathbb{R}$ thỏa mãn

(i) $\|\mathbf{x}\| \geq 0$, và $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$

(ii) $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$

(iii) $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (bất đẳng thức tam giác)

với mọi $\mathbf{x}, \mathbf{y} \in V$ và mọi $\alpha \in \mathbb{R}$. Không gian vectơ có chuẩn được gọi là **không gian vectơ định chuẩn**, hoặc đơn giản là một **không gian định chuẩn**.

Lưu ý rằng bất kỳ chuẩn nào trên V đều quy ra chỉ số khoảng cách trên V

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

Người ta có thể chứng minh rằng các tiên đề về metric đều thỏa mãn các tiên đề về chuẩn. Do đó, bất kỳ không gian định chuẩn nào cũng là một không gian metric.[1]

Chúng ta sẽ chỉ quan tâm tới một số chuẩn đặc biệt trên $\mathbb{R}^n$:

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^{n} |x_i| \\
\|\mathbf{x}\|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} \\
\|\mathbf{x}\|_p &= \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \qquad (p \geq 1) \\
\|\mathbf{x}\|_\infty &= \max_{1 \leq i \leq n} |x_i|
\end{aligned}$$

Lưu ý chuẩn 1 và chuẩn 2 là trường hợp đặc biệt của chuẩn p, và chuẩn ∞ là giới hạn của chuẩn p khi $p \to \infty$ (p tiến ra vô cùng). Chúng ta cần $p \geq 1$ cho định nghĩa về chuẩn p bởi vì bất đẳng thức tam giác không đúng trong trường hợp $p < 1$. (Hãy thử tìm một phản ví dụ!)

5.4 Không gian tích trong

Một **tích trong (tích vô hướng)** trên không gian trường số thực V là một hàm $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ thỏa mãn

(i) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, và $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$

(ii) $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$

(iii) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$

với mọi $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ và với mọi $\alpha, \beta \in \mathbb{R}$. Một không gian vectơ được trang bị bởi một tích vô hướng được gọi là một **không gian tích trong**.

Bất kỳ không gian tích trong nào trên V thì đều đưa ra một chuẩn trên V:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

Và người ta có thể chứng minh được rằng các tiên đề về chuẩn đều thỏa mãn theo các tiên đề về tích vô hướng. Do đó, bất kỳ không gian tích trong nào cũng đều là không gian định chuẩn (và do đó cũng là không gian metric).[2]

[1] Nếu một không gian định chuẩn là đầy đủ đối với metric khoảng cách, được tạo ra bởi chuẩn của nó, thì ta gọi đó là một **không gian Banach**.

[2] Nếu một không gian tích trong là đầy đủ đối với metric khoảng cách được tạo ra bởi tích vô hướng của nó, thì ta gọi đó là một **không gian Hilbert**.

Hai vectơ $\mathbf{x}$ và $\mathbf{y}$ được gọi là **trực giao (orthogonal)** nếu $\langle \mathbf{x}, \mathbf{y} \rangle = 0$; để cho ngắn gọn thì ta viết $\mathbf{x} \perp \mathbf{y}$. Trực giao là sự khái quát hơn về định nghĩa của tính vuông góc từ không gian Euclide. Nếu hai vectơ trực giao $\mathbf{x}$ và $\mathbf{y}$ có độ dài đơn vị (i.e. $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$), thì chúng được gọi là **trực chuẩn (orthonormal)**.

Tích vô hướng chính tắc trên $\mathbb{R}^n$ được cho bởi

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^\top \mathbf{y}$$

Ký hiệu ma trận ở vế phải (xem phần chuyển vị nếu bạn không quen thuộc) phát sinh bởi vì tích trong là một trường hợp đặc biệt của phép nhân ma trận trong đó chúng ta coi ma trận 1×1 thu được là một số vô hướng. Tích trong trên $\mathbb{R}^n$ cũng thường được viết $\mathbf{x} \cdot \mathbf{y}$ (do đó có tên thay thế là **dot product**). Bạn đọc có thể xác minh rằng chuẩn bậc hai $\| \cdot \|_2$ trên $\mathbb{R}^n$ là được tạo ra từ tích trong.

5.4.1 Định lý Pytago

Định lý Pytago la một định lý nổi tiếng bắt nguồn tự nhiên từ không gian các tích vô hướng.

Định lý 1. *Nếu* $\mathbf{x} \perp \mathbf{y}$*, thì*

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

Chứng minh. Giả sử $\mathbf{x} \perp \mathbf{y}$, i.e. $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Thì

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

□

5.4.2 Bất đẳng thức Cauchy-Schwarz

Bất đẳng thức này đôi khi hữu ích trong việc chứng minh các giới hạn:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

với mọi $\mathbf{x}, \mathbf{y} \in V$.

Đẳng thức xảy ra khi $\mathbf{x}$ và $\mathbf{y}$ là bội số vô hướng của nhau (hoặc tương đương với, chúng là phụ thuộc tuyến tính).

5.5 Chuyển vị

Nếu ma trận $\mathbf{A} \in \mathbb{R}^{m \times n}$, thì **chuyển vị** $\mathbf{A}^\top \in \mathbb{R}^{n \times m}$ là được xác định bởi $(\mathbf{A}^\top)_{ij} = A_{ji}$ với mỗi (i, j). Nói cách khác, các cột của ma trận của $\mathbf{A}$ trở thành các hàng của ma trận $\mathbf{A}^\top$, và các hàng của ma trận $\mathbf{A}$ trở thành các cột của ma trận $\mathbf{A}^\top$.

Phép chuyển vị có một số tính chất đại số mà có thể chứng minh từ định nghĩa như:

(i) $(\mathbf{A}^\top)^\top = \mathbf{A}$

(ii) $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

(iii) $(\alpha \mathbf{A})^\top = \alpha \mathbf{A}^\top$

(iv) $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

5.6 Giá trị riêng và vectơ riêng

Với một ma trận vuông $\mathbf{A} \in \mathbb{R}^{n\times n}$, vectơ $\mathbf{x} \neq \mathbf{0}$, nếu

$$\mathbf{Ax} = \lambda\mathbf{x}$$

thì ta nói $\mathbf{x} \in \mathbb{R}^n$ là **vectơ riêng (eigenvector)** của $\mathbf{A}$, và λ là **giá trị riêng (eigenvalue)** của $\mathbf{A}$. Vectơ $\mathbf{0}$ là không bao gồm trong định nghĩa trên vì $\mathbf{A0} = \mathbf{0} = \lambda\mathbf{0}$ với mọi λ.

Mệnh đề 1. *Cho* $\mathbf{x}$ *là một vectơ riêng của* $\mathbf{A}$ *tương ứng với các giá trị riêng* λ*. Thì*

(i) Bất kỳ $\gamma \in \mathbb{R}$, $\mathbf{x}$ *là một vectơ riêng của* $\mathbf{A} + \gamma\mathbf{I}$ *với giá trị riêng* $\lambda + \gamma$*.*

(ii) Nếu $\mathbf{A}$ *là khả nghịch, thì* $\mathbf{x}$ *cũng là một vectơ riêng của* $\mathbf{A}^{-1}$ *với giá trị riêng* λ^{-1}*.*

(iii) $\mathbf{A}^k\mathbf{x} = \lambda^k\mathbf{x}$ *với bất kỳ* $k \in \mathbb{Z}$ *(trong đó* $\mathbf{A}^0 = \mathbf{I}$ *theo định nghĩa).*

Chứng minh. (i) Ta có:

$$(\mathbf{A} + \gamma\mathbf{I})\mathbf{x} = \mathbf{Ax} + \gamma\mathbf{Ix} = \lambda\mathbf{x} + \gamma\mathbf{x} = (\lambda + \gamma)\mathbf{x}$$

(ii) Giả sử $\mathbf{A}$ khả nghịch. Thì

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}(\lambda\mathbf{x}) = \lambda\mathbf{A}^{-1}\mathbf{x}$$

Chia cả 2 vế λ, bởi vì $\mathbf{A}$ khả nghịch nên $\lambda \neq 0$, do đó $\lambda^{-1}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$.

(iii) Ta có: Với bất kỳ $k \in \mathbb{Z}$

$$\mathbf{A}^k\mathbf{x} = \mathbf{A}^{k-1}(\mathbf{Ax}) = \mathbf{A}^{k-1}(\lambda\mathbf{x}) = \lambda\mathbf{A}^{k-1}\mathbf{x} = \cdots = \lambda^k(\mathbf{A}^0\mathbf{x}) = \lambda^k(\mathbf{Ix}) = \lambda^k\mathbf{x}$$

□

5.7 Vết

Vết của một ma trận vuông $\mathbf{A}$, kí hiệu $\operatorname{tr}(\mathbf{A})$, là tổng của tất cả các phần tử trên đường chéo chính:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$$

Vết có một số tính chất như sau:

(i) $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B})$

(ii) $\operatorname{tr}(\alpha\mathbf{A}) = \alpha\operatorname{tr}(\mathbf{A})$

(iii) $\operatorname{tr}\left(\mathbf{A}^\top\right) = \operatorname{tr}(\mathbf{A})$

(iv) $\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{DABC})$

Ba tính chất đầu được suy ra trực tiếp từ định nghĩa. Tính chất cuối được xem như là tính **bất biến dưới hoán vị tuần hoàn**. Chú ý rằng thứ tự của các ma trận không được sắp xếp lại một cách tùy ý. Hay nói cách khác là $\operatorname{tr}(\mathbf{ABCD}) \neq \operatorname{tr}(\mathbf{BACD})$. Ngoài ra, không có gì đặc biệt về tích của bốn ma trận trên - các quy tắc tương tự áp dụng cho nhiều hoặc ít ma trận hơn.

Đặc biệt, vết của một ma trận còn bằng tổng của tất cả các giá trị riêng của ma trận đó

$$\operatorname{tr}(\mathbf{A}) = \sum_i \lambda_i(\mathbf{A})$$

5.8 Định thức

Định thức của một ma trận vuông có thể được định nghĩa theo nhiều cách khác nhau, nhưng đôi khi ta không xem xét tới điều đó, mà chỉ cần xem xét một số tính chất của nó:

(i) $\det(\mathbf{I}) = 1$

(ii) $\det\left(\mathbf{A}^\top\right) = \det(\mathbf{A})$

(iii) $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$

(iv) $\det\left(\mathbf{A}^{-1}\right) = \det(\mathbf{A})^{-1}$

(v) $\det(\alpha^n \mathbf{A}) = \alpha^n \det(\mathbf{A})$

Đặc biệt, định thức của một ma trận bằng tích của tất cả các giá trị riêng của ma trận đó:

$$\det(\mathbf{A}) = \prod_i \lambda_i(\mathbf{A})$$

5.9 Ma trận trực giao

Ma trận $\mathbf{Q} \in \mathbb{R}^{n\times n}$ được gọi là **trực giao** nếu các cột của nó là **trực chuẩn** theo từng cặp. Hay

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$$

hoặc, $\mathbf{Q}^\top = \mathbf{Q}^{-1}$. Một điều thú vị về ma trận trực giao là chúng bảo toàn tích vô hướng:

$$(\mathbf{Qx})^\top(\mathbf{Qy}) = \mathbf{x}^\top\mathbf{Q}^\top\mathbf{Qy} = \mathbf{x}^\top\mathbf{Iy} = \mathbf{x}^\top\mathbf{y}$$

Hơn nữa, các ma trận trực giao cũng bảo toàn chuẩn bậc 2:

$$\|\mathbf{Qx}\|_2 = \sqrt{(\mathbf{Qx})^\top(\mathbf{Qx})} = \sqrt{\mathbf{x}^\top\mathbf{x}} = \|\mathbf{x}\|_2$$

Do đó phép nhân với một ma trận trực giao có thể được coi là một phép biến đổi bảo toàn độ dài, nhưng có thể quay hoặc chiếu vectơ về gốc.

