Vũ Hữu Tiệp

Machine Learning cơ bản

machinelearningcoban.com

Vũ Hữu Tiệp

Machine Learning cơ bản

Order ebook tại *https://machinelearningcoban.com/ebook/*

Blog: *https://machinelearningcoban.com*
Facebook Page: *https://www.facebook.com/machinelearningbasicvn/*
Facebook Group: *https://www.facebook.com/groups/machinelearningcoban/*
Interactive Learning: *https:fundaml.com*

Last update:

May 23, 2018

0.1 Bảng các ký hiệu

Các ký hiệu sử dụng trong sách được liệt kê trong Bảng 0.1

Bảng 0.1: Bảng các ký hiệu

Ký hiệu	Ý nghĩa
x, y, N, k	in nghiêng, thường hoặc hoa, là các số vô hướng
$\mathbf{x}, \mathbf{y}$	in đậm, chữ thường, là các vector
$\mathbf{X}, \mathbf{Y}$	in đậm, chữ hoa, là các ma trận
$\mathbb{R}$	tập hợp các số thực
$\mathbb{N}$	tập hợp các số tự nhiên
$\mathbb{C}$	tập hợp các số phức
$\mathbb{R}^m$	tập hợp các vector thực có m phần tử
$\mathbb{R}^{m\times n}$	tập hợp các ma trận thực có m hàng, n cột
$\mathbb{S}^n$	tập hợp các ma trận vuông đối xứng bậc n
$\mathbb{S}^n_+$	tập hợp các ma trận nửa xác định dương bậc n
$\mathbb{S}^n_{++}$	tập hợp các ma trận xác định dương bậc n
$\in$	phần tử thuộc tập hợp
$\exists$	tồn tại
$\forall$	mọi
$\triangleq$	ký hiệu là/bởi. Ví dụ $a \triangleq f(x)$ nghĩa là "ký hiệu $f(x)$ bởi a".
x_i	phần tử thứ i (tính từ 1) của vector $\mathbf{x}$
$\text{sgn}(x)$	hàm xác định dấu. Bằng 1 nếu $x \geq 0$, bằng -1 nếu $x < 0$.
$\exp(x)$	e^x
$\log(x)$	logarit *tự nhiên* của số thực dương x
a_{ij}	phần tử hàng thứ i, cột thứ j của ma trận $\mathbf{A}$
$\mathbf{A}^T$	chuyển vị của ma trận $\mathbf{A}$
$\mathbf{A}^H$	chuyển vị liên hợp (Hermitian) của ma trận phức $\mathbf{A}$
$\mathbf{A}^{-1}$	nghịch đảo của ma trận vuông $\mathbf{A}$, nếu tồn tại
$\mathbf{A}^\dagger$	giả nghịch đảo của ma trận không nhất thiết vuông $\mathbf{A}$
$\mathbf{A}^{-T}$	chuyển vị của nghịch đảo của ma trận $\mathbf{A}$, nếu tồn tại
$\|\mathbf{x}\|_p$	ℓ_p norm của vector $\mathbf{x}$
$\|\mathbf{A}\|_F$	Frobenius norm của ma trận $\mathbf{A}$
$\text{diag}(\mathbf{A})$	đường chéo chính của ma trận $\mathbf{A}$
$\text{trace}(\mathbf{A})$	trace của ma trận $\mathbf{A}$
$\det(\mathbf{A})$	định thức của ma trận vuông $\mathbf{A}$
$\text{rank}(\mathbf{A})$	hạng của ma trận $\mathbf{A}$
o.w	*otherwise* – trong các trường hợp còn lại
$\frac{\partial f}{\partial x}$	đạo hàm của hàm số f theo $x \in \mathbb{R}$
$\nabla_\mathbf{x} f$	gradient (đạo hàm) của hàm số f theo $\mathbf{x}$ ($\mathbf{x}$ là vector hoặc ma trận)
$\nabla^2_\mathbf{x} f$	đạo hàm bậc hai của hàm số f theo $\mathbf{x}$, còn được gọi là *Hessian*
$\odot$	Hadamard product (elemenwise product). Phép nhân từng phần tử của hai vector hoặc ma trận cùng kích thước.
$\propto$	tỉ lệ với
v.v.	vân vân

Mục lục

Phần I

Kiến thức toán cơ bản cho machine learning

Chương 1

Ôn tập Đại số tuyến tính

1.1 Lưu ý về ký hiệu

Trong các bài viết của tôi, các số vô hướng được biểu diễn bởi các chữ cái viết ở dạng in nghiêng, có thể viết hoa, ví dụ x_1, N, y, k. Các vector được biểu diễn bằng các chữ cái thường in đậm, ví dụ $\mathbf{y}, \mathbf{x}_1$. Nếu không giải thích gì thêm, các vector được mặc định hiểu là các vector cột. Các ma trận được biểu diễn bởi các chữ viết hoa in đậm, ví dụ $\mathbf{X}, \mathbf{Y}, \mathbf{W}$.

Đối với vector, $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ được hiểu là một vector hàng, và $\mathbf{x} = [x_1; x_2; \ldots; x_n]$ được hiểu là vector cột. Chú ý sự khác nhau giữa dấu phẩy (,) và dấu chấm phẩy (;). Đây chính là ký hiệu được Matlab sử dụng. Nếu không giải thích gì thêm, một chữ cái viết thường in đậm được hiểu là một vector cột.

Tương tự, trong ma trận, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ được hiểu là các vector cột $\mathbf{x}_j$ được đặt cạnh nhau theo thứ tự từ trái qua phải để tạo ra ma trận $\mathbf{X}$. Trong khi $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_m]$ được hiểu là các vector $\mathbf{x}_i$ được đặt chồng lên nhau theo thứ tự từ trên xuống dưới để tạo ra ma trận $\mathbf{X}$. Các vector được ngầm hiểu là có kích thước phù hợp để có thể xếp cạnh hoặc xếp chồng lên nhau. Phần tử ở hàng thứ i, cột thứ j được ký hiệu là x_{ij}.

Cho một ma trận $\mathbf{W}$, nếu không giải thích gì thêm, chúng ta hiểu rằng $\mathbf{w}_i$ là **vector cột** thứ i của ma trận đó. Chú ý sự tương ứng giữa ký tự viết hoa và viết thường.

1.2 Chuyển vị và Hermitian

Một toán tử quan trọng của ma trận hay vector là toán tử *chuyển vị* (*transpose*).

Cho $\mathbf{A} \in \mathbb{R}^{m \times n}$, ta nói $\mathbf{B} \in \mathbb{R}^{n \times m}$ là chuyển vị của $\mathbf{A}$ nếu $b_{ij} = a_{ji},\ \forall 1 \leq i \leq n, 1 \leq j \leq m$.

Một cách ngắn gọn, chuyển vị của một ma trận là một ma trận nhận được từ ma trận cũ thông qua phép phản xạ gương qua đường chéo chính của ma trận ban đầu. Toán tử chuyển

vị thường được ký hiểu bởi chữ T, t hoặc ký tự $\top$. Trong cuốn sách này, chúng ta sẽ sử dụng chữ cái T. Ví dụ, chuyển vị của một vector $\mathbf{x}$ được ký hiệu là $\mathbf{x}^T$; chuyển vị của một ma trận $\mathbf{A}$ được ký hiệu là $\mathbf{A}^T$. Cụ thể:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \Rightarrow \mathbf{x}^T = \begin{bmatrix} x_1 \; x_2 \; \dots \; x_m \end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \ddots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

Nếu $\mathbf{A} \in \mathbb{R}^{m \times n}$ thì $\mathbf{A}^T \in \mathbb{R}^{n \times m}$. Nếu $\mathbf{A}^T = \mathbf{A}$, ta nói $\mathbf{A}$ là một *ma trận đối xứng* (*symmetric matrix*).

Trong trường hợp vector hay ma trận có các phần tử là số phức, việc lấy chuyển vị thường đi kèm với việc lấy liên hợp phức. Tức là ngoài việc đổi vị trí của các phần tử, ta còn lấy liên hợp phức của các phần tử đó. Tên gọi của phép toán chuyển vị và lấy liên hợp này còn được gọi là *chuyển vị liên hợp* (*conjugate transpose*), và thường được ký hiệu bằng chữ H thay cho chữ T. Chuyển vị liên hợp của một ma trận $\mathbf{A}$ được ký hiệu là $\mathbf{A}^H$ (cũng được đọc là $\mathbf{A}$ Hermitian).

Cho $\mathbf{A} \in \mathbb{C}^{m \times n}$, ta nói $\mathbf{B} \in \mathbb{C}^{n \times m}$ là chuyển vị liên hợp của $\mathbf{A}$ nếu $b_{ij} = \overline{a_{ji}}, \; \forall 1 \leq i \leq n, 1 \leq j \leq m$, trong đó $\bar{a}$ là liên hiệp phức của a.

Ví dụ:

$$\mathbf{A} = \begin{bmatrix} 1+2i & 3-4i \\ i & 2 \end{bmatrix} \Rightarrow \mathbf{A}^H = \begin{bmatrix} 1-2i & -i \\ 3+4i & 2 \end{bmatrix}; \mathbf{x} = \begin{bmatrix} 2+3i \\ 2i \end{bmatrix} \Rightarrow \mathbf{x}^H = \begin{bmatrix} 2-3i & -2i \end{bmatrix} \quad (1.1)$$

Nếu $\mathbf{A}, \mathbf{x}$ là các ma trận và vector thực thì $\mathbf{A}^H = \mathbf{A}^T, \mathbf{x}^H = \mathbf{x}^T$.

Nếu chuyển vị liên hợp của một ma trận phức bằng với chính nó, $\mathbf{A}^H = \mathbf{A}$, thì ta nói ma trận đó là *Hermitian*.

1.3 Phép nhân hai ma trận

Cho hai ma trận $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$, tích của hai ma trận được ký hiệu là $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times p}$ trong đó phần tử ở hàng thứ i, cột thứ j của ma trận kết quả được tính bởi:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \; \forall 1 \leq i \leq m, 1 \leq j \leq p \quad (1.2)$$

Để nhân được hai ma trận, số cột của ma trận thứ nhất phải bằng số hàng của ma trận thứ hai. Trong ví dụ trên, chúng đều bằng n.

Một vài tính chất của phép nhân hai ma trận (giả sử kích thước các ma trận là phù hợp để các phép nhân ma trận tồn tại):

1. Phép nhân ma trận không có tính chất giao hoán. Thông thường (không phải luôn luôn), $\mathbf{AB} \neq \mathbf{BA}$. Thậm chí, trong nhiều trường hợp, các phép tính này không tồn tại vì kích thước các ma trận lệch nhau.