5.10 Ma trận đối xứng

Ma trận $\mathbf{A} \in \mathbb{R}^{n\times n}$ được gọi là **đối xứng** nếu ($\mathbf{A} = \mathbf{A}^\top$), nghĩa là $A_{ij} = A_{ji}$ với mọi (i, j). Định nghĩa này dường như đơn giản nhưng lại có một số ứng dụng quan trọng.

Định lý 2. *(Định lý về phổ) Nếu $\mathbf{A} \in \mathbb{R}^{n\times n}$ là ma trận đối xứng, thì tồn tại một cơ sở trực chuẩn trên $\mathbb{R}^n$ chứa các vectơ riêng của $\mathbf{A}$.*

Định lý này được dùng cho việc phân tích một ma trận đối xứng (hay **phân tích riêng** hoặc **phân tích phổ**). Cơ sở trực chuẩn của các vectơ riêng $\mathbf{q}_1, \ldots, \mathbf{q}_n$ tương ứng với các giá trị riêng $\lambda_1, \ldots, \lambda_n$. $\mathbf{Q}$ là một ma trận trực giao với các cột $\mathbf{q}_1, \ldots, \mathbf{q}_n$, và $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Theo định nghĩa $\mathbf{Aq}_i = \lambda_i\mathbf{q}_i$ với mọi i, từ đó ta có:

$$\mathbf{AQ} = \mathbf{Q\Lambda}$$

Nhân vào phía phải của 2 vế bởi $\mathbf{Q}^\top$, ta được

$$\mathbf{A} = \mathbf{Q\Lambda Q}^\top$$

5.10.1 Thương Rayleigh

Cho $\mathbf{A} \in \mathbb{R}^{n\times n}$ là một ma trận đối xứng. Biểu diễn $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ được gọi là một **dạng bậc 2**.

Hóa ra là có một sự liên kết thú vị giữa dạng bậc 2 của ma trận đối xứng và các giá trị riêng của nó. Điều này được thể hiện qua thương **Rayleigh**

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

Thương Rayleigh có hai tính chất quan trọng mà người đọc có thể (và nên) chứng minh từ định nghĩa

(i) **Tính bất biến**: với bất kỳ vectơ $\mathbf{x} \neq \mathbf{0}$ và vô hướng $\alpha \neq 0$, $R_{\mathbf{A}}(\mathbf{x}) = R_{\mathbf{A}}(\alpha\mathbf{x})$.

(ii) Nếu $\mathbf{x}$ là một vectơ riêng của $\mathbf{A}$ cùng với giá trị riêng λ, thì $R_{\mathbf{A}}(\mathbf{x}) = \lambda$.

Chúng ta có thể chứng minh thêm rằng thương số Rayleigh được giới hạn bởi các giá trị riêng lớn nhất và nhỏ nhất của $\mathbf{A}$. Nhưng trước tiên, ta cần chỉ ra một trường hợp đặc biệt hữu ích của kết quả cuối cùng.

Mệnh đề 2. *Với bất kỳ* $\mathbf{x}$ *sao cho* $\|\mathbf{x}\|_2 = 1$*,*

$$\lambda_{\min}(\mathbf{A}) \leq \mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$$

dấu $=$ *xảy ra khi và chỉ khi* $\mathbf{x}$ *là vectơ riêng tương ứng.*

Chứng minh. Chúng ta chỉ chứng minh với trường hợp max bởi vì với trường hợp min thì hoàn toàn tương tự.

Bởi vì $\mathbf{A}$ đối xứng, ta có thể phân tích $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$. Sau đó sử dụng phép đổi biến $\mathbf{y} = \mathbf{Q}^\top \mathbf{x}$, chú ý mỗi quan hệ giữa $\mathbf{x}$ và $\mathbf{y}$ là một-một và $\|\mathbf{y}\|_2 = 1$ do $\mathbf{Q}$ là trực giao. Do đó

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^\top \mathbf{\Lambda}\mathbf{y} = \max_{y_1^2+\cdots+y_n^2=1} \sum_{i=1}^{n} \lambda_i y_i^2$$

Với cách viết như trên, hiển nhiên $\mathbf{y}$ tối đa hóa biểu thức trên khi và chỉ khi $\sum_{i\in I} y_i^2 = 1$ trong đó $I = \{i : \lambda_i = \max_{j=1,\ldots,n} \lambda_j = \lambda_{\max}(\mathbf{A})\}$ và $y_j = 0$ với $j \notin I$. Khi đó, I chứa chỉ số của giá trị riêng lớn nhất. Trong trường hợp này, giá trị lớn nhất của biểu thức là

$$\sum_{i=1}^{n} \lambda_i y_i^2 = \sum_{i\in I} \lambda_i y_i^2 = \lambda_{\max}(\mathbf{A}) \sum_{i\in I} y_i^2 = \lambda_{\max}(\mathbf{A})$$

Và do $\mathbf{q}_1, \ldots, \mathbf{q}_n$ là các vectơ cột của $\mathbf{Q}$, ta có

$$\mathbf{x} = \mathbf{Q}\mathbf{Q}^\top \mathbf{x} = \mathbf{Q}\mathbf{y} = \sum_{i=1}^{n} y_i \mathbf{q}_i = \sum_{i\in I} y_i \mathbf{q}_i$$

Nhớ rằng $\mathbf{q}_1, \ldots, \mathbf{q}_n$ là các vectơ của $\mathbf{A}$ và tạo thành một cơ sở trực chuẩn trong $\mathbb{R}^n$. Bằng cách xây dựng tập $\{\mathbf{q}_i : i \in I\}$ tạo thành một cơ sở trực chuẩn cho không gian riêng của $\lambda_{\max}(\mathbf{A})$. Do đó $\mathbf{x}$ chính là một tổ hợp tuyến tính trong không gian riêng này, đồng thời cũng là một vectơ riêng của $\mathbf{A}$ tương ứng với $\lambda_{\max}(\mathbf{A})$.

Như vậy, ta đã chứng minh $\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^\top \mathbf{A}\mathbf{x} = \lambda_{\max}(\mathbf{A})$, từ bất đẳng thức $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq \lambda_{\max}(\mathbf{A})$ với $\mathbf{x}$ có độ dài đơn vị. □

Theo tính bất biến của thương Rayleigh, chúng ta ngay lập tức có hệ quả (do $\mathbf{x}^\top\mathbf{A}\mathbf{x} = R_\mathbf{A}(\mathbf{x})$ với vectơ đơn vị $\mathbf{x}$)

Định lý 3. *(Min-max theorem) Với mọi* $\mathbf{x} \neq \mathbf{0}$*,*

$$\lambda_{\min}(\mathbf{A}) \leq R_\mathbf{A}(\mathbf{x}) \leq \lambda_{\max}(\mathbf{A})$$

dấu "=" xảy ra khi và chỉ khi $\mathbf{x}$ *là một vectơ riêng tương ứng.*

5.11 Ma trận xác định dương (Bán xác định dương)

Ma trận đối xứng $\mathbf{A}$ là **bán xác định dương** nếu với mọi $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$. Người ta thường viết $\mathbf{A} \succeq 0$ để chỉ $\mathbf{A}$ là ma trận **bán xác định dương**.

Ma trận đối xứng $\mathbf{A}$ là **xác định dương** nếu với mọi $\mathbf{x} \neq 0, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$. Người ta thường viết $\mathbf{A} \succ 0$ để chỉ $\mathbf{A}$ là ma trận **xác định dương**. Lưu ý rằng tính xác định dương là một thuộc tính mạnh hơn bán xác định dương, theo nghĩa là mọi ma trận xác định dương đều là bán xác định dương nhưng điều ngược lại thì không đúng.

Mệnh đề 3. *Ma trận đối xứng là bán xác định dương khi và chỉ khi tất cả các giá trị riêng của nó là không âm và xác định dương khi và chỉ khi tất cả các giá trị riêng của nó đều dương.*

Chứng minh. Giả sử A là ma trận nửa xác định dương, và $\mathbf{x}$ là một vectơ riêng của $\mathbf{A}$ cùng với giá trị riêng λ. Thì

$$0 \leq \mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top(\lambda\mathbf{x}) = \lambda\mathbf{x}^\top\mathbf{x} = \lambda\|\mathbf{x}\|_2^2$$

Vì $\mathbf{x} \neq \mathbf{0}$ (theo định nghĩa vectơ riêng), ta có $\|\mathbf{x}\|_2^2 > 0$, chia cả 2 vế cho $\|\mathbf{x}\|_2^2$ dẫn đến $\lambda \geq 0$. Nếu $\mathbf{A}$ là xác định dương, ta cần chứng minh $\lambda > 0$.

Để đơn giản hóa việc chứng minh này, ta sẽ sử dụng các tính chất của thương số Rayleigh. Giả sử rằng $\mathbf{A}$ là đối xứng và tất cả các giá trị riêng của nó là không âm. Thì với mọi $\mathbf{x} \neq \mathbf{0}$,

$$0 \leq \lambda_{\min}(\mathbf{A}) \leq R_\mathbf{A}(\mathbf{x})$$

Vì $\mathbf{x}^\top\mathbf{A}\mathbf{x}$ cùng dấu với $R_\mathbf{A}(\mathbf{x})$, ta kết luận rằng $\mathbf{A}$ là bán xác định dương. Nếu tất cả các giá trị riêng của $\mathbf{A}$ đều dương, thì $0 < \lambda_{\min}(\mathbf{A})$, từ đó ta kết luận được $\mathbf{A}$ là ma trận xác định dương. □

Mệnh đề 4. *Giả sử* $\mathbf{A} \in \mathbb{R}^{m\times n}$*. Thì* $\mathbf{A}^\top\mathbf{A}$ *là ma trận bán xác định dương. Nếu* $\text{null}(\mathbf{A}) = \{\mathbf{0}\}$*, thì* $\mathbf{A}^\top\mathbf{A}$ *là ma trận xác định dương.*

Chứng minh. Với bất kỳ $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} = (\mathbf{A}\mathbf{x})^\top(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0$$

nên $\mathbf{A}^\top\mathbf{A}$ là bán xác định dương.

Để ý rằng $\|\mathbf{A}\mathbf{x}\|_2^2 = 0$ thì $\|\mathbf{A}\mathbf{x}\|_2 = 0$, đồng nghĩa với $\mathbf{A}\mathbf{x} = \mathbf{0}$ (theo tính chất của chuẩn). Nếu $\text{null}(\mathbf{A}) = \{\mathbf{0}\}$, $\mathbf{A}\mathbf{x} = \mathbf{0}$ ta suy ra $\mathbf{x} = \mathbf{0}$, nên $\mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} = 0$ khi và chỉ khi $\mathbf{x} = \mathbf{0}$,và do đó $\mathbf{A}^\top\mathbf{A}$ là ma trận xác định dương. □

Các ma trận xác định dương thì khả nghịch (vì các giá trị riêng khác 0), trong khi đó đối với các ma trận bán xác định dương thì có thể không đúng. Tuy nhiên, nếu ta đã có ma trận bán xác định dương, ta có thể điều chỉnh đường chéo của nó một chút để tạo ra ma trận xác định dương.

Mệnh đề 5. *Nếu* $\mathbf{A}$ *là ma trận nửa xác định dương và* $\epsilon > 0$*, thì* $\mathbf{A} + \epsilon\mathbf{I}$ *là ma trận xác định dương.*

Chứng minh. Giả sử $\mathbf{A}$ là bán xác định dương và $\epsilon > 0$, với bất kỳ $\mathbf{x} \neq \mathbf{0}$ ta có

$$\mathbf{x}^\top(\mathbf{A} + \epsilon\mathbf{I})\mathbf{x} = \mathbf{x}^\top\mathbf{A}\mathbf{x} + \epsilon\mathbf{x}^\top\mathbf{I}\mathbf{x} = \underbrace{\mathbf{x}^\top\mathbf{A}\mathbf{x}}_{\geq 0} + \underbrace{\epsilon\|\mathbf{x}\|_2^2}_{>0} > 0$$

được chứng minh. □

Một hệ quả hiển nhiên nhưng rất hữu ích của hai mệnh đề trên mà chúng ta vừa chỉ ra là $\mathbf{A}^\top\mathbf{A} + \epsilon\mathbf{I}$ là ma trận xác định dương (đồng nghĩa khả nghịch) với *bất kỳ* ma trận $\mathbf{A}$ và $\epsilon > 0$.

5.11.1 Hình học của các dạng bậc hai xác định dương

Một cách hữu ích để hiểu các dạng bậc hai là sử dụng hình học của các tập đồng mức của chúng. Một tập **đồng mức** hay **isocontour** của một hàm là tập hợp tất cả các đầu vào sao cho giá trị đầu ra của chúng là như nhau. Về mặt toán học, c-isocontour của f là $\{\mathbf{x} \in \operatorname{dom} f : f(\mathbf{x}) = c\}$.

Cùng xem xét một trường hợp $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ với $\mathbf{A}$ là ma trận xác định dương. Vì $\mathbf{A}$ xác định dương, nên có một căn bậc hai duy nhất $\mathbf{A}^{\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$, với $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ là phép chéo hóa ma trận (phân tích riêng) của $\mathbf{A}$ và $\mathbf{\Lambda}^{\frac{1}{2}} = \operatorname{diag}(\sqrt{\lambda_1}, \dots \sqrt{\lambda_n})$. Dễ thấy $\mathbf{A}^{\frac{1}{2}}$ là xác định dương và thỏa mãn $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$. Cố định $c \geq 0$, tập đồng mức c của f là tập $\mathbf{x} \in \mathbb{R}^n$ sao cho

$$c = \mathbf{x}^\top\mathbf{A}\mathbf{x} = \mathbf{x}^\top\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}\mathbf{x} = \|\mathbf{A}^{\frac{1}{2}}\mathbf{x}\|_2^2$$

với ma trận đối xứng $\mathbf{A}^{\frac{1}{2}}$. Đổi biến $\mathbf{z} = \mathbf{A}^{\frac{1}{2}}\mathbf{x}$, thì $\|\mathbf{z}\|_2 = \sqrt{c}$. Do đó $\mathbf{z}$ tạo thành hình cầu bán kính $\sqrt{c}$. Biến đổi $\mathbf{z} = \sqrt{c}\hat{\mathbf{z}}$ trong đó $\hat{\mathbf{z}}$ có $\|\hat{\mathbf{z}}\|_2 = 1$. Từ $\mathbf{A}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top$, ta có

$$\mathbf{x} = \mathbf{A}^{-\frac{1}{2}}\mathbf{z} = \mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^\top\sqrt{c}\hat{\mathbf{z}} = \sqrt{c}\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\tilde{\mathbf{z}}$$

với $\tilde{\mathbf{z}} = \mathbf{Q}^\top\hat{\mathbf{z}}$ cũng thỏa mãn $\|\tilde{\mathbf{z}}\|_2 = 1$ do $\mathbf{Q}$ là trực giao. Sử dụng cách tham số hóa này, ta thấy rằng tập nghiệm $\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) = c\}$ là ảnh của các khối cầu đơn vị $\{\tilde{\mathbf{z}} \in \mathbb{R}^n : \|\tilde{\mathbf{z}}\|_2 = 1\}$ dưới phép biến đổi $\mathbf{x} = \sqrt{c}\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\tilde{\mathbf{z}}$.

Qua các phân tích trên, dễ thấy tập đồng mức c-isocontour của hàm f chính là đạt được qua việc ứng dụng một chuỗi các phép biến đổi tuyến tính.

Tóm lại: tập đồng mức của hàm $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{A}\mathbf{x}$ là các ellipsoid sao cho các trục hướng theo hướng các vectơ riêng của ma trận $\mathbf{A}$, và bán kính của các trục này tỷ lệ với căn bậc hai nghịch đảo của các giá trị riêng tương ứng.

5.12 Phân tích giá trị suy biến

Phân tích giá trị suy biến (SVD - Singular value decomposition) là một công cụ được ứng dụng phổ biến trong đại số tuyến tính. Điểm mạnh của nó bắt nguồn từ thực tế là *mọi ma trận* $\mathbf{A} \in \mathbb{R}^{m\times n}$ thì đều có phân tích SVD (có thể không vuông)! Cụ thể:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

trong đó $\mathbf{U} \in \mathbb{R}^{m\times m}$ và $\mathbf{V} \in \mathbb{R}^{n\times n}$ là ma trận trực giao và $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ là một ma trận chéo với **các giá trị suy biến** của $\mathbf{A}$ (ký hiệu σ_i) trên đường chéo.

Theo quy ước, các giá trị suy biến có thứ tự

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$$

Chỉ có r giá trị suy biến đầu tiên khác 0, trong đó r là hạng của ma trận $\mathbf{A}$.

Quan sát phép phân tích giá trị suy biến cho $\mathbf{A}^\top\mathbf{A}$ và $\mathbf{A}\mathbf{A}^\top$:

$$\mathbf{A}^\top\mathbf{A} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^\top\mathbf{\Sigma}\mathbf{V}^\top$$
$$\mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top$$

Theo đó, các cột của $\mathbf{V}$ (**các vectơ suy biến bên phải** của $\mathbf{A}$) là các vectơ riêng của $\mathbf{A}^\top\mathbf{A}$, và các cột của $\mathbf{U}$ (**các vectơ suy biến bên trái** của $\mathbf{A}$) là các vectơ riêng của $\mathbf{A}\mathbf{A}^\top$.

Các ma trận $\mathbf{\Sigma}^\top\mathbf{\Sigma}$ và $\mathbf{\Sigma}\mathbf{\Sigma}^\top$ không nhất thiết phải cùng cỡ, nhưng cả hai có đường chéo là bình phương giá trị suy biến σ_i^2 (có thể có một vài phần tử 0).

5.13 Một số nhận dạng hữu ích về ma trận

5.13.1 Tích ma trận - vectơ như là một tổ hợp tuyến tính các cột của ma trận

Mệnh đề 6. *Cho $\mathbf{x} \in \mathbb{R}^n$ là một vectơ và $\mathbf{A} \in \mathbb{R}^{m\times n}$ là một ma trận với các cột $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Thì*

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i\mathbf{a}_i$$

Điều này thực sự hữu ích trong việc hiểu các toán tử tuyến tính trên ma trận.

5.13.2 Tổng các tích ngoài như là tích của hai ma trận

Một **tích ngoài** là một cách biểu diễn $\mathbf{a}\mathbf{b}^\top$, trong đó $\mathbf{a} \in \mathbb{R}^m$ và $\mathbf{b} \in \mathbb{R}^n$. Bằng cách kiểm tra nó không khó để thấy rằng một biểu thức như vậy tạo ra một ma trận cỡ $m \times n$ sao cho

$$[\mathbf{a}\mathbf{b}^\top]_{ij} = a_i b_j$$

Mệnh đề 7. *Cho $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^m$ và $\mathbf{b}_1, \ldots, \mathbf{b}_k \in \mathbb{R}^n$. Thì*

$$\sum_{\ell=1}^{k} \mathbf{a}_\ell\mathbf{b}_\ell^\top = \mathbf{A}\mathbf{B}^\top$$

trong đó

$$\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_k\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}\mathbf{b}_1 & \cdots & \mathbf{b}_k\end{bmatrix}$$

Chứng minh. Với mỗi cặp (i, j), ta có

$$\left[\sum_{\ell=1}^{k} \mathbf{a}_\ell\mathbf{b}_\ell^\top\right]_{ij} = \sum_{\ell=1}^{k}[\mathbf{a}_\ell\mathbf{b}_\ell^\top]_{ij} = \sum_{\ell=1}^{k}[\mathbf{a}_\ell]_i[\mathbf{b}_\ell]_j = \sum_{\ell=1}^{k} A_{i\ell}B_{j\ell}$$

Biểu diễn biểu thức cuối cùng bên trên thể xem như là tích trong giữa hàng i của $\mathbf{A}$ và hàng j của $\mathbf{B}$, hay là các cột j của $\mathbf{B}^\top$. Do đó theo định nghĩa về phép nhân hai ma trận, điều đó tương đương $[\mathbf{A}\mathbf{B}^\top]_{ij}$. □

5.13.3 Dạng bậc hai

Cho $\mathbf{A} \in \mathbb{R}^{n \times n}$ là một ma trận đối xứng, biểu diễn $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ được gọi là một dạng bậc hai của $\mathbf{A}$. Trong một số trường hợp sẽ hữu ích hơn khi viết dạng bậc hai giữa $\mathbf{A}$ và $\mathbf{x}$ theo từng phần tử:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

Điều này đúng với mọi ma trận vuông bất kỳ (không cần phải đối xứng), dù dạng bậc hai thường chỉ được đề cập đến trong các nội dung về ma trận đối xứng.

6 Giải tích và Tối ưu hóa

Phần lớn các bài toán trong học máy sẽ đi cực tiểu hóa **hàm mất mát** (hay còn được gọi là **hàm mục tiêu** trong cộng đồng tối ưu hóa), một hàm vô hướng của nhiều biến để đo mức độ khớp của mô hình với dữ liệu mà ta đang có.

6.1 Cực trị

Tối ưu hóa là đi tìm điểm **cực trị**, sao cho giá trị tại điểm đó là bé nhất hoặc lớn nhất. Khi xác định cực trị, cần phải xem xét tập hợp các đầu vào mà chúng ta đang tối ưu. Tập $\mathcal{X} \subseteq \mathbb{R}^d$ được gọi là **tập chấp nhận được**. Nếu $\mathcal{X}$ là toàn bộ miền xác định của hàm được tối ưu hóa, ta gọi đây bài toán **không ràng buộc**. Ngược lại, bài toán thì là **bị ràng buộc** và sẽ khó hơn trong việc đi tìm nghiệm tối ưu, phụ thuộc vào tập chấp nhận được.

Giả sử $f : \mathbb{R}^d \to \mathbb{R}$. Điểm $\mathbf{x}$ được gọi là **cực tiểu địa phương** (tương ứng **cực đại địa phương**) của f trong $\mathcal{X}$ nếu $f(\mathbf{x}) \leq f(\mathbf{y})$ (tương ứng $f(\mathbf{x}) \geq f(\mathbf{y})$) với mọi $\mathbf{y}$ trong vùng lân cận $N \subseteq \mathcal{X}$ của $\mathbf{x}$.[3] Hơn nữa, nếu $f(\mathbf{x}) \leq f(\mathbf{y})$ với mọi $\mathbf{y} \in \mathcal{X}$, thì $\mathbf{x}$ là một **cực tiểu toàn cục** của f trong $\mathcal{X}$ (tương tự với cực đại toàn cục).

Định nghĩa **ngặt (strict)** (ví dụ, cực tiểu địa phương ngặt) nếu bất đẳng thức trên không xảy ra trường hợp dấu "=". Điều này đồng nghĩa với các cực trị là duy nhất trong vùng lân cận của chúng.

Thấy rằng cực đại hóa hàm f cũng tương đương với việc ta đi cực tiểu hóa hàm $-f$, vì vậy bài toán tối ưu hóa thường được diễn đạt theo nghĩa cực tiểu hóa mà không làm mất đi tính tổng quát.

6.2 Gradients

Khái niệm quan trọng nhất từ giải tích được sử dụng trong lĩnh vực học máy là **gradient**. Gradient tổng quát hóa các đạo hàm thành các hàm vô hướng của nhiều biến. Gradient của hàm $f : \mathbb{R}^d \to \mathbb{R}$, ký hiệu ∇f, được định nghĩa bởi

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad \text{i.e.} \quad [\nabla f]_i = \frac{\partial f}{\partial x_i}$$

Gradient có một số tính chất quan trọng: $\nabla f(\mathbf{x})$ chỉ theo hướng **tăng nhiều nhất** từ $\mathbf{x}$. Tương tự, $-\nabla f(\mathbf{x})$ chỉ theo hướng **giảm nhiều nhất** từ $\mathbf{x}$. Chúng ta sẽ sử dụng tính chất này thường xuyên và lặp đi lặp lại khi cực tiểu hóa một hàm, gọi là **gradient descent**.