2. Phép nhân ma trận có tính chất kết hợp: $\mathbf{ABC} = (\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$

3. Phép nhân ma trận có tính chất phân phối đối với phép cộng: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.

4. Chuyển vị của một tích bằng tích các chuyển vị theo thứ tự ngược lại. Điều tương tự xảy ra với Hermitian của một tích:

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T; \quad (\mathbf{AB})^H = \mathbf{B}^H\mathbf{A}^H \tag{1.3}$$

Theo định nghĩa trên, bằng cách coi vector là một trường hợp đặc biệt của ma trận, tích vô hướng của hai vector (*inner product*) $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ được định nghĩa là:

$$\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = \sum_{i=1}^{n} x_i y_i \tag{1.4}$$

Chú ý, $\mathbf{x}^H\mathbf{y} = (\mathbf{y}^H\mathbf{x})^H = \mathbf{y}^H\mathbf{x}$. Chúng bằng nhau khi và chỉ khi chúng là các số thực. Nếu tích vô hướng của hai vector khác không bằng không, hai vector đó vuông góc với nhau.

$\mathbf{x}^H\mathbf{x} \geq 0, \ \forall \mathbf{x} \in \mathbb{C}^n$ vì tích của một số phức với liên hiệp của nó luôn là một số không âm.

Phép nhân của một ma trận $\mathbf{A} \in \mathbb{R}^{m \times n}$ với một vector $\mathbf{x} \in \mathbb{R}^n$ là một vector $\mathbf{b} \in \mathbb{R}^m$:

$$\mathbf{Ax} = \mathbf{b}, \ \text{với } b_i = \mathbf{A}_{:,i}\mathbf{x} \tag{1.5}$$

với $\mathbf{A}_{:,i}$ là vector hàng thứ i của $\mathbf{A}$.

Ngoài ra, một phép nhân khác được gọi là *Hadamard* (hay *element-wise*) hay được sử dụng trong Machine Learning. Tích Hadamard của hai ma trận **cùng kích thước** $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, ký hiệu là $\mathbf{C} = \mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{m \times n}$, trong đó:

$$c_{ij} = a_{ij}b_{ij} \tag{1.6}$$

1.4 Ma trận đơn vị và ma trận nghịch đảo

1.4.1 Ma trận đơn vị

Đường chéo chính của một ma trận là tập hợp các điểm có chỉ số hàng và cột là như nhau. Cách định nghĩa này cũng có thể được định nghĩa cho một ma trận không vuông. Cụ thể, nếu $\mathbf{A} \in \mathbb{R}^{m \times n}$ thì đường chéo chính của $\mathbf{A}$ bao gồm $\{a_{11}, a_{22}, \ldots, a_{pp}\}$, trong đó $p = \min\{m, n\}$.

Một ma trận đơn vị bậc n là một ma trận đặc biệt trong $\mathbb{R}^{n \times n}$ với các phần tử trên đường chéo chính bằng 1, các phần tử còn lại bằng 0. Ma trận đơn vị thường được ký hiệu là $\mathbf{I}$

(identity matrix). Nếu làm việc với nhiều ma trận đơn vị với bậc khác nhau, ta thường ký kiệu $\mathbf{I}_n$ cho ma trận đơn vị bậc n. Dưới đây là ma trận đơn vị bậc 3 và bậc 4:

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1.7}$$

Ma trận đơn vị có tính chất đặc biệt trong phép nhân. Nếu $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times m}$ và $\mathbf{I}$ là ma trận đơn vị bậc n, ta có: $\mathbf{AI} = \mathbf{A}, \quad \mathbf{IB} = \mathbf{B}$.

Với mọi vector $\mathbf{x} \in \mathbb{R}^n$, ta có $\mathbf{I}_n\mathbf{x} = \mathbf{x}$.

1.4.2 Ma trận nghịch đảo

Cho một ma trận vuông $\mathbf{A} \in \mathbb{R}^{n \times n}$, nếu tồn tại ma trận vuông $\mathbf{B} \in \mathbb{R}^{n \times n}$ sao cho $\mathbf{AB} = \mathbf{I}_n$, thì ta nói $\mathbf{A}$ là *khả nghịch* (*invertible, nonsingular* hoặc *nondegenerate*), và $\mathbf{B}$ được gọi là *ma trận nghịch đảo* (*inverse matrix*) của $\mathbf{A}$. Nếu không tồn tại ma trận $\mathbf{B}$ thoả mãn điều kiện trên, ta nói rằng ma trận $\mathbf{A}$ là không khả nghịch (*singular* hoặc *degenerate*).

Nếu $\mathbf{A}$ là khả nghịch, ma trận nghịch đảo của nó thường được ký hiệu là $\mathbf{A}^{-1}$. Ta cũng có:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \tag{1.8}$$

Ma trận nghịch đảo thường được sử dụng để giải hệ phương trình tuyến tính. Giả sử rằng $\mathbf{A} \in \mathbb{R}^{n \times n}$ là một ma trận khả nghịch và một vector bất kỳ $\mathbf{b} \in \mathbb{R}^n$. Khi đó, phương trình:

$$\mathbf{Ax} = \mathbf{b} \tag{1.9}$$

có nghiệm duy nhất là $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Thật vậy, nhân bên trái cả hai vế của phương trình với $\mathbf{A}^{-1}$, ta có $\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Nếu $\mathbf{A}$ không khả nghịch, thậm chí không vuông, phương trình tuyến tính (1.9) có thể không có nghiệm hoặc có vô số nghiệm.

Giả sử các ma trận vuông $\mathbf{A}, \mathbf{B}$ là khả nghịch, khi đó tích của chúng cũng khả nghịch, và $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. Quy tắc này cũng khá giống với cách tính ma trận chuyển vị của tích các ma trận.

1.5 Một vài ma trận đặc biệt khác

1.5.1 Ma trận đường chéo

Ma trận đường chéo (*diagonal matrix*) là ma trận chỉ có các thành phần trên đường chéo chính là khác không. Định nghĩa này cũng có thể được áp dụng lên các ma trận không vuông. Ma trận không (tất cả các phần tử bằng 0) và đơn vị là các ma trận đường chéo. Một vài ví dụ về các ma trận đường chéo $\begin{bmatrix} 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}$.

Với các ma trận đường chéo vuông, thay vì viết cả ma trận, ta có thể chỉ liệt kê các thành phần trên đường chéo. Ví dụ, một ma trận đường chéo vuông $\mathbf{A} \in \mathbb{R}^{m \times m}$ có thể được ký hiệu là $\text{diag}(a_{11}, a_{22}, \ldots, a_{mm})$ với a_{ii} là phần tử hàng thứ i, cột thứ i của ma trận $\mathbf{A}$.

Tích, tổng của hai ma trận đường chéo vuông cùng bậc là một ma trận đường chéo. Một ma trận đường chéo vuông là khả nghịch nếu và chỉ nếu mọi phần tử trên đường chéo chính là khác không. Nghịch đảo của một ma trận đường chéo khả nghịch cũng là một ma trận đường chéo. Cụ thể hơn, $(\text{diag}(a_1, a_2, \ldots, a_n))^{-1} = \text{diag}(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1})$.

1.5.2 Ma trận tam giác

Một ma trận vuông được gọi là *ma trận tam giác trên* (*upper triangular matrix*) nếu tất cả các thành phần *nằm phía dưới đường chéo chính bằng 0*. Tương tự, một ma trận vuông được gọi là *ma trận tam giác dưới* (*lower triangular matrix*) nếu tất cả các thành phần *nằm phía trên đường chéo chính bằng 0*.

Các hệ phương trình tuyến tính mà ma trận hệ số có dạng tam giác thường được quan tâm vì chúng có thể được giải với chi phí tính toán thấp (*low computational cost*). Xét hệ:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n = b_1 \\ a_{22}x_2 + \cdots + a_{2,n-1}x_{n-2} + a_{2n}x_n = b_2 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ a_{nn}x_n = b_n \end{cases} \tag{1.10}$$

Hệ này có thể được viết gọn dưới dạng $\mathbf{Ax} = \mathbf{b}$ với $\mathbf{A}$ là một ma trận tam giác trên. Nhận thấy rằng phương trình này có thể giải mà không cần tính ma trận nghịch đảo $\mathbf{A}^{-1}$ (quá trình tính ma trận nghịch đảo thường tốn khá nhiều thời gian), thay vào đó, ta có thể giải x_n dựa vào phương trình cuối cùng. Sau khi có x_n, ta có thể thay nó vào phương trình gần cuối để suy ra x_{n-1}. Tiếp tục quá trình này, ta sẽ có nghiệm cuối cùng $\mathbf{x}$. Quá trình giải từ cuối lên đầu và thay toàn bộ các thành phần đã tìm được vào phương trình hiện tại được gọi là *back substitution*. Nếu ma trận hệ số là một ma trận tam giác dưới, hệ phương trình có thể được giải bằng một quá trình ngược lại – lần lượt tính x_1 rồi $x_2, \ldots, x_n$, quá trình này được gọi là *forward substitution*.

1.6 Định thức

1.6.1 Định nghĩa

Định thức của một ma trận vuông $\mathbf{A}$ được ký hiệu là $\det(\mathbf{A})$ hoặc $\det \mathbf{A}$. Có nhiều cách định nghĩa khác nhau của *định thức* (determinant). Chúng ta sẽ sử dụng cách định nghĩa dựa trên quy nạp theo bậc n của ma trận.

Với $n = 1$, $\det(\mathbf{A})$ chính là phần tử duy nhất của ma trận đó.

Với một ma trận vuông bậc $n > 1$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \Rightarrow \det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij}) \tag{1.11}$$

Trong đó $1 \leq i \leq n$ bất kỳ và $\mathbf{A}_{ij}$ là *phần bù đại số của* $\mathbf{A}$ *ứng với phần tử ở hàng* i*, cột* j. Phần bù đại số này là một *ma trận con* của $\mathbf{A}$ nhận được từ $\mathbf{A}$ bằng cách xoá hàng thứ i và cột thứ j của nó. Đây chính là cách tính định thức dựa trên cách khai triển hàng thứ i của ma trận[1].

1.6.2 Tính chất

1. $\det(\mathbf{A}) = \det(\mathbf{A}^T)$: *Một ma trận bất kỳ và chuyển vị của nó có định thức như nhau.*

2. *Định thức của một ma trận đường chéo (và vuông) bằng tích các phần tử trên đường chéo chính.* Nói cách khác, nếu $\mathbf{A} = \text{diag}(a_1, a_2, \dots, a_n)$, thì $\det(\mathbf{A}) = a_1 a_2 \dots a_n$.

3. *Định thức của một ma trận đơn vị bằng 1.*

4. *Định thức của một tích bằng tích các định thức.*

$$\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B}) \tag{1.12}$$

   với $\mathbf{A}, \mathbf{B}$ là hai ma trận vuông cùng chiều.

5. *Nếu một ma trận có một hàng hoặc một cột là một vector* $\mathbf{0}$*, thì định thức của nó bằng 0.*

6. *Một ma trận là khả nghịch khi và chỉ khi định thức của nó khác 0.*

7. *Nếu một ma trận khả nghịch, định thức của ma trận nghịch đảo của nó bằng nghịch đảo định thức của nó.*

$$\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})} \text{ vì } \det(\mathbf{A}) \det(\mathbf{A}^{-1}) = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{I}) = 1. \tag{1.13}$$

1.7 Tổ hợp tuyến tính, không gian sinh

1.7.1 Tổ hợp tuyến tính

Cho các vector khác không $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$ và các số thực $x_1, \dots, x_n \in \mathbb{R}$, vector:

$$\mathbf{b} = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \cdots + x_n \mathbf{a}_n \tag{1.14}$$

được gọi là một *tổ hợp tuyến tính* (linear combination) của $\mathbf{a}_1, \dots, \mathbf{a}_n$. Xét ma trận $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ và $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$, biểu thức (1.14) có thể được viết lại thành $\mathbf{b} = \mathbf{A}\mathbf{x}$. Ta có thể nói rằng $\mathbf{b}$ là một tổ hợp tuyến tính các cột của $\mathbf{A}$.