6.3 Jacobian

Jacobian của hàm $f : \mathbb{R}^n \to \mathbb{R}^m$ là một ma trận của các đạo hàm riêng bậc nhất:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad \text{i.e.} \quad [\mathbf{J}_f]_{ij} = \frac{\partial f_i}{\partial x_j}$$

Khi $m = 1$, khi đó $\nabla f = \mathbf{J}_f^\top$.

[3] Một **lân cận** của $\mathbf{x}$ là một tập mở có chứa $\mathbf{x}$.

6.4 Hessian

Ma trận **Hessian** của hàm $f : \mathbb{R}^d \to \mathbb{R}$ là một ma trận của các đạo hàm riêng cấp hai:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Lưu ý rằng các đạo hàm riêng liên tục, thì thứ tự đạo hàm có thể hoán đổi cho nhau, tức là $[\nabla^2 f]_{ij} = [\nabla^2 f]_{ji}$ (Định lý Clairaut), nên ma trận Hessian là đối xứng. Đây là tính chất phổ biến trên các hàm khả vi mà chúng ta làm việc.

Ma trận Hessian được sử dụng trong một số thuật toán tối ưu hóa như phương pháp Newton. Phương pháp này mất "nhiều chi phí" nhưng có thể giảm số bước lặp cần hội tụ đến điểm cực tiểu địa phương bằng cách sử dụng thông tin độ cong của hàm f.

6.5 Giải tích ma trận

Bởi vì rất nhiều phương pháp tối ưu hóa đi các điểm, mà tại điểm đó gradient giảm, nên sẽ hữu ích khi ta biết về các quy tắc đạo hàm của biểu thức chứa ma trận và vector. Ở đây, chúng ta đưa ra một số quy tắc phổ biến. Trong đó, hai quy tắc có thể xem là quan trọng nhất là

$$\nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \mathbf{a}$$
$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

Lưu ý rằng ở quy tắc thứ hai chỉ xảy ra khi $\mathbf{A}$ là ma trận vuông. Hơn nữa, nếu $\mathbf{A}$ là ma trận đối xứng, chúng ta có thể đơn giản hóa kết quả nhận về là $2\mathbf{A}\mathbf{x}$.

6.5.1 Quy tắc chuỗi

Hầu hết các hàm chúng ta tối ưu hóa sẽ là các hàm phức tạp. Chúng thường là hàm hợp của các hàm đơn giản. Quy tắc chuỗi cho ta một cách tính về đạo hàm của một hàm hợp từ đạo hàm của các hàm đơn lẻ tạo nên chúng.

Quy tắc chuỗi đối với đạo hàm một biến:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

trong đó $\circ$ ký hiệu cho hàm hợp. Có một sự tổng quát hóa tự nhiên của quy tắc này thành các hàm nhiều biến.

Mệnh đề 8. *Giả sử $f : \mathbb{R}^m \to \mathbb{R}^k$ và $g : \mathbb{R}^n \to \mathbb{R}^m$. Thì $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ và*

$$\mathbf{J}_{f \circ g}(\mathbf{x}) = \mathbf{J}_f(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$$

Trong trường hợp $k = 1$ ta có hệ quả sau $\nabla f = \mathbf{J}_f^\top$.

Hệ quả 1. *Giả sử $f : \mathbb{R}^m \to \mathbb{R}$ và $g : \mathbb{R}^n \to \mathbb{R}^m$. Thì $f \circ g : \mathbb{R}^n \to \mathbb{R}$ và*

$$\nabla(f \circ g)(\mathbf{x}) = \mathbf{J}_g(\mathbf{x})^\top \nabla f(g(\mathbf{x}))$$

6.6 Định lý Taylor

Định lý Taylor là sự tổng quát hóa với các hàm nhiều biến thay vì một biến.

Định lý 4. *(Định lý Taylor) Giả sử $f : \mathbb{R}^d \to \mathbb{R}$ là khả vi liên tục, và cho $\mathbf{h} \in \mathbb{R}^d$. Thì tồn tại $t \in (0, 1)$ sao cho*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + t\mathbf{h})^\top \mathbf{h}$$

Hơn nữa, nếu f khả vi liên tục 2 lần, thì

$$\nabla f(\mathbf{x} + \mathbf{h}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}\, \mathrm{d}t$$

và tồn tại $t \in (0, 1)$ sao cho

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Định lý này được sử dụng trong chứng minh về các điều kiện cho cực tiểu cục bộ của các bài toán tối ưu hóa không ràng buộc. Một số kết quả quan trọng nhất được đưa ra trong phần tiếp theo.

6.7 Điều kiện cực tiểu địa phương

Mệnh đề 9. *Nếu $\mathbf{x}^*$ là một cực tiểu địa phương của f và f khả vi liên tục trong một lân cận của $\mathbf{x}^*$, thì $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Chứng minh. Cho $\mathbf{x}^*$ là một cực tiểu địa phương của f, và giả sử $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Chọn $\mathbf{h} = -\nabla f(\mathbf{x}^*)$, do ∇f liên tục nên ta có

$$\lim_{t \to 0} -\nabla f(\mathbf{x}^* + t\mathbf{h}) = -\nabla f(\mathbf{x}^*) = \mathbf{h}$$

Do đó

$$\lim_{t \to 0} \mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) = \mathbf{h}^\top \nabla f(\mathbf{x}^*) = -\|\mathbf{h}\|_2^2 < 0$$

Như vậy có tồn tại $T > 0$ sao cho $\mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) < 0$ với mọi $t \in [0, T]$. Áp dụng định lý Taylor: với bất kỳ $t \in (0, T]$, tồn tại $t' \in (0, t)$ thỏa mãn

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + t\mathbf{h}^\top \nabla f(\mathbf{x}^* + t'\mathbf{h}) < f(\mathbf{x}^*)$$

Từ đó, $\mathbf{x}^*$ không phải là cực tiểu địa phương, mẫu thuẫn với giả thiết. Vậy $\nabla f(\mathbf{x}^*) = \mathbf{0}$. □

Bằng chứng cho chúng ta thấy lý do tại sao gradient biến mất ($= 0$ lại cần thiết cho một điểm cực trị: nếu $\nabla f(\mathbf{x})$ khác 0, thì luôn tồn tại $\alpha > 0$ để $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}))) < f(\mathbf{x})$. Vì thế, $-\nabla f(\mathbf{x})$ được gọi là **hướng giảm**.

Các điểm mà gradient biến mất được gọi là **điểm dừng**. Lưu ý rằng không phải tất cả điểm dừng thì đều là cực trị.

Ta thấy rằng các thông tin đạo hàm bậc nhất (như gradient) là không đủ đối với cực tiểu toàn cục. Nhưng ta có thể nói nhiều hơn về thông tin bậc hai (Hessian). Đầu tiên, chúng ta chứng minh một điều kiện cần bậc hai cho cực tiểu địa phương.

Mệnh đề 10. *Nếu $\mathbf{x}^*$ là một cực tiểu địa phương f và f khả vi liên tục hai lần trong một lân cận của $\mathbf{x}^*$, thì $\nabla^2 f(\mathbf{x}^*)$ là nửa xác định dương.*

Chứng minh. Cho $\mathbf{x}^*$ là một cực tiểu địa phương của hàm f, và giả sử rằng $\nabla^2 f(\mathbf{x}^*)$ không phải là ma trận nửa xác định dương. Chọn $\mathbf{h}$ sao cho $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$, $\nabla^2 f$ liên tục nên ta có

$$\lim_{t\to 0} \nabla^2 f(\mathbf{x}^* + t\mathbf{h}) = \nabla^2 f(\mathbf{x}^*)$$

Do đó

$$\lim_{t\to 0} \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} = \mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$$

Như vậy tồn tại $T > 0$ sao cho $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} < 0$ với mọi $t \in [0, T]$. Áp dụng định lý Taylor: với bất kỳ $t \in (0, T]$, tồn tại $t' \in (0, t)$ thỏa mãn

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + \underbrace{t\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2} t^2 \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t'\mathbf{h})\mathbf{h} < f(\mathbf{x}^*)$$

Nên $\mathbf{x}^*$ không phải là cực tiểu địa phương (mâu thuẫn). Vậy $\nabla^2 f(\mathbf{x}^*)$ là ma trận nửa xác định dương. □

Bây giờ chúng ta đưa ra các điều kiện đủ cho cực tiểu địa phương.

Mệnh đề 11. *Giả sử f là hàm khả vi liên tục hai lần với $\nabla^2 f$ là ma trận nửa xác định dương trong lân cận của $\mathbf{x}^*$, và $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Thì $\mathbf{x}^*$ là một cực tiểu địa phương của hàm f. Hơn nữa nếu $\nabla^2 f(\mathbf{x}^*)$ là ma trận xác định dương, thì $\mathbf{x}^*$ là cực tiểu địa phương ngặt.*

Chứng minh. Cho B là một hình cầu mở với bán kính $r > 0$ tâm $\mathbf{x}^*$, trong lân cận của $\mathbf{x}^*$. Áp dụng định lý Taylor, ta có với bất kỳ $\mathbf{h}$, $\|\mathbf{h}\|_2 < r$, tồn tại $t \in (0, 1)$ sao cho

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \underbrace{\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2} \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq f(\mathbf{x}^*)$$

Bất đẳng thức cuối cùng tồn tại vì $\nabla^2 f(\mathbf{x}^* + t\mathbf{h})$ là nửa xác định dương (do $\|t\mathbf{h}\|_2 = t\|\mathbf{h}\|_2 < \|\mathbf{h}\|_2 < r$), nên $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq 0$. Vì $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \mathbf{h})$ với mọi hướng $\mathbf{h}$, $\|\mathbf{h}\|_2 < r$, ta kết luận rằng $\mathbf{x}^*$ là một cực tiểu địa phương.

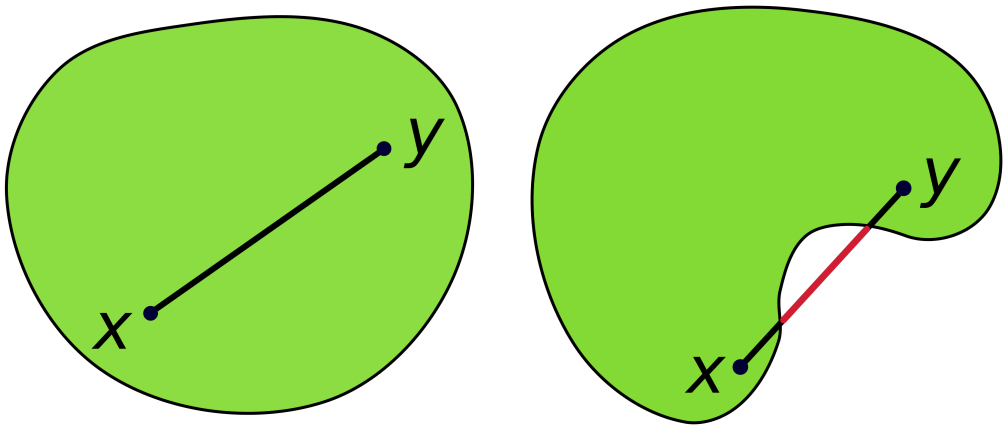
Giờ ta giả sử thêm rằng $\nabla^2 f(\mathbf{x}^*)$ là ma trận xác định dương thực sự. Do Hessian liên tục, ta có thể chọn một hình cầu khác là B' với bán kính $r' > 0$ tại tâm $\mathbf{x}^*$ sao cho $\nabla^2 f(\mathbf{x})$ là xác định dương với mọi $\mathbf{x} \in B'$. Theo lập luận như trên (và ma trận Hessian là xác định dương) ta có $f(\mathbf{x}^* + \mathbf{h}) > f(\mathbf{x}^*)$ với mọi $\mathbf{h}$ với $0 < \|\mathbf{h}\|_2 < r'$. Do đó $\mathbf{x}^*$ là cực tiểu địa phương ngặt. □

Lưu ý rằng, trong điều ngược lại, các điều kiện $\nabla f(\mathbf{x}^*) = \mathbf{0}$ và $\nabla^2 f(\mathbf{x}^*)$ nửa xác định dương không đủ để kết luận $\mathbf{x}^*$ là cực tiểu địa phương!