[1] Việc ghi nhớ định nghĩa này không thực sự quan trọng bằng việc ta cần nhớ một vài tính chất của nó.

Tập hợp tất cả các vector có thể biểu diễn được dưới dạng một tổ hợp tuyến tính của các cột của một ma trận được gọi là *không gian sinh* (*span space*, hoặc gọn là *span*) các cột của ma trận đó. Không gian sinh của một hệ các vector thường được ký hiệu là $\text{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. Nếu phương trình:

$$\mathbf{0} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n \tag{1.15}$$

có nghiệm duy nhất $x_1 = x_2 = \cdots = x_n = 0$, ta nói rằng hệ $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ là một hệ *độc lập tuyến tính* (linear independence). Ngược lại, Nếu tồn tại $x_i \neq 0$ sao cho phương trình trên thoả mãn, ta nói rằng đó là một hệ *phụ thuộc tuyến tính* (linear dependence).

1.7.2 Tính chất

1. *Một hệ là phụ thuộc tuyến tính nếu và chỉ nếu tồn tại một vector trong hệ đó là tổ hợp tuyến tính của các vector còn lại.* Thật vậy, giả sử phương trình (1.15) có nghiệm khác không. Giả sử hệ số khác không là x_i, ta sẽ có:

$$\mathbf{a}_i = \frac{-x_1}{x_i}\mathbf{a}_1 + \cdots + \frac{-x_{i-1}}{x_i}\mathbf{a}_{i-1} + \frac{-x_{i+1}}{x_i}\mathbf{a}_{i+1} + \ldots \frac{-x_n}{x_i}\mathbf{a}_n \tag{1.16}$$

   tức $\mathbf{a}_i$ là một tổ hợp tuyến tính của các vector còn lại.

2. *Tập con khác rỗng của một hệ độc lập tuyến tính là một hệ độc lập tuyến tính.*

3. *Tập hợp các cột của một ma trận khả nghịch tạo thành một hệ độc lập tuyến tính.*

   Giả sử ma trận $\mathbf{A}$ khả nghịch, phương trình $\mathbf{Ax} = \mathbf{0}$ có nghiệm *duy nhất* $\mathbf{x} = \mathbf{A}^{-1}\mathbf{0} = \mathbf{0}$. Vì vậy, các cột của $\mathbf{A}$ tạo thành một hệ độc lập tuyến tính.

4. *Nếu $\mathbf{A}$ là một ma trận cao (tall matrix), tức số hàng lớn hơn số cột, $m > n$, thì tồn tại vector $\mathbf{b}$ sao cho $\mathbf{Ax} = \mathbf{b}$ vô nghiệm.*

   Việc này có thể dễ hình dung trong không gian ba chiều. Không gian sinh của một vector là một đường thẳng, không gian sinh của hai vector độc lập tuyến tính là một mặt phẳng, tức chỉ biểu diễn được các vector nằm trong mặt phẳng đó.

   Ta cũng có thể chứng minh tính chất này bằng phản chứng. Giả sử mọi vector trong không gian m chiều đều nằm trong không gian sinh của một hệ $n < m$ vector là các cột của một ma trận $\mathbf{A}$. Xét các cột của ma trận đơn vị bậc m. Vì mọi cột của ma trận này đều có thể biểu diễn dưới dạng một tổ hợp tuyến tính của n vector đã cho nên phương trình $\mathbf{AX} = \mathbf{I}$ có nghiệm. Nếu ta thêm các vào các cột bằng 0 và các hàng bằng 0 vào $\mathbf{A}$ và $\mathbf{X}$ để được các ma trận vuông, ta sẽ có $\begin{bmatrix}\mathbf{A} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{X} \\ \mathbf{0}\end{bmatrix} = \mathbf{I}$. Việc này chỉ ra rằng $\begin{bmatrix}\mathbf{A} & \mathbf{0}\end{bmatrix}$ là một ma trận khả nghịch trong khi nó có các cột bằng 0. Đây là một điều vô lý vì theo tính chất của định thức, định thức của $\begin{bmatrix}\mathbf{A} & \mathbf{0}\end{bmatrix}$ bằng 0.

5. *Nếu $n > m$, thì n vector bất kỳ trong không gian m chiều tạo thành một hệ phụ thuộc tuyến tính.* Xin được bỏ qua phần chứng minh.

1.7.3 Cơ sở của một không gian

Một hệ các vector $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ trong không gian vector m chiều $V = \mathbb{R}^m$ được gọi là một *cơ sở* (basic) nếu hai điều kiện sau được thoả mãn:

1. $V \equiv \text{span}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$
2. $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ là một hệ độc lập tuyến tính.

Khi đó, mọi vector $\mathbf{b} \in V$ đều có thể biểu diễn *duy nhất* dưới dạng một tổ hợp tuyến tính của các $\mathbf{a}_i$.

Từ hai tính chất cuối ở mục trước, ta có thể suy ra rằng $m = n$.

1.7.4 Range và Null space

Với mỗi ma trận $\mathbf{A} \in \mathbb{R}^{m \times n}$, có hai không gian con quan trọng ứng với ma trận này.

1. Range của $\mathbf{A}$. Range của $\mathbf{A}$, được định nghĩa là:

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \exists \mathbf{x} \in \mathbb{R}^n, \mathbf{A}\mathbf{x} = \mathbf{y}\} \tag{1.17}$$

   Nói cách khác, $\mathcal{R}(\mathbf{A})$ là tập hợp các điểm là tổ hợp tuyến tính của các cột của $\mathbf{A}$, hay chính là không gian sinh (span) của các cột của $\mathbf{A}$. $\mathcal{R}(\mathbf{A})$ là một không gian con của $\mathbb{R}^m$ với số chiều chính bằng sô lượng lớn nhất các cột của $\mathbf{A}$ độc lập tuyến tính.

2. Null của $\mathbf{A}$, ký hiệu là $\mathcal{N}(\mathbf{A})$, được định nghĩa là:

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}\} \tag{1.18}$$

   Mỗi vector trong $\mathcal{N}(\mathbf{A})$ chính là một bộ các hệ số làm cho tổ hợp tuyến tính các cột của $\mathbf{A}$ tạo thành một vector 0. $\mathcal{N}(\mathbf{A})$ có thể được chứng minh là một không gian con trong $\mathbb{R}^n$. Khi các cột của $\mathbf{A}$ là độc lập tuyến tính, theo định nghĩa, $\mathcal{N}(\mathbf{A}) = \{\mathbf{0}\}$ (chỉ gồm vector $\mathbf{0}$).

$\mathcal{R}(\mathbf{A})$ và $\mathcal{N}(\mathbf{A})$ là các không gian con vector với số chiều lần lượt là $\dim(\mathcal{R}(\mathbf{A}))$ và $\dim(\mathcal{N}(\mathbf{A}))$, ta có tính chất quan trọng sau đây:

$$\dim(\mathcal{R}(\mathbf{A})) + \dim(\mathcal{N}(\mathbf{A})) = n \tag{1.19}$$

1.8 Hạng của ma trận

Xét một ma trận $\mathbf{A} \in \mathbb{R}^{m \times n}$. *Hạng* (*rank*) của ma trận này, ký hiệu là $\text{rank}(\mathbf{A})$, được định nghĩa là số lượng lớn nhất các cột của nó tạo thành một hệ độc lập tuyến tính.

Các tính chất quan trọng của hạng:

1. Một ma trận có hạng bằng 0 khi và chỉ khi nó là ma trận 0.
2. $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^T)$. Hạng của một ma trận bằng hạng của ma trận chuyển vị. Nói cách khác, số lượng lớn nhất các cột độc lập tuyến tính của một ma trận bằng với số lượng lớn nhất các hàng độc lập tuyến tính của ma trận đó. Từ đây ta suy ra:
3. Nếu $\mathbf{A} \in \mathbb{R}^{m \times n}$, thì $\text{rank}(\mathbf{A}) \leq \min(m, n)$ vì theo định nghĩa, hạng của một ma trận không thể lớn hơn số hàng hoặc số cột của nó.
4. $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$
5. $\text{rank}(\mathbf{A} + \mathbf{B}) \leq \text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B})$. Điều này chỉ ra rằng một ma trận có hạng bằng k không được biểu diễn dưới dạng ít hơn k ma trận có hạng bằng 1. Đến bài Singular Value Decomposition, chúng ta sẽ thấy rằng một ma trận có hạng bằng k có thể biểu diễn được dưới dạng đúng k ma trận có hạng bằng 1.
6. Bất đẳng thức Sylvester về hạng: Nếu $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times k}$, thì

$$\text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B}) - n \leq \text{rank}(\mathbf{AB})$$

Xét một ma trận vuông $\mathbf{A} \in \mathbb{R}^{n \times}$, hai điều kiện bất kỳ dưới đây là tương đương:

1. $\mathbf{A}$ là một ma trận khả nghịch.
2. Các cột của $\mathbf{A}$ tạo thành một cơ sở trong không gian n chiều.
3. $\det(\mathbf{A}) \neq 0$.
4. $\text{rank}(\mathbf{A}) = n$

1.9 Hệ trực chuẩn, ma trận trực giao

1.9.1 Định nghĩa

Một hệ cơ sở $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m \in \mathbb{R}^m\}$ được gọi là *trực giao* (*orthogonal*) nếu mỗi vector là khác 0 và tích của hai vector khác nhau bất kỳ bằng 0:

$$\mathbf{u}_i \neq \mathbf{0}; \quad \mathbf{u}_i^T \mathbf{u}_j = 0 \; \forall \; 1 \leq i \neq j \leq m \tag{1.20}$$

Một hệ cơ sở $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m \in \mathbb{R}^m\}$ được gọi là *trực chuẩn* (*orthonormal*) nếu nó là một hệ *trực giao* và độ dài Euclidean (xem thêm phần ℓ_2 norm) của mỗi vector bằng 1:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & \text{nếu } i = j \\ 0 & \text{o.w.} \end{cases} \tag{1.21}$$

(o.w. là cách viết ngắn gọn của *trong các trường hợp còn lại* (viết tắt của *otherwise*).)

Gọi $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m]$ với $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m \in \mathbb{R}^m\}$ là *trực chuẩn*, từ (1.21) có thể suy ra:

$$\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{1.22}$$

trong đó $\mathbf{I}$ là ma trận đơn vị bậc m. Nếu một ma trận thoả mãn điều kiện 1.22, ta gọi nó là *ma trận trực giao* (*orthogonal matrix*). *Ma trận loại này không không được gọi là ma trận trực chuẩn, không có định nghĩa cho ma trận trực chuẩn.*

Nếu một ma trận vuông phức $\mathbf{U}$ thoả mãn $\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}$, ta nói rằng $\mathbf{U}$ là một ma *trận unitary* (*unitary matrix*).

1.9.2 Tính chất của ma trận trực giao

1. $\mathbf{U}^{-1} = \mathbf{U}^T$: nghịch đảo của một ma trận trực giao chính là chuyển vị của nó.

2. Nếu $\mathbf{U}$ là ma trận trực giao thì chuyển vị của nó $\mathbf{U}^T$ cũng là một ma trận trực giao.

3. Định thức của ma trận trực giao bằng 1 hoặc -1. Điều này có thể suy ra từ việc $\det(\mathbf{U}) = \det(\mathbf{U}^T)$ và $\det(\mathbf{U})\det(\mathbf{U}^T) = \det(\mathbf{I}) = 1$.

4. Ma trận trực giao thể hiện cho phép xoay một vector (xem thêm mục 1.10). Giả sử có hai vector $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ và một ma trận trực giao $\mathbf{U} \in \mathbb{R}^{m\times m}$. Dùng ma trận này để xoay hai vector trên ta được $\mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{y}$. Tích vô hướng của hai vector mới là:

   $$(\mathbf{U}\mathbf{x})^T(\mathbf{U}\mathbf{y}) = \mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{y} = \mathbf{x}^T\mathbf{y} \tag{1.23}$$

   như vậy *phép xoay không làm thay đổi tích vô hướng giữa hai vector*.

5. Giả sử $\hat{\mathbf{U}} \in \mathbb{R}^{m\times r}, r < m$ là môt ma trận con của ma trận trực giao $\mathbf{U}$ được tạo bởi r cột của $\mathbf{U}$, ta sẽ có $\hat{\mathbf{U}}^T\hat{\mathbf{U}} = \mathbf{I}_r$. Việc này có thể được suy ra từ (1.21).

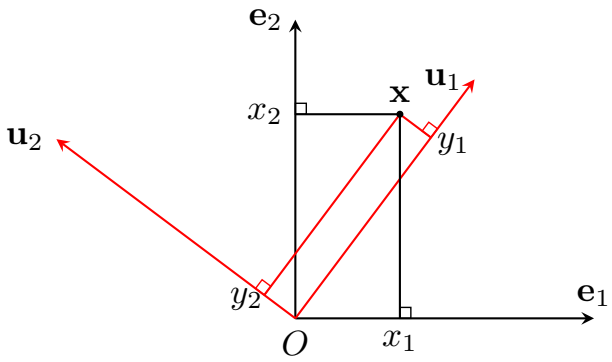
1.10 Biểu diễn vector trong các hệ cơ sở khác nhau

Trong không gian m chiều, toạ độ của mỗi điểm được xác định dựa trên một hệ toạ độ nào đó. Ở các hệ toạ độ khác nhau, hiển nhiên là toạ độ của mỗi điểm cũng khác nhau.

Tập hợp các vector $\mathbf{e}_1, \ldots, \mathbf{e}_m$ mà mỗi vector $\mathbf{e}_i$ có đúng 1 phần tử khác 0 ở thành phần thứ i và phần tử đó bằng 1, được gọi là hệ cơ sở đơn vị (hoặc hệ đơn vị, hoặc hệ chính tắc) trong không gian m chiều. Nếu xếp các vector $\mathbf{e}_i, i = 1, 2, \ldots, m$ theo đúng thứ tự đó, ta sẽ được ma trận đơn vị m chiều.

Mỗi vector cột $\mathbf{x} = [x_1, x_2, \ldots, x_m] \in \mathbb{R}^m$ có thể coi là một tổ hợp tuyến tính của các vector trong hệ cơ sở chính tắc:

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_m\mathbf{e}_m \tag{1.24}$$



Hình 1.1: Chuyển đổi toạ độ trong các hệ cơ sở khác nhau. Trong hệ toạ độ $O\mathbf{e}_1\mathbf{e}_2$, $\mathbf{x}$ có toạ độ là (x_1, x_2). Trong hệ toạ độ $O\mathbf{u}_1\mathbf{u}_2$, $\mathbf{x}$ có toạ độ là (y_2, y_2).

Giả sử có một hệ cơ sở khác $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ (các vector này độc lập tuyến tính), biểu diễn của vector $\mathbf{x}$ trong hệ cơ sở mới này có dạng:

$$\mathbf{x} = y_1\mathbf{u}_1 + y_2\mathbf{u}_2 + \cdots + y_m\mathbf{u}_m = \mathbf{U}\mathbf{y} \tag{1.25}$$

với $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 \ldots \mathbf{u}_m\end{bmatrix}$. Lúc này, vector $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$ chính là biểu diễn của $\mathbf{x}$ trong hệ cơ sở mới. Biểu diễn này là duy nhất vì $\mathbf{y} = \mathbf{U}^{-1}\mathbf{x}$.

Trong các ma trận đóng vai trò như hệ cơ sở, các ma trận trực giao, tức $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, được quan tâm nhiều hơn vì nghịch đảo của chúng chính là chuyển vị của chúng: $\mathbf{U}^{-1} = \mathbf{U}^T$. Khi đó, $\mathbf{y}$ có thể được tính một cách nhanh chóng $\mathbf{y} = \mathbf{U}^T\mathbf{x}$. Từ đó suy ra: $y_i = \mathbf{x}^T\mathbf{u}_i = \mathbf{u}_i^T\mathbf{x}, i = 1, \ldots, m$. Dưới góc nhìn hình học, hệ trực giao tạo thành một hệ trục toạ độ Descartes vuông góc mà chúng ta đã quen thuộc trong không gian hai chiều hoặc ba chiều.

Có thể nhận thấy rằng vector $\mathbf{0}$ được biểu diễn như nhau trong mọi hệ cơ sở. Hình 1.1 là một ví dụ về việc chuyển hệ cơ sở trong không gian hai chiều.

Việc chuyển đổi hệ cơ sở sử dụng ma trận trực giao có thể được coi như một phép xoay trục toạ độ. Nhìn theo một cách khác, đây cũng chính là một phép xoay vector dữ liệu theo chiều ngược lại, nếu ta coi các trục toạ độ là cố định. Trong chương Principle Component Analysis, chúng ta sẽ thấy được một ứng dụng quan trọng của việc đổi hệ cơ sở.

1.11 Trị riêng và vector riêng

1.11.1 Định nghĩa

Cho một ma trận vuông $\mathbf{A} \in \mathbb{R}^{n\times n}$, một vector $\mathbf{x} \in \mathbb{R}^n (\mathbf{x} \neq \mathbf{0})$ và một số vô hướng (có thể thực hoặc phức) λ. Nếu $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, thì ta nói λ và $\mathbf{x}$ là một cặp *trị riêng, vector riêng* (*eigenvalue, eigenvector*) của ma trận $\mathbf{A}$.

Từ định nghĩa ta cũng có $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$, tức $\mathbf{x}$ nằm trong null space của $\mathbf{A} - \lambda\mathbf{I}$. Vì $\mathbf{x} \neq 0$, $\mathbf{A} - \lambda\mathbf{I}$ là một ma trận không khả nghịch. Nói cách khác $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, tức λ là nghiệm của phương trình $\det(\mathbf{A} - t\mathbf{I}) = 0$. Định thức này là một đa thức bậc n của t, được gọi là *đa thức đặc trưng* (*characteristic polynomial*) của $\mathbf{A}$, được ký hiệu là $p_{\mathbf{A}}(t)$. Tập hợp tất cả các trị riêng của một ma trận vuông còn được gọi là *phổ* (*spectrum*) của ma trận đó.

1.11.2 Tính chất

1. Nếu $\mathbf{x}$ là một vector riêng của $\mathbf{A}$ ứng với λ thì $k\mathbf{x}$, $\forall k \neq 0$ cũng là vector riêng ứng với trị riêng đó. Nếu $\mathbf{x}_1, \mathbf{x}_2$ là hai vector riêng ứng với cùng trị riêng λ, thì tổng của chúng cũng là một vector ứng với trị riêng đó. Từ đó suy ra tập hợp các vector riêng ứng với một trị riêng của một ma trận vuông tạo thành một không gian vector con, thường được gọi là *không gian riêng* (*eigenspace*) ứng với trị riêng đó.

2. Mọi ma trận vuông bậc n đều có n trị riêng (kể cả lặp) và có thể là các số phức.

3. Tích của tất cả các trị riêng của một ma trận bằng định thức của ma trận đó. Tổng tất cả các trị riêng của một ma trận bằng tổng các phần tử trên đường chéo của ma trận đó.

4. Phổ của một ma trận bằng phổ của ma trận chuyển vị của nó.

5. Nếu $\mathbf{A}, \mathbf{B}$ là các ma trận vuông cùng bậc thì $p_{\mathbf{AB}}(t) = p_{\mathbf{BA}}(t)$. Điều này nghĩa là, mặc dù tích của hai ma trận không có tính chất giao hoán, đa thức đặc trưng của $\mathbf{AB}$ và $\mathbf{BA}$ là như nhau. Tức phổ của hai tích này là trùng nhau.

6. Với ma trận đối xứng (hoặc tổng quát, Hermitian), tất cả các trị riêng của nó đều là các số thực. Thật vậy, giả sử λ là một trị riêng của một ma trận Hermitian $\mathbf{A}$ và $\mathbf{x}$ là một vector riêng ứng với trị riêng đó. Từ định nghĩa ta suy ra:

$$\mathbf{Ax} = \lambda\mathbf{x} \Rightarrow (\mathbf{Ax})^H = \bar{\lambda}\mathbf{x}^H \Rightarrow \bar{\lambda}\mathbf{x}^H = \mathbf{x}^H\mathbf{A} \tag{1.26}$$

   với $\bar{\lambda}$ là liên hiệp phức của số vô hướng λ. Nhân cả hai vế vào bên phải với $\mathbf{x}$ ta có:

$$\bar{\lambda}\mathbf{x}^H\mathbf{x} = \mathbf{x}^H\mathbf{Ax} = \lambda\mathbf{x}^H\mathbf{x} \Rightarrow (\lambda - \bar{\lambda})\mathbf{x}^H\mathbf{x} = 0 \tag{1.27}$$

   vì $\mathbf{x} \neq 0$ nên $\mathbf{x}^H\mathbf{x} \neq 0$. Từ đó suy ra $\bar{\lambda} = \lambda$, tức λ phải là một số thực.

7. Nếu $(\lambda, \mathbf{x})$ là một cặp trị riêng, vector riêng của một ma trận khả nghịch $\mathbf{A}$, thì $(\frac{1}{\lambda}, \mathbf{x})$ là một cặp trị riêng, vector riêng của $\mathbf{A}^{-1}$, vì $\mathbf{Ax} = \lambda\mathbf{x} \Rightarrow \frac{1}{\lambda}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$.