Xét hàm $f(x) = x^3$. Ta có $f'(0) = 0$ và $f''(0) = 0$ (nên ma trận Hessian trong trường hợp này là $\big[0\big]$, nửa xác định dương). Nhưng f có một điểm yên ngựa tại $x = 0$. Hàm $f(x) = -x^4$ có gradient và ma trận Hessian giống nhau tại $x = 0$, nhưng $x = 0$ là cực đại ngặt của hàm này!

6.8 Tính lồi

Tính lồi là một thuật ngữ dùng để chỉ cho cả tập lồi và hàm lồi. Đối với các hàm, có các mức độ lồi khác nhau và độ lồi của một hàm cho chúng ta biết rất nhiều về cực tiểu của nó: chúng có tồn tại không, chúng có duy nhất không, chúng ta có thể tìm thấy chúng nhanh như thế nào bằng cách sử dụng các thuật toán tối ưu hóa, v.v. Trong phần này, chúng ta trình bày các kết quả cơ bản liên quan đến độ lồi, độ lồi nghiêm ngặt và độ lồi mạnh.



(a) Một tập lồi (b) Một tập không lồi

Hình 1: Tập hợp lồi trông như thế nào

6.8.1 Tập lồi

Một tập hợp $\mathcal{X} \subseteq \mathbb{R}^d$ là **lồi (convex)** nếu

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{X}$$

với mọi $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ và với mọi $t \in [0, 1]$.

Về mặt hình học, điều này có nghĩa là tất cả các điểm trên đoạn thẳng giữa hai điểm bất kỳ trong $\mathcal{X}$ thì cũng đều nằm trong $\mathcal{X}$. Nhìn Hình 1 để hiểu rõ.

Tại sao chúng ta quan tâm đến việc một tập hợp có phải là tập lồi hay không? Chúng ta sẽ thấy ở phần sau rằng bản chất của cực tiểu có thể phụ thuộc rất nhiều vào việc tập chấp nhận được có phải là tập lồi hay không. Các kết quả không mong muốn có thể xảy ra khi chúng ta cho phép tập chấp nhận được là tùy ý, vì vậy đối với các chứng minh, chúng ta sẽ cần giả định rằng nó là tập hợp lồi. May thay, chúng ta thường cực tiểu hóa trên hầu hết các bài toán là miền $\mathbb{R}^d$, dễ dàng thấy đây là một tập lồi.

6.8.2 Khái niệm cơ bản về hàm lồi

Trong phần còn lại của phần này, giả định $f : \mathbb{R}^d \to \mathbb{R}$ trừ khi có ghi chú khác. Chúng ta sẽ bắt đầu với các định nghĩa và sau đó đưa ra một số kết quả.

Một hàm f là **lồi** nếu

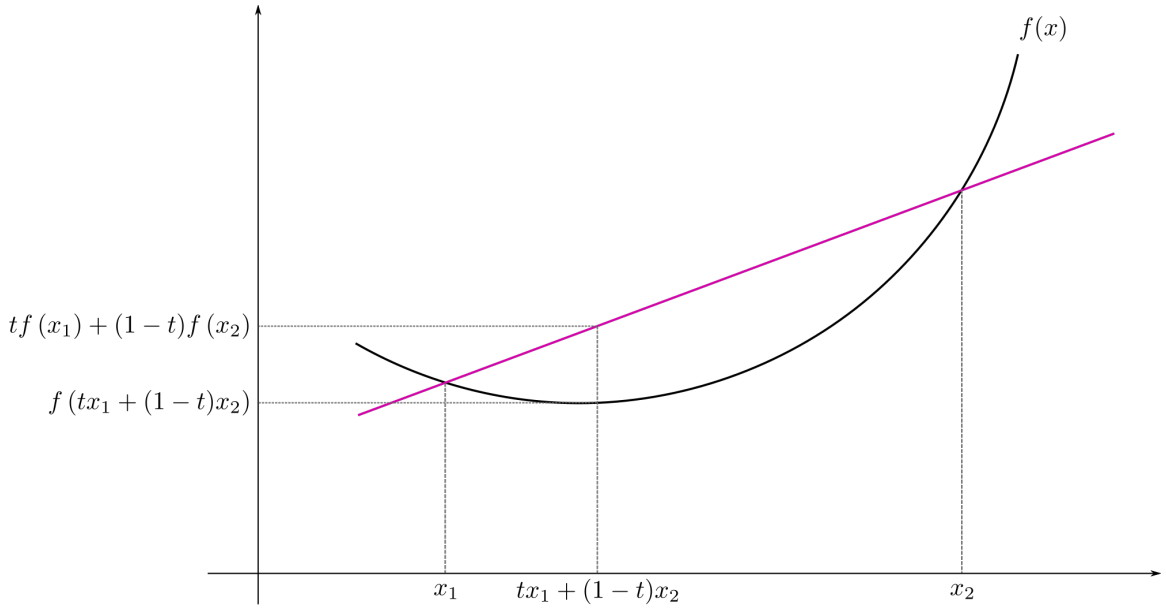
$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ và với mọi $t \in [0, 1]$.

Nếu bất đẳng thức ngặt hơn (tức là $<$ thay vì $\leq$) với mọi $t \in (0, 1)$ và $\mathbf{x} \neq \mathbf{y}$, thì ta nói rằng f là **lồi ngặt (strictly convex)**.

Một hàm f là **lồi mạnh với tham số** m nếu hàm

$$\mathbf{x} \mapsto f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$



Hình 2: Hàm lỗi trông như thế nào?

là hàm lồi.

Những điều kiện này được đưa ra theo thứ tự tăng dần; lồi mạnh dẫn đến lồi ngặt, lồi ngặt dẫn đến lồi.

Về mặt hình học, độ lồi có nghĩa là đoạn thẳng giữa hai điểm trên đồ thị f nằm trên hoặc trên chính đồ thị. Nhìn hình 2 để hình dung rõ.

Độ lồi ngặt có nghĩa là đồ thị của f nằm hoàn toàn phía trên đoạn thẳng, ngoại trừ ở các điểm cuối (Vì vậy, thực sự hàm trong hình có vẻ là hàm lồi ngặt.)

6.8.3 Một số hệ quả của tính lồi

Tại sao chúng ta lại quan tâm đến việc một hàm là lồi (lồi ngặt/lồi mạnh)?

Về cơ bản, các khái niệm khác nhau của chúng ta về độ lồi có ý nghĩa về bản chất của cực tiểu. Không có gì ngạc nhiên khi các điều kiện mạnh hơn cho chúng ta biết nhiều hơn về cực tiểu.

Mệnh đề 12. *Cho $\mathcal{X}$ là một tập lồi. Nếu f là hàm lồi, thì mọi cực tiểu địa phương của f trong $\mathcal{X}$ cũng là cực tiểu toàn cục.*

Chứng minh. Giả sử f là lồi, và cho $\mathbf{x}^*$ là một cực tiểu địa phương của f trong $\mathcal{X}$. Thì với một số lân cận $N \subseteq \mathcal{X}$ của $\mathbf{x}^*$, ta có $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ với mọi $\mathbf{x} \in N$. Giả sử tồn tại mâu thuẫn $\tilde{\mathbf{x}} \in \mathcal{X}$ sao cho $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$.

Xét đoạn thẳng $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, lưu ý rằng $\mathbf{x}(t) \in \mathcal{X}$ do tính lồi $\mathcal{X}$. Thì với f là hàm lồi,

$$f(\mathbf{x}(t)) \leq tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) < tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

với mọi $t \in (0,1)$.

Chúng ta có thể chọn t đủ gần 1 để $\mathbf{x}(t) \in N$; thì $f(\mathbf{x}(t)) \geq f(\mathbf{x}^*)$ theo định nghĩa của N, nhưng $f(\mathbf{x}(t)) < f(\mathbf{x}^*)$ theo bất đăng thức bên trên (mâu thuẫn).

Như vậy $f(\mathbf{x}^*) \leq f(\mathbf{x})$ với mọi $\mathbf{x} \in \mathcal{X}$, thì $\mathbf{x}^*$ là cực tiểu toàn cục của f in $\mathcal{X}$. □

Mệnh đề 13. *Cho $\mathcal{X}$ là một tập lồi. Nếu f là lồi ngặt, thì tồn tại ít nhất là một cực tiểu địa phương f trên miền $\mathcal{X}$. Do đó, nếu tồn tại thì đó là cực tiểu toàn cục duy nhất f in $\mathcal{X}$.*

Chứng minh. Ý thứ hai theo ý đầu tiên, vì vậy tất cả những gì chúng ta phải chỉ ra là nếu tồn tạimột cực tiểu địa phương trong $\mathcal{X}$ thì nó là duy nhất.

Giả sử $\mathbf{x}^*$ là một cực tiểu địa phương của f in $\mathcal{X}$, và giả sử là tồn tại một cực tiểu địa phương $\tilde{\mathbf{x}} \in \mathcal{X}$ sao cho $\tilde{\mathbf{x}} \neq \mathbf{x}^*$.

Vì f là lồi ngặt, nên nó lồi, nên $\mathbf{x}^*$ và $\tilde{\mathbf{x}}$ đều là cực tiểu toàn cục củaf in $\mathcal{X}$ theo định lý phía trước. Do đó $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$. Xét đoạn thẳng $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, nằm hoàn toàn trong $\mathcal{X}$. Theo tính lồi ngặt của f,

$$f(\mathbf{x}(t)) < tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) = tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

với mọi $t \in (0,1)$. Nhưng điều này mâu thuẫn với thực tế là $\mathbf{x}^*$ là một cực tiểu toàn cục. Vì vậy nếu $\tilde{\mathbf{x}}$ là cực tiểu địa phương của f in $\mathcal{X}$, thì $\tilde{\mathbf{x}} = \mathbf{x}^*$, nên $\mathbf{x}^*$ là cực tiểu duy nhất trong $\mathcal{X}$. □

Điều đáng nói ở đây là phải kiểm tra xem tập chấp nhận được ảnh hưởng đến vấn đề tối ưu hóa như thế nào. Chúng ta sẽ thấy lý do tại sao giả định rằng $\mathcal{X}$ là cần lồi trong các kết quả trên.

Xét hàm số $f(x) = x^2$, là một hàm lồi ngặt. Điểm cực tiểu duy nhất của hàm này trong $\mathbb{R}$ là $x = 0$. Nhưng hãy xem điều gì sẽ xảy ra khi chúng ta thay đổi tập chấp nhận được $\mathcal{X}$.

(i) $\mathcal{X} = \{1\}$: Tập hợp này thực sự là lồi, vì vậy chúng ta vẫn có một cực tiểu duy nhất. Nhưng nó không giống như là cực tiểu hóa không ràng buộc!

(ii) $\mathcal{X} = \mathbb{R} \setminus \{0\}$: Không phải là một tập lồi, và chúng ta có thể thấy rằng f không có cực trị $\mathcal{X}$. Với bất kỳ $x \in \mathcal{X}$, ta cũng có thể tìm một điểm khác $y \in \mathcal{X}$ sao cho $f(y) < f(x)$.

(iii) $\mathcal{X} = (-\infty, -1] \cup [0, \infty)$: Không phải là một tập lồi, và chúng ta có một cực tiểu địa phương $(x = -1)$, khác với điểm cực tiểu toàn cục $(x = 0)$.

(iv) $\mathcal{X} = (-\infty, -1] \cup [1, \infty)$: Không phải là một tập lồi, và chúng ta có thể thấy rằng có hai điểm cực tiểu toàn cục là $(x = \pm 1)$.

6.8.4 Chứng minh một hàm là hàm lồi

Hy vọng rằng phần trước đã thuyết phục người đọc rằng độ lồi là một tính chất quan trọng. Tiếp theo, chúng ta chuyển sang vấn đề chỉ ra rằng một hàm là lồi (lồi ngặt/lồi mạnh). Tất nhiên là có thể (về nguyên tắc) trực tiếp chỉ ra rằng điều kiện trong định nghĩa là đúng, nhưng đây thường không phải là cách dễ nhất.