1.12 Chéo hoá ma trận

Việc phân tích một đại lượng toán học ra thành các đại lượng nhỏ hơn mang lại nhiều hiệu quả. Phân tích một số thành tích các thừa số nguyên tố giúp kiểm tra một số có bao nhiêu ước số. Phân tích đa thức thành nhân tử giúp tìm nghiệm của đa thức. Việc phân tích một ma trận thành tích của các ma trận có dạng đặc biệt khác (quá trình này được gọi là *matrix decomposition*) cũng mang lại nhiều lợi ích trong việc giải hệ phương trình một cách hiệu quả, tính luỹ thừa của ma trận, xấp xỉ ma trận, nén dữ liệu, phân cụm dữ liệu, v.v. Trong mục này, chúng ta sẽ ôn lại một phương pháp matrix decomposition quen thuộc–phương pháp chéo hoá ma trận (*diagonalization* hoặc *eigendecomposition*).

Giả sử $\mathbf{x}_1, \ldots, \mathbf{x}_n \neq \mathbf{0}$ là các vector riêng của một ma trận vuông $\mathbf{A}$ ứng với các trị riêng $\lambda_1, \ldots, \lambda_n$ (có thể lặp hoặc là các số phức) của nó. Tức là $\mathbf{Ax}_i = \lambda_i\mathbf{x}_i$, $\forall i = 1, \ldots, n$.

Đặt $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, và $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\end{bmatrix}$, ta sẽ có $\mathbf{AX} = \mathbf{X}\boldsymbol{\Lambda}$. Hơn nữa, nếu các trị riêng $\mathbf{x}_1, \ldots, \mathbf{x}_n$ là độc lập tuyến tính, ma trận $\mathbf{X}$ là một ma trận khả nghịch. Khi đó ta có thể viết $\mathbf{A}$ dưới dạng tích của ba ma trận:

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \tag{1.28}$$

Các vector riêng $\mathbf{x}_i$ thường được chọn sao cho $\mathbf{x}_i^T\mathbf{x}_i = 1$. Cách biểu diễn một ma trận như (1.28) được gọi là *eigendecomposition* vì nó tách ra thành tích của các ma trận đặc biệt dựa trên vector riêng (eigenvectors) và trị riêng (eigenvalues). Ma trận các trị riêng $\boldsymbol{\Lambda}$ là một ma trận đường chéo. Vì vậy, cách khai triển này cũng có tên gọi là chéo hoá ma trận.

Tính chất:

1. Khái niệm chéo hoá ma trận chỉ áp dụng với ma trận vuông. Vì không có định nghĩa vector riêng hay trị riêng cho ma trận không vuông.

2. Không phải ma trận vuông nào cũng có thể *chéo hoá được* (*diagonalizable*). Một ma trận vuông bậc n là chéo hoá được nếu và chỉ nếu nó có đủ n trị riêng độc lập tuyến tính.

3. Nếu một ma trận là chéo hoá được, có nhiều hơn một cách chéo hoá ma trận đó. Chỉ cần đổi vị trí của các λ_i và vị trí tương ứng các cột của $\mathbf{X}$, ta sẽ có một cách chéo hoá mới.

4. Nếu $\mathbf{A}$ có thể viết được dưới dạng (1.28), khi đó các luỹ thừa có nó cũng chéo hoá được. Cụ thể:

$$\mathbf{A}^2 = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})(\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}) = \mathbf{X}\boldsymbol{\Lambda}^2\mathbf{X}^{-1}; \quad \mathbf{A}^k = \mathbf{X}\boldsymbol{\Lambda}^k\mathbf{X}^{-1},\ \forall k \in \mathbb{N} \tag{1.29}$$

   Xin chú ý rằng nếu λ và $\mathbf{x}$ là một cặp (trị riêng, vector riêng) của $\mathbf{A}$, thì λ^k và $\mathbf{x}$ là một cặp (trị riêng, vector riêng) của $\mathbf{A}^k$. Thật vậy, $\mathbf{A}^k\mathbf{x} = \mathbf{A}^{k-1}(\mathbf{Ax}) = \lambda\mathbf{A}^{k-1}\mathbf{x} = \cdots = \lambda^k\mathbf{x}$.

5. Nếu $\mathbf{A}$ khả nghịch, thì $\mathbf{A}^{-1} = (\mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1})^{-1} = \mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^{-1}$. Vậy chéo hoá ma trận cũng có ích trong việc tính ma trận nghịch đảo.

1.13 Ma trận xác định dương

1.13.1 Định nghĩa

Một ma trận đối xứng[2] $\mathbf{A} \in \mathbb{R}^{n\times n}$ được gọi là *xác định dương* (*positive definite*) nếu:

$$\mathbf{x}^T\mathbf{Ax} > 0, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}. \tag{1.30}$$

Một ma trận đối xứng $\mathbf{A} \in \mathbb{R}^{n\times n}$ được gọi là *nửa xác định dương* (*positive semidefinite*) nếu:

$$\mathbf{x}^T\mathbf{Ax} \geq 0, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}. \tag{1.31}$$

Trên thực tế, ma trận nửa xác định dương, được viết tắt là *PSD*, được sử dụng nhiều hơn.

[2] Chú ý, tồn tại những ma trận không đối xứng thoả mãn điều kiện (1.30). Ta sẽ không xét những ma trận này.

Ma trận *xác định âm* (*negative definite*) và *nửa xác định âm* (*negative semi-definite*) cũng được định nghĩa tương tự.

Ký hiệu $\mathbf{A} \succ 0, \succeq 0, \prec 0, \preceq 0$ được dùng để chỉ một ma trận là xác định dương, nửa xác định dương, xác định âm, nửa xác định âm, theo thứ tự đó. Ký hiệu $\mathbf{A} \succ \mathbf{B}$ cũng được dùng để chỉ ra rằng $\mathbf{A} - \mathbf{B} \succ 0$.

Mở rộng, một ma trận phức, Hermitian $\mathbf{A} \in \mathbb{C}^{n \times n}$ được gọi là xác định dương nếu:

$$\mathbf{x}^H \mathbf{A} \mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{C}^n, \mathbf{x} \neq \mathbf{0}. \tag{1.32}$$

Ví dụ, $\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ là nửa xác định dương vì với mọi vector $\mathbf{x} = \begin{bmatrix} u \\ v \end{bmatrix}$, ta có:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = u^2 + v^2 - 2uv = (u - v)^2 \geq 0, \forall u, v \in \mathbb{R} \tag{1.33}$$

1.13.2 Tính chất

1. *Mọi trị riêng của một ma trận xác định dương đều là một số thực dương.*

   Trước hết, các trị riêng của các ma trận dạng này là số thực vì các ma trận đều là đối xứng. Để chứng minh chúng là các số thực dương, ta giả sử λ là một trị riêng của một ma trận xác định dương $\mathbf{A}$ và $\mathbf{x} \neq \mathbf{0}$ là một vector riêng ứng với trị riêng đó. Nhân vào bên trái cả hai vế của $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ với $\mathbf{x}^H$ ta có:

   $$\lambda \mathbf{x}^H \mathbf{x} = \mathbf{x}^H \mathbf{A} \mathbf{x} > 0 \tag{1.34}$$

   (ở đây Hermitian được dùng để xét tổng quát cho cả trường hợp ma trận phức). Vì $\mathbf{x}^H \mathbf{x}$ luôn dương với mọi $\mathbf{x}$ nên ta phải có $\lambda > 0$. Tương tự, ta có thể chứng minh được rằng mọi trị riêng của một ma trận nửa xác định dương là không âm.

2. *Mọi ma trận xác định dương là khả nghịch. Hơn nữa, định thức của nó là một số dương.*

   Điều này được trực tiếp suy ra từ tính chất 1. Nhắc lại rằng định thức của một ma trận bằng tích tất cả các trị riêng của nó.

3. Tiêu chuẩn Sylvester: *Một ma trận Hermitian là xác định dương nếu và chỉ nếu mọi leading principal minors của nó là dương. Một ma trận Hermitian là nửa xác định dương nếu mọi principal minors của nó là không âm.* Đây là một tiêu chuẩn để kiểm tra một ma trận Hermitian $\mathbf{A} \in \mathbb{R}^n$ có là (nửa) xác định dương hay không. Ở đây, *leading principal minors* và *principal minors* được định nghĩa như sau:

   Gọi $\mathcal{I}$ là một tập con bất kỳ của $\{1, 2, \ldots, n\}$, $\mathbf{A}_\mathcal{I}$ là ma trận con của $\mathbf{A}$ nhận được bằng cách trích ra các hàng và cột có chỉ số nằm trong $\mathcal{I}$ của $\mathbf{A}$. Khi đó, $\mathbf{A}_\mathcal{I}$ và $\det(\mathbf{A}_\mathcal{I})$ lần lượt được gọi là một *ma trận con chính* (*principal submatrix*) và *principal minor* của $\mathbf{A}$. Nếu $\mathcal{I}$ chỉ bao gồm các số tự nhiên liên tiếp từ 1 đến $k \leq n$, ta nói $\mathbf{A}_\mathcal{I}$ và $\det(\mathbf{A}_\mathcal{I})$ lần lượt là một *leading principal submatrix* và *leading principal minor* bậc k của $\mathbf{A}$.

4. $\mathbf{A} = \mathbf{B}^H\mathbf{B}$ *là nửa xác định dương với mọi ma trận* $\mathbf{B}$ *(*$\mathbf{B}$ *không nhất thiết vuông).*

   Thật vậy, với mọi vector $\mathbf{x} \neq 0$ với chiều phù hợp, $\mathbf{x}^H\mathbf{A}\mathbf{x} = \mathbf{x}^H\mathbf{B}^H\mathbf{B}\mathbf{x} = (\mathbf{B}\mathbf{x})^H(\mathbf{B}\mathbf{x}) \geq 0$.

5. Khai triển Cholesky (Cholesky decomposition): *Mọi ma trận Hermitian, nửa xác định dương* $\mathbf{A}$ *đều biểu diễn được duy nhất dưới dạng* $\mathbf{A} = \mathbf{L}\mathbf{L}^H$*, trong đó* $\mathbf{L}$ *là một ma trận tam giác dưới với các thành phần trên đường chéo là thực dương.*

6. *Nếu* $\mathbf{A}$ *là một ma trận nửa xác định dương, thì* $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0 \Leftrightarrow \mathbf{A}\mathbf{x} = 0$.

   Nếu $\mathbf{A}\mathbf{x} = 0 \Rightarrow \mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ một cách hiển nhiên.

   Nếu $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$. Với vector $\mathbf{y} \neq \mathbf{0}$ bất kỳ có cùng kích thước với $\mathbf{x}$, xét hàm số sau đây:

   $$f(\lambda) = (\mathbf{x} + \lambda\mathbf{y})^T\mathbf{A}(\mathbf{x} + \lambda\mathbf{y}) \tag{1.35}$$

   Hàm số này không âm với mọi λ vì $\mathbf{A}$ là một ma trận nửa xác định dương. Đây là một tam thức bậc hai của λ:

   $$f(\lambda) = \mathbf{y}^T\mathbf{A}\mathbf{y}\lambda^2 + 2\mathbf{y}^T\mathbf{A}\mathbf{x}\lambda + \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{y}\lambda^2 + 2\mathbf{y}^T\mathbf{A}\mathbf{x}\lambda \tag{1.36}$$

   Xét hai trường hợp:

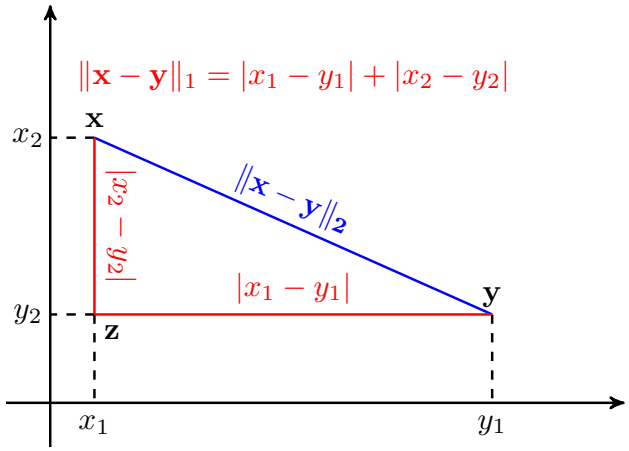
   - $\mathbf{y}^T\mathbf{A}\mathbf{y} = 0$. Khi đó, $f(\lambda) = 2\mathbf{y}^T\mathbf{A}\mathbf{x}\lambda \geq 0, \forall\lambda$ nếu và chỉ nếu $\mathbf{y}^T\mathbf{A}\mathbf{x} = 0$.
   - $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$. Khi đó tam thức bậc hai $f(\lambda) \geq 0, \forall\lambda$ nếu và chỉ nếu $\Delta' = (\mathbf{y}^T\mathbf{A}\mathbf{x})^2 \leq 0$ vì hệ số ứng với thành phần bậc hai bằng $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$. Điều này cũng đồng nghĩa với việc $\mathbf{y}^T\mathbf{A}\mathbf{x} = 0$

   Tóm lại, $\mathbf{y}^T\mathbf{A}\mathbf{x} = 0,\ \forall\mathbf{y} \neq \mathbf{0}$. Điều này chỉ xảy ra nếu $\mathbf{A}\mathbf{x} = 0$. □

1.14 Chuẩn của vector và ma trận

Trong không gian một chiều, khoảng cách giữa hai điểm là trị tuyệt đối của hiệu giữa hai giá trị đó. Trong không gian hai chiều, tức mặt phẳng, chúng ta thường dùng khoảng cách Euclid để đo khoảng cách giữa hai điểm. Khoảng cách này chính là đại lượng chúng ta thường nói bằng ngôn ngữ thông thường là *đường chim bay*. Đôi khi, để đi từ một điểm này tới một điểm kia, con người chúng ta không thể đi bằng đường chim bay được mà còn phụ thuộc vào việc đường đi nối giữa hai điểm có dạng như thế nào.

Việc đo khoảng cách giữa hai điểm dữ liệu nhiều chiều, tức hai vector, là rất cần thiết trong Machine Learning. Và đó chính là lý do mà khái niệm *chuẩn* (*norm*) ra đời. Để xác định khoảng cách giữa hai vector $\mathbf{y}$ và $\mathbf{z}$, người ta thường áp dụng một hàm số lên vector hiệu $\mathbf{x} = \mathbf{y} - \mathbf{z}$. Hàm số này cần có một vài tính chất đặc biệt.



Hình 1.2: Minh họa ℓ_1 norm và ℓ_2 norm trong không gian hai chiều. ℓ_2 norm chính là khoảng cách giữa hai điểm trong mặt phẳng. Trong khi đó ℓ_1 norm là quãng đường ngắn nhất giữa hai điểm nếu chỉ được đi theo các đường song song với các trục toạ độ.

Định nghĩa 1.1: Norm

Một hàm số $f : \mathbb{R}^n \to \mathbb{R}$ được gọi là một norm nếu nó thỏa mãn ba điều kiện sau đây:

1. $f(\mathbf{x}) \geq 0$. Dấu bằng xảy ra $\Leftrightarrow \mathbf{x} = \mathbf{0}$.
2. $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x}), \quad \forall \alpha \in \mathbb{R}$
3. $f(\mathbf{x}_1) + f(\mathbf{x}_2) \geq f(\mathbf{x}_1 + \mathbf{x}_2), \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$

Điều kiện thứ nhất là dễ hiểu vì khoảng cách không thể là một số âm. Hơn nữa, khoảng cách giữa hai điểm $\mathbf{y}$ và $\mathbf{z}$ bằng 0 nếu và chỉ nếu hai điểm nó trùng nhau, tức $\mathbf{x} = \mathbf{y} - \mathbf{z} = \mathbf{0}$.

Điều kiện thứ hai cũng có thể được lý giải như sau. Nếu ba điểm $\mathbf{y}, \mathbf{v}$ và $\mathbf{z}$ thẳng hàng, hơn nữa $\mathbf{v} - \mathbf{y} = \alpha(\mathbf{v} - \mathbf{z})$ thì khoảng cách giữa $\mathbf{v}$ và $\mathbf{y}$ gấp $|\alpha|$ lần khoảng cách giữa $\mathbf{v}$ và $\mathbf{z}$.

Điều kiện thứ ba chính là bất đẳng thức tam giác nếu ta coi $\mathbf{x}_1 = \mathbf{y} - \mathbf{w}, \mathbf{x}_2 = \mathbf{w} - \mathbf{z}$ với $\mathbf{w}$ là một điểm bất kỳ trong cùng không gian.

1.14.1 Một số chuẩn vector thường dùng

Độ dài Euclid của một vector $\mathbf{x} \in \mathbb{R}^n$ chính là một norm, norm này được gọi là ℓ_2 norm hoặc Euclidean norm:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{1.37}$$

Bình phương của ℓ_2 norm chính là tích vô hướng của một vector với chính nó, $\|\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{x}$. Với p **là một số không nhỏ hơn 1** bất kỳ, hàm số:

$$\|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \ldots |x_n|^p)^{\frac{1}{p}} \tag{1.38}$$

được chứng minh thỏa mãn ba điều kiện của norm, và được gọi là ℓ_p norm.

Có một vài giá trị của p thường được dùng:

1. Khi $p = 2$ chúng ta có ℓ_2 norm như ở trên.

2. Khi $p = 1$ chúng ta có ℓ_1 norm: $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$ là tổng các trị tuyệt đối của từng phần tử của $\mathbf{x}$. Hình 1.2 là một ví dụ sánh ℓ_1 norm và ℓ_2 norm trong không gian hai chiều. Norm 2 (màu xanh) chính là đường thẳng *chim bay* nối giữa hai vector $\mathbf{x}$ và $\mathbf{y}$. Khoảng cách ℓ_1 norm giữa hai điểm này (màu đỏ) có thể diễn giải như là đường đi từ $\mathbf{x}$ tới $\mathbf{y}$ trong một thành phố mà đường phố tạo thành hình bàn cờ. Chúng ta chỉ có cách đi dọc theo cạnh của bàn cờ mà không được đi thẳng như đường chim bay.

3. Khi $p \to \infty$, giả sử $i = \arg\max_{j=1,2,\ldots,n} |x_j|$. Khi đó:

$$\|\mathbf{x}\|_p = |x_i| \left(1 + \left|\frac{x_1}{x_i}\right|^p + \cdots + \left|\frac{x_{i-1}}{x_i}\right|^p + \left|\frac{x_{i+1}}{x_i}\right|^p + \cdots + \left|\frac{x_n}{x_i}\right|^p\right)^{\frac{1}{p}} \tag{1.39}$$

Ta thấy rằng:

$$\lim_{p\to\infty} \left(1 + \left|\frac{x_1}{x_i}\right|^p + \cdots + \left|\frac{x_{i-1}}{x_i}\right|^p + \left|\frac{x_{i+1}}{x_i}\right|^p + \cdots + \left|\frac{x_n}{x_i}\right|^p\right)^{\frac{1}{p}} = 1 \tag{1.40}$$

vì đại lượng trong dấu ngoặc đơn không vượt quá n, ta sẽ có:

$$\|\mathbf{x}\|_\infty \triangleq \lim_{p\to\infty} \|\mathbf{x}\|_p = |x_i| = \max_{j=1,2,\ldots,n} |x_j| \tag{1.41}$$

1.14.2 Chuẩn Frobenius của ma trận

Với một ma trận $\mathbf{A} \in \mathbb{R}^{m\times n}$, chuẩn thường được dùng nhất là chuẩn Frobenius, ký hiệu là $\|\mathbf{A}\|_F$ là căn bậc hai của tổng bình phương tất cả các phần tử của ma trận đó.

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$$

Chú ý rằng ℓ_2 norm $\|\mathbf{A}\|_2$ là một norm khác của ma trận, không phổ biến bằng Frobenius norm. Bạn đọc có thể xem ℓ_2 norm của ma trận trong Phụ lục **??**.

1.14.3 Vết của ma trận

Vết (*trace*) của một ma trận vuông là tổng tất cả cả phần tử trên đường chéo chính của nó.

Vết của một ma trận được $\mathbf{A}$ được ký hiệu là $\text{trace}(\mathbf{A})$. Hàm số trace xác định trên tập các ma trận vuông được sử dụng rất nhiều trong tối ưu vì những tính chất đẹp của nó.

Các tính chất quan trọng của hàm trace, với giả sử rằng các ma trận trong hàm trace là vuông và các phép nhân ma trận thực hiện được:

- *Một ma trận vuông bất kỳ và chuyển vị của nó có trace bằng nhau* $\text{trace}(\mathbf{A}) = \text{trace}(\mathbf{A}^T)$. Việc này khá hiển nhiên vì phép chuyển vị không làm thay đổi các phần tử trên đường chéo chính của một ma trận.

- *trace của một tổng bằng tổng các trace*: $\text{trace}(\sum_{i=1}^{k} \mathbf{A}_i) = \sum_{i=1}^{k} \text{trace}(\mathbf{A}_i)$.
- $\text{trace}(k\mathbf{A}) = k\text{trace}(\mathbf{A})$ với k là một số vô hướng bất kỳ.
- $\text{trace}(\mathbf{A}) = \sum_{i=1}^{D} \lambda_i$ với $\mathbf{A}$ là một ma trận vuông và $\lambda_i, i = 1, 2, \ldots, N$ là toàn bộ các trị riêng của nó, có thể phức hoặc lặp. Việc chứng minh tính chất này có thể được dựa trên ma trận đặc trưng của $\mathbf{A}$ và định lý Viète.
- $\text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})$. Đẳng thức này được suy ra từ việc đa thức đặc trưng của $\mathbf{AB}$ và $\mathbf{BA}$ là như nhau. Bạn đọc cũng có thể chứng minh bằng cách tính trực tiếp các phần tử trên đường chéo chính của $\mathbf{AB}$ và $\mathbf{BA}$.
- $\text{trace}(\mathbf{ABC}) = \text{trace}(\mathbf{BCA})$ nhưng $\text{trace}(\mathbf{ABC})$ không đồng nhất với $\text{trace}(\mathbf{ACB})$.
- Nếu $\mathbf{X}$ là một ma trận khả nghịch cùng chiều với $\mathbf{A}$:

$$\text{trace}(\mathbf{XAX}^{-1}) = \text{trace}(\mathbf{X}^{-1}\mathbf{XA}) = \text{trace}(\mathbf{A}) \tag{1.42}$$

- $\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^T\mathbf{A}) = \text{trace}(\mathbf{AA}^T)$ với $\mathbf{A}$ là một ma trận bất kỳ. Từ đây ta cũng suy ra $\text{trace}(\mathbf{AA}^T) \geq 0$ với mọi ma trận $\mathbf{A}$.