Mệnh đề 14. *Các **chuẩn** là lồi.*

Chứng minh. Cho $\|\cdot\|$ là một chuẩn trên không gian vector V. Thì với mọi $\mathbf{x}, \mathbf{y} \in V$ và $t \in [0,1]$,

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\| = |t|\|\mathbf{x}\| + |1-t|\|\mathbf{y}\| = t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

trong đó chúng ta đã sử dụng bất đẳng thức tam giác tương ứng, tính đồng nhất của các chuẩn và thực tế là t và $1-t$ là không âm.. Do đó $\|\cdot\|$ là lồi. □

Mệnh đề 15. *Giả sử f là khả vi. Thì f là lồi khi và chỉ khi*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

Chứng minh. Tự chứng minh. □

Mệnh đề 16. *Giả sử f khả vi hai lần. Thì*

(i) f là lồi nếu và chỉ nếu $\nabla^2 f(\mathbf{x}) \succeq 0$ với mọi $\mathbf{x} \in \operatorname{dom} f$.

(ii) Nếu $\nabla^2 f(\mathbf{x}) \succ 0$ với mọi $\mathbf{x} \in \operatorname{dom} f$, thì f là lồi ngặt.

(iii) f là m-lồi mạnh nếu và chỉ nếu $\nabla^2 f(\mathbf{x}) \succeq mI$ với mọi $\mathbf{x} \in \operatorname{dom} f$.

Chứng minh. Bỏ qua. □

Mệnh đề 17. *Nếu f là lồi và $\alpha \geq 0$, thì αf cũng lồi.*

Chứng minh. Giả sử f là lồi và $\alpha \geq 0$. Thì với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(\alpha f) = \operatorname{dom} f$,

$$\begin{aligned}
(\alpha f)(t\mathbf{x} + (1-t)\mathbf{y}) &= \alpha f(t\mathbf{x} + (1-t)\mathbf{y}) \\
&\leq \alpha \left(t f(\mathbf{x}) + (1-t) f(\mathbf{y}) \right) \\
&= t(\alpha f(\mathbf{x})) + (1-t)(\alpha f(\mathbf{y})) \\
&= t(\alpha f)(\mathbf{x}) + (1-t)(\alpha f)(\mathbf{y})
\end{aligned}$$

nên αf lồi. □

Mệnh đề 18. *Nếu f và g là các hàm lồi, thì $f + g$ là lồi. Hơn nữa, nếu g là lồi ngặt, thì $f + g$ là lồi ngặt, và nếu g là lồi mạnh với tham số m, thì $f + g$ cúng lồi mạnh với tham số m.*

Chứng minh. Giả sử f và g là lồi. Thì với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f+g) = \operatorname{dom} f \cap \operatorname{dom} g$,

$$\begin{aligned}
(f+g)(t\mathbf{x} + (1-t)\mathbf{y}) &= f(t\mathbf{x} + (1-t)\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) \\
&\leq t f(\mathbf{x}) + (1-t) f(\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) && \text{vì } f \text{ là hàm lồi} \\
&\leq t f(\mathbf{x}) + (1-t) f(\mathbf{y}) + t g(\mathbf{x}) + (1-t) g(\mathbf{y}) && \text{vì } g \text{ là hàm lồi} \\
&= t(f(\mathbf{x}) + g(\mathbf{x})) + (1-t)(f(\mathbf{y}) + g(\mathbf{y})) \\
&= t(f+g)(\mathbf{x}) + (1-t)(f+g)(\mathbf{y})
\end{aligned}$$

nên $f + g$ là lồi.

Nếu g là lồi ngặt, thì bất đẳng thức thứ hai là ngặt với $\mathbf{x} \neq \mathbf{y}$ và $t \in (0,1)$, nên $f + g$ là lồi ngặt.

Nếu g là lồi mạnh với tham số m, thì hàm số $h(\mathbf{x}) \equiv g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ là lồi, nên $f + h$ là lồi. Nhưng

$$(f+h)(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2 \equiv (f+g)(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

nên $f + g$ là lồi mạnh với tham số m. □

Mệnh đề 19. *Nếu $f_1, \ldots, f_n$ là các hàm lồi và $\alpha_1, \ldots, \alpha_n \geq 0$, thì*

$$\sum_{i=1}^{n} \alpha_i f_i$$

là lồi.

Chứng minh. Chứng minh từ hai hệ quả trước bằng cách quy nạp. □

Mệnh đề 20. *Nếu f là lồi, thì $g(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x} + \mathbf{b})$ là lồi đối với bất kỳ kích thước thích hợp nào $\mathbf{A}$ và $\mathbf{b}$.*

Chứng minh. Giả sử f là lồi và g được định nghĩa như trên. Thì với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom} g$,

$$\begin{aligned}
g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(\mathbf{A}(t\mathbf{x} + (1-t)\mathbf{y}) + \mathbf{b}) \\
&= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + \mathbf{b}) \\
&= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + t\mathbf{b} + (1-t)\mathbf{b}) \\
&= f(t(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\
&\leq tf(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)f(\mathbf{A}\mathbf{y} + \mathbf{b}) && \text{vì } f \text{ là hàm lồi} \\
&= tg(\mathbf{x}) + (1-t)g(\mathbf{y})
\end{aligned}$$

Như vậy g là lồi. □

Mệnh đề 21. *Nếu f và g lồi, thì $h(\mathbf{x}) \equiv \max\{f(\mathbf{x}), g(\mathbf{x})\}$ là lồi.*

Chứng minh. Giả sử f và g là lồi và h được định nghĩa như trên. Thì với mọi $\mathbf{x}, \mathbf{y} \in \operatorname{dom} h$,

$$\begin{aligned}
h(t\mathbf{x} + (1-t)\mathbf{y}) &= \max\{f(t\mathbf{x} + (1-t)\mathbf{y}), g(t\mathbf{x} + (1-t)\mathbf{y})\} \\
&\leq \max\{tf(\mathbf{x}) + (1-t)f(\mathbf{y}), tg(\mathbf{x}) + (1-t)g(\mathbf{y})\} \\
&\leq \max\{tf(\mathbf{x}), tg(\mathbf{x})\} + \max\{(1-t)f(\mathbf{y}), (1-t)g(\mathbf{y})\} \\
&= t\max\{f(\mathbf{x}), g(\mathbf{x})\} + (1-t)\max\{f(\mathbf{y}), g(\mathbf{y})\} \\
&= th(\mathbf{x}) + (1-t)h(\mathbf{y})
\end{aligned}$$

Lưu ý rằng trong bất đẳng thức đầu tiên, chúng ta đã sử dụng tính lồi của f và g cùng với $a \leq c, b \leq d$ dẫn đến $\max\{a, b\} \leq \max\{c, d\}$. Trong bất đẳng thức thứ hai, chúng ta đã sử dụng bất đẳng thức $\max\{a + b, c + d\} \leq \max\{a, c\} + \max\{b, d\}$.

Như vậy h là lồi. □

6.8.5 Một số ví dụ

Một cách tốt để có được trực giác về sự phân biệt giữa hàm lồi, hàm lồi ngặt và hàm lồi mạnh là xem xét các ví dụ mà thuộc tính mạnh hơn không giữ được.

Các hàm lồi nhưng không lồi ngặt:

(i) $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \alpha$ với bất kỳ $\mathbf{w} \in \mathbb{R}^d, \alpha \in \mathbb{R}$. Một hàm như vậy được gọi là **hàm affine**, và nó vừa lồi vừa lõm. (Thực tế, một hàm là affine nếu và chỉ khi nó vừa lồi vừa lõm.) Lưu ý rằng hàm tuyến tính và hàm hằng là những trường hợp đặc biệt của hàm affine.

(ii) $f(\mathbf{x}) = \|\mathbf{x}\|_1$

Các hàm lồi ngặt nhưng không lồi mạnh:

(i) $f(x) = x^4$. Ví dụ này rất thú vị vì nó lồi ngặt nhưng bạn không thể hiển thị sự thật này thông qua đối số bậc hai (vì $f''(0) = 0$).

(ii) $f(x) = \exp(x)$. Ví dụ này rất thú vị vì nó được giới hạn bên dưới nhưng không có giá trị cực tiểu địa phương.

(iii) $f(x) = -\log x$. Ví dụ này rất thú vị vì nó lồi ngặt nhưng không bị giới hạn bên dưới.

Các hàm lồi mạnh:

(i) $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$

7 Xác suất

Lý thuyết xác suất cung cấp các công cụ mạnh mẽ để mô hình hóa và giải quyết với điều không chắc chắn.

7.1 Khái niệm cơ bản

Giả sử chúng ta có một số loại thử nghiệm ngẫu nhiên (ví dụ như tung đồng xu) có một tập hợp cố định các kết quả có thể xảy ra. Tập hợp này được gọi là **không gian mẫu** và được ký hiệu là Ω.

Chúng ta muốn xác định xác suất cho một số **biến cố**, là các tập con của Ω. Tập hợp các sự kiện được ký hiệu $\mathcal{F}$. Phần **bù** của biến cố A là một biến cố khác, $A^{\mathrm{c}} = \Omega \setminus A$.

Sau đó, chúng ta có thể định nghĩa một **độ đo xác xuất** $\mathbb{P} : \mathcal{F} \to [0, 1]$ thỏa mãn

(i) $\mathbb{P}(\Omega) = 1$

(ii) **Cộng tính đếm được**: với bất kì tập rời rạc đếm được $\{A_i\} \subseteq \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$$

Bộ ba $(\Omega, \mathcal{F}, \mathbb{P})$ được gọi là **không gian xác suất**.[4]

Nếu $\mathbb{P}(A) = 1$, ta nói A **gần như chắc chắn xảy ra**. , ngược lại A **gần như không xảy ra** $\mathbb{P}(A) = 0$.

Từ những tiên đề này, một số quy tắc hữu ích có thể được rút ra.

Mệnh đề 22. *Cho A là một biến cố. Thì*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(ii) Nếu B là một biến cố và $B \subseteq A$, thì $\mathbb{P}(B) \leq \mathbb{P}(A)$.

(iii) $0 = \mathbb{P}(\varnothing) \leq \mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$

Chứng minh. (i) Dùng cộng tính đếm được của $\mathbb{P}$, ta có

$$\mathbb{P}(A) + \mathbb{P}(A^{\mathrm{c}}) = \mathbb{P}(A \dot\cup A^{\mathrm{c}}) = \mathbb{P}(\Omega) = 1$$

Để chứng minh (ii), giả sử $B \in \mathcal{F}$ và $B \subseteq A$. Thì

$$\mathbb{P}(A) = \mathbb{P}(B \dot\cup (A \setminus B)) = \mathbb{P}(B) + \mathbb{P}(A \setminus B) \geq \mathbb{P}(B)$$

được chứng minh.

Với (iii): bất đẳng thức giữa từ (ii) vì $\varnothing \subseteq A \subseteq \Omega$. Ta cũng có

$$\mathbb{P}(\varnothing) = \mathbb{P}(\varnothing \dot\cup \varnothing) = \mathbb{P}(\varnothing) + \mathbb{P}(\varnothing)$$

bằng cộng tính đếm được, ta chứng minh được $\mathbb{P}(\varnothing) = 0$. □

[4]Một không gian xác xuất đơn giản chỉ là một không gian độ đo, trong đó độ đo của toàn bộ không gian bằng 1.

Mệnh đề 23. *Nếu A và B là hai biến cố, thì $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Chứng minh. Điều quan trọng là chia các biến cố thành nhiều phần khác nhau và không chồng chéo.

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B) \dot\cup (A \setminus B) \dot\cup (B \setminus A)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\end{aligned}$$

□

Mệnh đề 24. *Nếu $\{A_i\} \subseteq \mathcal{F}$ là một tập các biến cố có thể đếm được, độc lập hoặc không, thì*

$$\mathbb{P}\Big(\bigcup_i A_i\Big) \leq \sum_i \mathbb{P}(A_i)$$

Bất đẳng thức này đôi khi được gọi là **bất đẳng thức Boole**.

Chứng minh. Định nghĩa $B_1 = A_1$ và $B_i = A_i \setminus (\bigcup_{j<i} A_j)$ với $i > 1$, lưu ý rằng $\bigcup_{j\leq i} B_j = \bigcup_{j\leq i} A_j$ với mọi i và B_i là độc lập. Thì

$$\mathbb{P}\Big(\bigcup_i A_i\Big) = \mathbb{P}\Big(\bigcup_i B_i\Big) = \sum_i \mathbb{P}(B_i) \leq \sum_i \mathbb{P}(A_i)$$

trong đó bất đẳng thức cuối cùng theo sau bởi tính đơn điệu do $B_i \subseteq A_i$ với mọi i. □

7.1.1 Xác suất có điều kiện

Xác suất có điều kiện (conditional probability) của biến cố A phụ thuộc vào biến cố B đã xảy ra được ký hiệu $\mathbb{P}(A|B)$ và được định nghĩa bởi

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

với $\mathbb{P}(B) > 0$.[5]

7.1.2 Quy tắc chuỗi

Một công cụ rất hữu ích khác, **quy tắc chuỗi**, theo sau ngay từ định nghĩa này:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

[5]Trong một số trường hợp, có thể xác định xác suất có điều kiện trên các biến cố có xác suất bằng không, nhưng điều này mang tính kỹ thuật cao hơn đáng kể nên chúng ta bỏ qua nó.

7.1.3 Quy tắc Bayes

Biến đổi biểu thức ở trên thêm một bước nữa, chúng ta có một công thức đơn giản nhưng quan trọng là **quy tắc Bayes**:

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

Đôi khi có lợi nếu bỏ qua hằng số chuẩn hóa và viết

$$\mathbb{P}(A|B) \propto \mathbb{P}(A)\mathbb{P}(B|A)$$

Theo công thức này, $\mathbb{P}(A)$ thường được gọi là **xác suất tiên nghiệm (prior)**, $\mathbb{P}(A|B)$ thường được gọi là **xác suất hậu nghiệm (posterior)**, và $\mathbb{P}(B|A)$ là **xác suất khả năng xảy ra (likelihood)**.

7.2 Biến ngẫu nhiên

Một **biến ngẫu nhiên** là đại lượng nào đó không chắc chắn với phân phối xác suất liên kết trên các giá trị có thể của biến.

Về mặt hình thức, một biến ngẫu nhiên trên không gian xác suất $(\Omega, \mathcal{F}, \mathbb{P})$ là một hàm[6] $X : \Omega \to \mathbb{R}$.[7]

Chúng ta ký hiệu miền X bởi $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. Để đưa ra một ví dụ cụ thể (lấy từ [1]), giả sử X là số lần mặt ngửa trong hai lần tung của một đồng xu đồng chất. Không gian mẫu là

$$\Omega = \{hh, tt, ht, th\}$$

và X được xác định hoàn toàn bởi kết quả ω, i.e. $X = X(\omega)$. Ví dụ, biến cố $X = 1$ là tập hợp các kết quả $\{ht, th\}$.

Người ta thường nói về các giá trị của một biến ngẫu nhiên mà không tham chiếu trực tiếp đến không gian mẫu của nó. Cả hai có liên quan với nhau theo định nghĩa sau: biến cố mà giá trị của X phụ thuộc vào một số tập $S \subseteq \mathbb{R}$ là

$$X \in S = \{\omega \in \Omega : X(\omega) \in S\}$$

Lưu ý rằng các trường hợp đặc biệt của định nghĩa này bao gồm X bằng, nhỏ hơn hoặc lớn hơn một số giá trị được chỉ định. Ví dụ,

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

Một chút chú ý về ký hiệu: chúng ta viết $p(X)$ biểu thị cho toàn bộ phân phối xác suất của X và $p(x)$ cho đánh giá của hàm p tại một giá trị cụ thể $x \in X(\Omega)$. Hy vọng rằng việc lạm dụng ký hiệu (phù hợp với quy chuẩn) này không quá gây mất tập trung với bạn đọc. Nếu p được tham số hóa bởi một số tham số θ, ta viết $p(X; \theta)$ hoặc $p(x; \theta)$, trừ khi chúng ta đang ở trong một quy ước Bayes, nơi các tham số được coi là một biến ngẫu nhiên, trong trường hợp đó, chúng ta có các điều kiện về các tham số.

7.2.1 Hàm phân phối tích lũy

Hàm phân phối tích lũy (c.d.f.) cho ta xác suất lớn nhất mà một biến ngẫu nhiên có khả năng xảy ra tại một giá trị nhất định:

$$F(x) = \mathbb{P}(X \leq x)$$

[6] Hàm này phải đo được.

[7] Nói chung, tên miền có thể là bất kỳ không gian có thể đo được, nhưng $\mathbb{R}$ là trường hợp phổ biến nhất cho đến nay và đủ cho các mục đích của chúng ta.

Hàm phân phối tích lũy có thể được sử dụng để đưa ra xác suất một biến nằm trong một phạm vi nhất định:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a)$$

7.2.2 Các biến ngẫu nhiên rời rạc

Một **biến ngẫu nhiên rời rạc** là một biến ngẫu nhiên có phạm vi đếm được và giả định mỗi giá trị trong phạm vi này với xác suất dương. Các biến ngẫu nhiên rời rạc được xác định hoàn toàn bởi **hàm khối xác suất** (p.m.f.) $p : X(\Omega) \to [0, 1]$ thỏa mãn

$$\sum_{x \in X(\Omega)} p(x) = 1$$

Đối với X rời rạc, xác suất của một giá trị cụ thể được đưa ra chính xác bởi p.m.f của nó:

$$\mathbb{P}(X = x) = p(x)$$

7.2.3 Biến ngẫu nhiên liên tục

Một **biến ngẫu nhiên liên tục** là một biến ngẫu nhiên có phạm vi không đếm được và giả định mỗi giá trị trong phạm vi này có xác suất bằng không. Hầu hết các biến ngẫu nhiên liên tục mà ta sẽ gặp trong thực tế là **các biến ngẫu nhiên hoàn toàn liên tục**[8], có nghĩa là tồn tại một hàm $p : \mathbb{R} \to [0, \infty)$ thỏa mãn

$$F(x) \equiv \int_{-\infty}^{x} p(z)\,\mathrm{d}z$$

Hàm p được gọi là **hàm mật độ xác suất** (p.d.f.) và phải thỏa mãn

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Các giá trị của hàm này không phải là xác suất, vì chúng có thể vượt quá 1. Tuy nhiên, họ có một vài cách giải thích hợp lý. Một là xác suất tương đối; mặc dù xác suất của từng giá trị cụ thể được chọn về mặt kỹ thuật là 0, một số điểm vẫn có khả năng xảy ra cao hơn những điểm khác.

Người ta cũng có thể coi mật độ là xác định xác suất mà biến sẽ nằm trong một phạm vi nhỏ về một giá trị nhất định. Điều này là do, với $\epsilon > 0$ và nhỏ,

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} p(z)\,\mathrm{d}z \approx 2\epsilon p(x)$$

sử dụng xấp xỉ gần đúng điểm giữa cho tích phân.

Dưới đây là một số nhận dạng hữu ích tiếp theo từ các định nghĩa ở trên:

$$\mathbb{P}(a \leq X \leq b) = \int_{a}^{b} p(x)\,\mathrm{d}x$$
$$p(x) = F'(x)$$

[8] Các biến ngẫu nhiên liên tục nhưng không liên tục hoàn toàn được gọi là **biến ngẫu nhiên kỳ dị**. Chúng ta sẽ không thảo luận về chúng, giả sử đúng hơn là tất cả các biến ngẫu nhiên liên tục thừa nhận một hàm mật độ.

7.3 Phân phối đồng thời

Thường thì chúng ta có một số biến ngẫu nhiên và chúng ta muốn nhận được phân phối đồng thời của chúng. Một **phân phối xác suất đồng thời** chính là như thế. Đối với một số biến ngẫu nhiên $X_1, \ldots, X_n$, phân phối đồng thời được viết $p(X_1, \ldots, X_n)$ và cho ta biết xác xuất khi tất cả X_i xảy ra cùng lúc.

7.3.1 Tính độc lập của các biến ngẫu nhiên

Chúng ta nói rằng hai biến X và Y là **độc lập** nếu phân phối xác suất đồng thời của chúng bằng tích các phân phối xác suất thành phần,

$$p(X, Y) = p(X)p(Y)$$

Chúng ta cũng có thể xác định tính độc lập cho nhiều hơn hai biến ngẫu nhiên, mặc dù nó phức tạp hơn. Cho $\{X_i\}_{i \in I}$ là một tập hợp các biến ngẫu nhiên được lập chỉ mục bởi I, có thể là vô cùng. Thì $\{X_i\}$ độc lập nếu với mọi tập hợp con hữu hạn của chỉ số $i_1, \ldots, i_k \in I$ ta có

$$p(X_{i_1}, \ldots, X_{i_k}) = \prod_{j=1}^{k} p(X_{i_j})$$

Ví dụ, trong trường hợp ba biến ngẫu nhiên, X, Y, Z, chúng ta đòi hỏi $p(X, Y, Z) = p(X)p(Y)p(Z)$ cũng như $p(X, Y) = p(X)p(Y)$, $p(X, Z) = p(X)p(Z)$, và $p(Y, Z) = p(Y)p(Z)$.

Việc giả định rằng một loạt các biến ngẫu nhiên là **độc lập và phân phối giống nhau** (i.i.d.) như vậy để phân phối đồng thời của chúng có thể được tính toàn bộ:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i)$$

trong đó $X_1, \ldots, X_n$ tất cả đều giống nhau về p.m.f./p.d.f.

7.3.2 Phân phối biên

Nếu chúng ta có một phân phối đồng thời trên một số tập hợp các biến ngẫu nhiên, có thể có được một phân phối cho một tập hợp con của chúng bằng cách "tính tổng" (hoặc "tích phân" trong trường hợp liên tục) các biến mà chúng ta không quan tâm đến:

$$p(X) = \sum_{y} p(X, y)$$

7.4 Kỳ vọng

Nếu chúng ta có một số biến ngẫu nhiên X, chúng ta có thể muốn biết giá trị "trung bình " của X là gì. Khái niệm này được gọi là **giá trị kỳ vọng** (hay **trung bình**) $\mathbb{E}[X]$, được định nghĩa là

$$\mathbb{E}[X] = \sum_{x \in X(\Omega)} x p(x)$$

đối với biến X rời rạc,

và có giá trị là:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$$

đối với biến X liên tục.

Nói cách khác, chúng ta đang lấy tổng trọng số của các giá trị mà X có thể nhận, trong đó trọng số là xác suất của các giá trị tương ứng đó. Giá trị kỳ vọng có cách hiểu vật lý là "khối tâm" ("center of mass") của phân bố.

7.4.1 Tính chất của giá trị kỳ vọng

Một tính chất rất hữu ích của kỳ vọng là tính tuyến tính:

$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha_i X_i + \beta\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] + \beta$$

Lưu ý rằng điều này đúng ngay cả khi X_i không độc lập!

Nhưng nếu chúng độc lập, ta có quy tắc nhân:

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]$$

7.5 Phương sai

Sự kỳ vọng cung cấp một thước đo về "trung tâm" của một phân phối, nhưng thường thì chúng ta cũng quan tâm đến những gì "xung quanh" về trung tâm đó. Chúng ta định nghĩa phương sai $\mathrm{Var}(X)$ của một biến ngẫu nhiên X bởi

$$\mathrm{Var}(X) = \mathbb{E}\left[\big(X - \mathbb{E}[X]\big)^2\right]$$

Nói cách khác, đây là độ lệch bình phương trung bình của các giá trị X so với giá trị trung bình của X. Sử dụng một chút đại số và tuyến tính của kỳ vọng, thật dễ dàng để chỉ ra rằng

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

7.5.1 Tính chất của phương sai

Phương sai không phải là tuyến tính (vì xuất hiện bình phương trong định nghĩa), nhưng người ta có thể chỉ ra những điều sau:

$$\mathrm{Var}(\alpha X + \beta) = \alpha^2 \,\mathrm{Var}(X)$$

Về cơ bản, các hằng số nhân trở thành bình phương khi chúng được kéo ra và các hằng số cộng biến mất (vì phương sai do một hằng số đóng góp bằng 0).