Chương 2

Giải tích ma trận

Trong chương này, nếu không nói gì thêm, chúng ta giả sử rằng các đạo hàm tồn tại. Tài liệu tham khảo chính của chương là *Matrix calculus–Stanford* (*https://goo.gl/BjTPLr*).

2.1 Đạo hàm của hàm trả về một số vô hướng

Đạo hàm bậc nhất (*first-order gradient*) hay viết gọn là *đạo hàm* (*gradient*) của một hàm số $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ theo $\mathbf{x}$ được định nghĩa là

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \triangleq \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \dfrac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \tag{2.1}$$

trong đó $\dfrac{\partial f(\mathbf{x})}{\partial x_i}$ là *đạo hàm riêng* (*partial derivative*) của hàm số theo thành phần thứ i của vector $\mathbf{x}$. Đạo hàm này được lấy khi tất cả các biến, ngoài x_i, được giả sử là hằng số. Nếu không có thêm biến nào khác, $\nabla_{\mathbf{x}} f(\mathbf{x})$ thường được viết gọn là $\nabla f(\mathbf{x})$. **Đạo hàm của hàm số này là một vector có cùng chiều với vector đang được lấy đạo hàm**. Tức nếu vector được viết ở dạng cột thì đạo hàm cũng phải được viết ở dạng cột.

Đạo hàm bậc hai (*second-order gradient*) của hàm số trên còn được gọi là *Hessian* và được định nghĩa như sau, với $\mathbb{S}^n \in \mathbb{R}^{n \times n}$ là tập các ma trận vuông đối xứng bậc n.

$$\nabla^2 f(\mathbf{x}) \triangleq \begin{bmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{S}^n \tag{2.2}$$

Đạo hàm của một hàm số $f(\mathbf{X}) : \mathbb{R}^{n\times m} \to \mathbb{R}$ **theo ma trận X** được định nghĩa là

$$\nabla f(\mathbf{X}) = \begin{bmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \frac{\partial f(\mathbf{X})}{\partial x_{12}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1m}} \\ \frac{\partial f(\mathbf{X})}{\partial x_{21}} & \frac{\partial f(\mathbf{X})}{\partial x_{22}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{n1}} & \frac{\partial f(\mathbf{X})}{\partial x_{n2}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{nm}} \end{bmatrix} \in \mathbb{R}^{n\times m} \tag{2.3}$$

Chiều của đạo hàm

Đạo hàm của hàm số $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ là một ma trận trong $\mathbb{R}^{m\times n}$, $\forall m, n \in \mathbb{N}^*$.

Cụ thể, để tính đạo hàm của một hàm $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, ta tính đạo hàm riêng của hàm số đó theo từng thành phần của ma trận *khi toàn bộ các thành phần khác được giả sử là hằng số*. Tiếp theo, ta sắp xếp các đạo hàm riêng tính được theo đúng thứ tự trong ma trận.

Ví dụ: Xét hàm số $f : \mathbb{R}^2 \to \mathbb{R}$, $f(\mathbf{x}) = x_1^2 + 2x_1x_2 + \sin(x_1) + 2$.

Đạo hàm bậc nhất theo $\mathbf{x}$ của hàm số đó là

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + 2x_2 + \cos(x_1) \\ 2x_1 \end{bmatrix}$$

Đạo hàm bậc hai theo $\mathbf{x}$, hay *Hessian* là $\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial f^2(\mathbf{x})}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial f^2(\mathbf{x})}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 - \sin(x_1) & 2 \\ 2 & 0 \end{bmatrix}$

Chú ý rằng *Hessian* luôn là một ma trận đối xứng.

2.2 Đạo hàm của hàm trả về một vector

Những *hàm số trả về một vector*, hoặc gọn hơn *hàm trả về vector* được gọi là *vector-valued function* trong tiếng Anh.

Xét một hàm trả về vector với **đầu vào là một số thực** $v(x) : \mathbb{R} \to \mathbb{R}^n$:

$$v(x) = \begin{bmatrix} v_1(x) \\ v_2(x) \\ \vdots \\ v_n(x) \end{bmatrix} \tag{2.4}$$

Đạo hàm của hàm số này theo x là một **vector hàng** như sau:

$$\nabla v(x) \triangleq \left[\frac{\partial v_1(x)}{\partial x} \frac{\partial v_2(x)}{\partial x} \dots \frac{\partial v_n(x)}{\partial x}\right] \tag{2.5}$$

Đạo hàm bậc hai của hàm số này có dạng

$$\nabla^2 v(x) \triangleq \left[\frac{\partial^2 v_1(x)}{\partial x^2} \frac{\partial^2 v_2(x)}{\partial x^2} \dots \frac{\partial^2 v_n(x)}{\partial x^2}\right] \tag{2.6}$$

Ví dụ: Cho một vector $\mathbf{a} \in \mathbb{R}^n$ và một hàm số *vector-valued* $v(x) = x\mathbf{a}$, đạo hàm bậc nhất và Hession của nó lần lượt là

$$\nabla v(x) = \mathbf{a}^T, \quad \nabla^2 v(x) = \mathbf{0} \in \mathbb{R}^{1\times n} \tag{2.7}$$

Xét một hàm trả về vector với **đầu vào là một vector** $h(\mathbf{x}) : \mathbb{R}^k \to \mathbb{R}^n$, đạo hàm bậc nhất của nó là

$$\nabla h(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial h_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial h_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\mathbf{x})}{\partial x_k} & \frac{\partial h_2(\mathbf{x})}{\partial x_k} & \dots & \frac{\partial h_n(\mathbf{x})}{\partial x_k} \end{bmatrix} = \left[\nabla h_1(\mathbf{x}) \nabla h_2(\mathbf{x}) \dots \nabla h_n(\mathbf{x})\right] \in \mathbb{R}^{k\times n} \tag{2.8}$$

Nếu một hàm số $g : \mathbb{R}^m \to \mathbb{R}^n$, thì đạo hàm của nó là một ma trận thuộc $\mathbb{R}^{m\times n}$.

Đạo hàm bậc hai của hàm số trên là một *mảng ba chiều*, chúng ta sẽ không nhắc đến ở đây.

Trước khi đến phần tính đạo hàm của các hàm số thường gặp, chúng ta cần biết hai tính chất quan trọng khá giống với đạo hàm của hàm một biến.

2.3 Tính chất quan trọng của đạo hàm

2.3.1 Quy tắc tích (Product rule)

Để cho tổng quát, ta giả sử biến đầu vào là một ma trận. Giả sử rằng các hàm số có chiều phù hợp để các phép nhân thực hiện được. Ta có:

$$\nabla \left(f(\mathbf{X})^T g(\mathbf{X})\right) = \left(\nabla f(\mathbf{X})\right) g(\mathbf{X}) + \left(\nabla g(\mathbf{X})\right) f(\mathbf{X}) \tag{2.9}$$

Biểu thức này giống như biểu thức chúng ta đã quen thuộc:

$$(f(x)g(x))' = f'(x)g(x) + g'(x)f(x)$$

Chú ý rằng với tích của vector và ma trận, ta không được sử dụng tính chất giao hoán.

2.3.2 Quy tắc chuỗi (Chain rule)

Khi có các hàm hợp thì

$$\nabla_{\mathbf{X}} g(f(\mathbf{X})) = (\nabla_{\mathbf{X}} f)^T (\nabla_f g) \tag{2.10}$$

Quy tắc này cũng giống với quy tắc trong hàm một biến:

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

Một lưu ý nhỏ nhưng quan trọng khi làm việc với tích các ma trận là sự phù hợp về kích thước của các ma trận trong tích.

2.4 Đạo hàm của các hàm số thường gặp

2.4.1 $\boldsymbol{f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}}$

Giả sử $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, ta viết lại $\quad f(\mathbf{x}) = \mathbf{a}^T\mathbf{x} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$

Có thể nhận thấy rằng $\quad \dfrac{\partial f(\mathbf{x})}{\partial x_i} = a_i, \ \forall i = 1, 2 \ldots, n.$

Vậy, $\nabla(\mathbf{a}^T\mathbf{x}) = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}^T = \mathbf{a}$. Ngoài ra, vì $\mathbf{a}^T\mathbf{x} = \mathbf{x}^T\mathbf{a}$ nên $\nabla(\mathbf{x}^T\mathbf{a}) = \mathbf{a}$.

2.4.2 $\boldsymbol{f(\mathbf{x}) = \mathbf{A}\mathbf{x}}$

Đây là một hàm trả về vector $f : \mathbb{R}^n \to \mathbb{R}^m$ với $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m \times n}$. Giả sử rằng $\mathbf{a}_i$ là **hàng** thứ i của ma trận $\mathbf{A}$. Ta có

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1\mathbf{x} \\ \mathbf{a}_2\mathbf{x} \\ \vdots \\ \mathbf{a}_m\mathbf{x} \end{bmatrix}$$

Theo định nghĩa (2.8), và công thức đạo hàm của $\mathbf{a}_i\mathbf{x}$, ta có thể suy ra

$$\nabla_{\mathbf{x}}(\mathbf{A}\mathbf{x}) = \begin{bmatrix} \mathbf{a}_1^T & \mathbf{a}_2^T & \ldots & \mathbf{a}_m^T \end{bmatrix} = \mathbf{A}^T \tag{2.11}$$

Từ đây ta có thể suy ra đạo hàm của hàm số $f(\mathbf{x}) = \mathbf{x} = \mathbf{I}\mathbf{x}$, với $\mathbf{I}$ là ma trận đơn vị, là

$$\nabla \mathbf{x} = \mathbf{I}$$

2.4.3 $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$

với $\mathbf{x} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$. Áp dụng quy tắc tích (2.9) ta có

$$\begin{aligned}\nabla f(\mathbf{x}) &= \nabla\left(\left(\mathbf{x}^T\right)\left(\mathbf{A}\mathbf{x}\right)\right)\\ &= \left(\nabla(\mathbf{x})\right)\mathbf{A}\mathbf{x} + \left(\nabla(\mathbf{A}\mathbf{x})\right)\mathbf{x}\\ &= \mathbf{I}\mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}\\ &= (\mathbf{A} + \mathbf{A}^T)\mathbf{x}\end{aligned} \tag{2.12}$$

Từ (2.12) và (2.11), ta có thể suy ra $\nabla^2\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T + \mathbf{A}$

Nếu $\mathbf{A}$ là một ma trận đối xứng, ta sẽ có $\nabla\mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}, \quad \nabla^2\mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}$

Nếu $\mathbf{A}$ là ma trận đơn vị, tức $f(\mathbf{x}) = \mathbf{x}^T\mathbf{I}\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$, ta có

$$\nabla\|\mathbf{x}\|_2^2 = 2\mathbf{x}, \quad \nabla^2\|\mathbf{x}\|_2^2 = 2\mathbf{I} \tag{2.13}$$

2.4.4 $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$

Có hai cách tính đạo hàm của hàm số này:

Cách 1: Trước hết, biến đổi

$$\begin{aligned}f(\mathbf{x}) &= \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = (\mathbf{x}^T\mathbf{A}^T - \mathbf{b}^T)(\mathbf{A}\mathbf{x} - \mathbf{b})\\ &= \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b}\end{aligned}$$

Lấy đạo hàm cho từng số hạng rồi cộng lại ta có

$$\nabla\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = 2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b} = 2\mathbf{A}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

Cách 2: Sử dụng $\nabla(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{A}^T$ và $\nabla\|\mathbf{x}\|_2^2 = 2\mathbf{x}$ và quy tắc chuỗi (2.10), ta cũng sẽ thu được kết quả tương tự.