Hơn nữa, nếu $X_1, \ldots, X_n$ không tương quan[9], thì

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$$

[9]Chúng ta vẫn chưa xác định điều này; xem phần Tương quan bên dưới

7.5.2 Độ lệch chuẩn

Phương sai là một khái niệm hữu ích, nhưng nó mắc phải thực tế là các đơn vị của phương sai không giống với các đơn vị của biến ngẫu nhiên (một lần nữa vì bình phương). Để giải quyết vấn đề này, chúng ta có thể sử dụng **độ lệch chuẩn**, được định nghĩa là $\sqrt{\text{Var}(X)}$. Độ lệch chuẩn của X có cùng đơn vị với X.

7.6 Hiệp phương sai

Hiệp phương sai là thước đo mối quan hệ tuyến tính giữa hai biến ngẫu nhiên. Chúng ta định nghĩa hiệp phương sai giữa X và Y là $\text{Cov}(X, Y)$, và được xác định

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Lưu ý rằng kỳ vọng bên ngoài phải được thực hiện trên sự phân phối đồng thời của X và Y.

Một lần nữa, tính tuyến tính của kỳ vọng cho phép chúng ta viết lại điều này dưới dạng

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

So sánh các công thức này với công thức để tìm phương sai, không khó để thấy rằng $\text{Var}(X) = \text{Cov}(X, X)$.

Một tính chất hữu ích của hiệp phương sai là **song tính**:

$$\text{Cov}(\alpha X + \beta Y, Z) = \alpha \,\text{Cov}(X, Z) + \beta \,\text{Cov}(Y, Z)$$
$$\text{Cov}(X, \alpha Y + \beta Z) = \alpha \,\text{Cov}(X, Y) + \beta \,\text{Cov}(X, Z)$$

7.6.1 Tương quan

Chuẩn hóa hiệp phương sai cho ta khái niệm về **tương quan**:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Tương quan cũng đo lường mối quan hệ tuyến tính giữa hai biến, nhưng không giống như hiệp phương sai luôn nằm giữa -1 và 1.

Hai biến được cho là **không tương quan** nếu $\text{Cov}(X, Y) = 0$ bởi vì $\text{Cov}(X, Y) = 0$ dẫn đến rằng $\rho(X, Y) = 0$. Nếu hai biến độc lập, thì chúng không tương quan, nhưng ngược lại có thể không đúng.

7.7 Vectơ ngẫu nhiên

Cho đến nay chúng ta đã nói về các **phân phối đơn biến (univariate distributions)**, tức là các phân phối của các biến đơn lẻ. Nhưng chúng ta cũng có thể nói về các **phân phối đa biến (multivariate distributions)**, cung cấp các phân phối **vectơ ngẫu nhiên**:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

Các đại lượng tổng hợp mà chúng ta đã thảo luận cho các biến đơn có tổng quát tự nhiên cho trường hợp đa biến.

Kỳ vọng của một vectơ ngẫu nhiên chỉ đơn giản là kỳ vọng được áp dụng cho mỗi thành phần:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

Phương sai được tổng quát bởi **ma trận hiệp phương sai**:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Var}(X_n) \end{bmatrix}$$

Đó là, $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Vì hiệp phương sai là đối xứng trong các đối số của nó, nên ma trận hiệp phương sai cũng là đối xứng. Nó cũng bán xác định dương: với bất kỳ $\mathbf{x}$,

$$\mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x} = \mathbf{x}^\top \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top]\mathbf{x} = \mathbb{E}[\mathbf{x}^\top(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x}] = \mathbb{E}[((\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \mathbf{x})^2] \geq 0$$

Nghịch đảo của ma trận hiệp phương sai, $\boldsymbol{\Sigma}^{-1}$, đôi khi được gọi là **ma trận chính xác**.

7.8 Ước lượng các tham số

Bây giờ chúng ta đi vào một số chủ đề cơ bản từ số liệu thống kê. Chúng ta đưa ra một số giả định về vấn đề của mình bằng cách quy định một mô hình **tham số** (ví dụ: một phân phối mô tả cách dữ liệu được tạo ra), sau đó chúng ta khớp các tham số của mô hình với dữ liệu. Làm thế nào để chúng ta chọn các giá trị của các tham số?

7.8.1 Ước lượng hợp lý cực đại

Một cách phổ biến để điều chỉnh các tham số là **ước lượng hợp lý cực đại** (MLE). Nguyên tắc cơ bản của MLE là chọn các giá trị "giải thích" dữ liệu tốt nhất bằng cách tối đa hóa xác suất / mật độ của dữ liệu mà chúng ta đã xem như một hàm của các tham số. Giả sử chúng ta có các biến ngẫu nhiên $X_1, \dots, X_n$ và các quan sát tương ứng $x_1, \dots, x_n$. Thì

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \mathcal{L}(\theta)$$

trong đó $\mathcal{L}$ là **hàm hợp lý**

$$\mathcal{L}(\theta) = p(x_1, \dots, x_n; \theta)$$

Thông thường, chúng ta cho rằng $X_1, \dots, X_n$ là xác định và độc lập (kí hiệu: i.i.d). Thì chúng ta có thể viết

$$p(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta)$$

Tại thời điểm này, việc lấy *log* thường thuận tiện, dẫn đến **log-likelihood**

$$\log \mathcal{L}(\theta) = \sum_{i=1}^n \log p(x_i; \theta)$$

Đây là một phép toán hợp lệ vì xác suất / mật độ được giả định là dương và vì log là một hàm tăng đơn điệu nên nó bảo toàn thứ tự. Nói cách khác, cực đại hóa $\log \mathcal{L}$ cũng đồng nghĩa với cực đại hóa $\mathcal{L}$.

Đối với một số phân phối, có thể phân tích để tìm ra **ước lượng hợp lý cực đại (maximum likelihood estimator)**. Nếu $\log \mathcal{L}$ khả vi, cho đạo hàm bằng 0 và cố gắng tìm ra θ.

7.8.2 Ước lượng hậu nghiệm cực đại

Một cách Bayes khác để điều chỉnh các tham số là thông qua **ước lượng hậu nghiệm cực đại** (MAP). Trong kỹ thuật này, chúng ta giả định rằng các tham số là một biến ngẫu nhiên và chúng ta chỉ định một phân phối tiên nghiệm $p(\theta)$. Sau đó, chúng ta có thể sử dụng quy tắc Bayes để tính toán phân phối hậu nghiệm của các tham số dựa trên dữ liệu quan sát:

$$p(\theta|x_1, \ldots, x_n) \propto p(\theta)p(x_1, \ldots, x_n|\theta)$$

Việc tính toán **hằng số chuẩn hóa** (là mẫu số của xác suất Bayes) thường khó thực hiện, bởi vì nó liên quan đến việc tích hợp trên không gian tham số, có thể có chiều rất lớn. Thật may mắn, nếu chúng ta chỉ muốn ước tính MAP, chúng ta không quan tâm đến hằng số chuẩn hóa! Nó không ảnh hưởng đến giá trị θ để cực đại hóa xác suất hậu nghiệm. Nên ta có:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta)p(x_1, \ldots, x_n|\theta)$$

Một lần nữa, nếu chúng ta giả sử các quan sát là cùng phân phối, thì chúng ta có thể diễn đạt điều này ở dạng tương đương và có thể là thân thiện hơn:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left(\log p(\theta) + \sum_{i=1}^{n} \log p(x_i|\theta) \right)$$

7.9 Phân phối Gauss (phân phối chuẩn)

Có nhiều bản phân phối, nhưng một trong những phân phối quan trọng đặc biệt là **phân phối Gauss (Gaussian distribution)**, còn được gọi là **phân phối chuẩn (normal distribution)**. Nó là một phân phối liên tục, được tham số hóa bằng giá trị trung bình của nó $\boldsymbol{\mu} \in \mathbb{R}^d$ và ma trận hiệp phương sai xác định dương $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, với mật độ

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right)$$

Lưu ý rằng trong trường hợp đặc biệt $d = 1$, hàm mật độ được viết ở dạng dễ nhận biết hơn

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2} \right)$$

Chúng ta viết $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ để biểu thị rằng $\mathbf{X}$ là phân phối chuẩn với giá trị trung bình $\boldsymbol{\mu}$ và phương sai $\boldsymbol{\Sigma}$.

7.9.1 Hình học của phân phối Gauss đa biến

Hình học của hàm mật độ Gauss đa biến có liên quan mật thiết đến hình học của các dạng bậc hai xác định dương, vì vậy hãy đảm bảo rằng tài liệu trong phần đó được hiểu rõ trước khi giải quyết phần này.

Trước tiên, hãy quan sát rằng hàm mật độ xác suất của Gauss đa biến có thể được viết lại thành

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = g(\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}})$$

trong đó $\tilde{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}$ và $g(z) = [(2\pi)^d \det(\boldsymbol{\Sigma})]^{-\frac{1}{2}} \exp\left(-\frac{z}{2}\right)$. Viết mật độ theo cách này, chúng ta thấy rằng sau khi dịch chuyển theo giá trị trung bình $\boldsymbol{\mu}$, mật độ thực sự chỉ là một hàm đơn giản của dạng bậc hai ma trận của nó.

Đây là một quan sát quan trọng: hàm g là **đơn điệu giảm ngặt**. Vì $g(a) > g(b)$ khi và chỉ khi $a < b$. Do đó, các giá trị nhỏ của $\tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$ (tương ứng với những điểm mà ở đó $\tilde{\mathbf{x}}$ gần $\mathbf{0}$, i.e. $\mathbf{x} \approx \boldsymbol{\mu}$) có mật độ xác suất tương đối cao, và ngược lại. Hơn nữa, bởi vì g là đơn điệu *ngặt*, vì vậy tập cấp c của $p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ là tập cấp $g^{-1}(c)$ của hàm $\mathbf{x} \mapsto \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}}$. Do đó, với bất kỳ c,

$$\{\mathbf{x} \in \mathbb{R}^d : p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = c\} = \{\mathbf{x} \in \mathbb{R}^d : \tilde{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}} = g^{-1}(c)\}$$

Nói cách khác, các hàm này là các đường đồng mức nhưng các giá trị khác nhau.

Nhắc lại tóm tắt về hình học của các dạng bậc hai xác định dương: tập cấp $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ à các ellipsoids sao cho các trục hướng theo các hướng của các vector riêng của $\mathbf{A}$, và độ dài của các trục này tỷ lệ với căn bậc hai nghịch đảo của các giá trị riêng tương ứng. Do đó trong trường hợp này, các đường đồng mức của mật độ là một ellipsoid (tâm tại $\boldsymbol{\mu}$) với độ dài trục tỷ lệ với căn bậc hai nghịch đảo của các giá trị riêng của $\boldsymbol{\Sigma}^{-1}$, hoặc tương đương, căn bậc hai của các giá trị riêng của $\boldsymbol{\Sigma}$.

Tài liệu

[1] J. Pitman, *Probability*. New York: Springer-Verlag, 1993.

[2] S. Axler, *Linear Algebra Done Right (Third Edition)*. Springer International Publishing, 2015.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2009.

[4] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer Science+Business Media, 2006.

[5] J. S. Rosenthal, *A First Look at Rigorous Probability Theory (Second Edition)*. Singapore: World Scientific Publishing, 2006.

[6] J. A. Rice, *Mathematical Statistics and Data Analysis*. Belmont, California: Thomson Brooks/Cole, 2007.

[7] W. Cheney, *Analysis for Applied Mathematics*. New York: Springer Science+Business Medias, 2001.