2.4.5 $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$

Bằng cách viết lại $f(\mathbf{x}) = (\mathbf{a}^T\mathbf{x})(\mathbf{x}^T\mathbf{b})$, ta có thể dùng Quy tắc tích (2.9) và có kết quả

$$\nabla(\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}) = \mathbf{a}\mathbf{x}^T\mathbf{b} + \mathbf{b}\mathbf{a}^T\mathbf{x} = \mathbf{a}\mathbf{b}^T\mathbf{x} + \mathbf{b}\mathbf{a}^T\mathbf{x} = (\mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T)\mathbf{x},$$

ở đây ta đã sử dụng tính chất $\mathbf{y}^T\mathbf{z} = \mathbf{z}^T\mathbf{y}$.

2.4.6 $f(\mathbf{X}) = \text{trace}(\mathbf{A}\mathbf{X})$

Giả sử $\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{X} = \mathbb{R}^{m\times n}$, và $\mathbf{B} = \mathbf{A}\mathbf{X} \in \mathbb{R}^{n\times n}$. Theo định nghĩa của trace,

$$f(\mathbf{X}) = \text{trace}(\mathbf{A}\mathbf{X}) = \text{trace}(\mathbf{B}) = \sum_{j=1}^{n} b_{jj} = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ji}x_{ji} \tag{2.14}$$

Từ đây ta thấy rằng $\dfrac{\partial f(\mathbf{X})}{\partial x_{ij}} = a_{ji}$. Sử dụng định nghĩa (2.3) ta đạt được $\nabla_{\mathbf{X}}\text{trace}(\mathbf{A}\mathbf{X}) = \mathbf{A}^T$.

Bảng 2.1: Bảng các đạo hàm cơ bản.

$f(\mathbf{x})$	$\nabla f(\mathbf{x})$	$f(\mathbf{X})$	$\nabla_{\mathbf{X}} f(\mathbf{X})$
$\mathbf{x}$	$\mathbf{I}$	$\text{trace}(\mathbf{X})$	$\mathbf{I}$
$\mathbf{a}^T\mathbf{x}$	$\mathbf{a}$	$\text{trace}(\mathbf{A}^T\mathbf{X})$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$(\mathbf{A}+\mathbf{A}^T)\mathbf{x}$	$\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X})$	$(\mathbf{A}+\mathbf{A}^T)\mathbf{X}$
$\mathbf{x}^T\mathbf{x} = \\|\mathbf{x}\\|_2^2$	$2\mathbf{x}$	$\text{trace}(\mathbf{X}^T\mathbf{X}) = \\|\mathbf{X}\\|_F^2$	$2\mathbf{X}$
$\\|\mathbf{A}\mathbf{x}-\mathbf{b}\\|_2^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b})$	$\\|\mathbf{A}\mathbf{X}-\mathbf{B}\\|_F^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{X}-\mathbf{B})$
$\mathbf{a}^T\mathbf{x}^T\mathbf{x}\mathbf{b}$	$2\mathbf{a}^T\mathbf{b}\mathbf{x}$	$\mathbf{a}^T\mathbf{X}\mathbf{b}$	$\mathbf{a}\mathbf{b}^T$
$\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$	$(\mathbf{a}\mathbf{b}^T+\mathbf{b}\mathbf{a}^T)\mathbf{x}$	$\text{trace}(\mathbf{A}^T\mathbf{X}\mathbf{B})$	$\mathbf{A}\mathbf{B}^T$

2.4.7 $f(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{b}$

Giả sử rằng $\mathbf{a} \in \mathbb{R}^m, \mathbf{X} \in R^{m\times n}, \mathbf{b} \in \mathbb{R}^n$. Bạn đọc có thể chứng minh được

$$f(\mathbf{X}) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}a_ib_j$$

Từ đó, sử dụng định nghĩa (2.3) ta sẽ có $\nabla_{\mathbf{X}}(\mathbf{a}^T\mathbf{X}\mathbf{b}^T) = \begin{bmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \dots & \dots & \ddots & \dots \\ a_mb_1 & a_mb_2 & \dots & a_mb_n \end{bmatrix} = \mathbf{a}\mathbf{b}^T$.

2.4.8 $f(\mathbf{X}) = \|\mathbf{X}\|_F^2$

Giả sử $\mathbf{X} \in \mathbb{R}^{n\times n}$, bằng cách viết lại $\|\mathbf{X}\|_F^2 = \sum_{i=1}^n\sum_{j=1}^n x_{ij}^2$, ta có thể suy ra $\frac{\partial f}{\partial x_{ij}} = 2x_{ij}$. Và vì vậy, $\nabla\|\mathbf{X}\|_F^2 = 2\mathbf{X}$.

2.4.9 $f(\mathbf{X}) = \text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X})$

Giả sử rằng $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_m\end{bmatrix} \in \mathbb{R}^{m\times n}, \mathbf{A} \in \mathbb{R}^{m\times m}$. Bằng cách khai triển

$$\mathbf{X}^T\mathbf{A}\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T\end{bmatrix}\mathbf{A}\begin{bmatrix}\mathbf{x}_1 \mathbf{x}_2 \dots, \mathbf{x}_n\end{bmatrix} = \begin{bmatrix}\mathbf{x}_1^T\mathbf{A}\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{A}\mathbf{x}_2 & \dots & \mathbf{x}_1^T\mathbf{A}\mathbf{x}_n \\ \mathbf{x}_2^T\mathbf{A}\mathbf{x}_1 & \mathbf{x}_2^T\mathbf{A}\mathbf{x}_2 & \dots & \mathbf{x}_2^T\mathbf{A}\mathbf{x}_n \\ \dots & \dots & \ddots & \dots \\ \mathbf{x}_n^T\mathbf{A}\mathbf{x}_1 & \mathbf{x}_n^T\mathbf{A}\mathbf{x}_2 & \dots & \mathbf{x}_n^T\mathbf{A}\mathbf{x}_n\end{bmatrix}, \tag{2.15}$$

ta tính được $\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = \sum_{i=1}^n \mathbf{x}_i^T\mathbf{A}\mathbf{x}_i$. Nhắc lại rằng $\nabla_{\mathbf{x}_i}\mathbf{x}_i^T\mathbf{A}\mathbf{x}_i = (\mathbf{A}+\mathbf{A}^T)\mathbf{x}_i$, ta có

$$\nabla_{\mathbf{X}}\text{trace}(\mathbf{X}^T\mathbf{A}\mathbf{X}) = (\mathbf{A}+\mathbf{A}^T)\begin{bmatrix}\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n\end{bmatrix} = (\mathbf{A}+\mathbf{A}^T)\mathbf{X} \tag{2.16}$$

Bằng cách thay $\mathbf{A} = \mathbf{I}$, ta cũng thu được $\nabla_{\mathbf{X}}\text{trace}(\mathbf{X}^T\mathbf{X}) = \nabla_{\mathbf{X}}\|\mathbf{X}\|_F^2 = 2\mathbf{X}$.

2.4.10 $f(X) = \|AX - B\|_F^2$

Bằng kỹ thuật hoàn toàn tương tự như đã làm trong mục 2.4.4, ta thu được

$$\nabla_X \|AX - B\|_F^2 = 2A^T(AX - B)$$

2.5 Bảng các đạo hàm thường gặp

Bảng 2.1 bao gồm đạo hàm của các hàm số thường gặp với biến là vector hoặc đạo hàm.

2.6 Kiểm tra đạo hàm

Việc tính đạo hàm của hàm nhiều biến thông thường khá phức tạp và rất dễ mắc lỗi. Trong thực nghiệm, có một cách để kiểm tra liệu đạo hàm tính được có chính xác không. Cách này dựa trên định nghĩa của đạo hàm cho hàm một biến.

2.6.1 Xấp xỉ đạo hàm của hàm một biến

Theo định nghĩa,

$$f'(x) = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon) - f(x)}{\varepsilon} \tag{2.17}$$

Một cách thường được sử dụng là lấy một giá trị ε rất nhỏ, ví dụ 10^{-6}, và sử dụng công thức

$$f'(x) \approx \frac{f(x+\varepsilon) - f(x-\varepsilon)}{2\varepsilon} \tag{2.18}$$

Cách tính này được gọi là *numerical gradient*. Biểu thức (2.18) được sử dụng rộng rãi hơn để tính *numerical gradient*. Có hai cách giải thích cho vấn đề này.

Bằng giải tích

Chúng ta cùng quay lại một chút với khai triển Taylor. Với ε rất nhỏ, ta có hai xấp xỉ sau:

$$f(x+\varepsilon) \approx f(x) + f'(x)\varepsilon + \frac{f"(x)}{2}\varepsilon^2 + \frac{f^{(3)}}{6}\varepsilon^3 + \dots \tag{2.19}$$

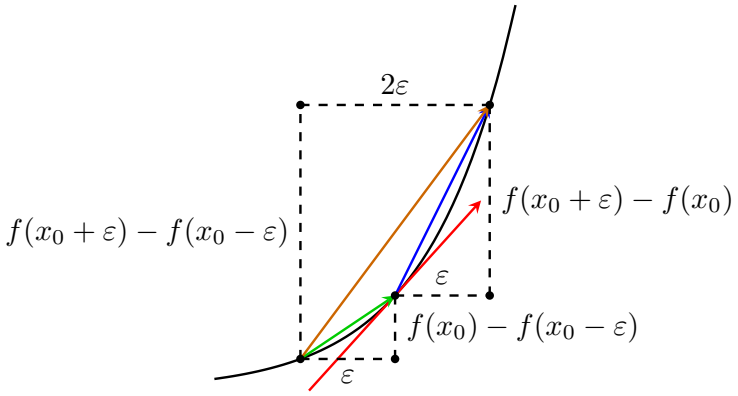
$$f(x-\varepsilon) \approx f(x) - f'(x)\varepsilon + \frac{f"(x)}{2}\varepsilon^2 - \frac{f^{(3)}}{6}\varepsilon^3 + \dots \tag{2.20}$$

Từ đó ta có:

$$\frac{f(x+\varepsilon) - f(x)}{\varepsilon} \approx f'(x) + \frac{f"(x)}{2}\varepsilon + \dots = f'(x) + O(\varepsilon) \tag{2.21}$$

$$\frac{f(x+\varepsilon) - f(x-\varepsilon)}{2\varepsilon} \approx f'(x) + \frac{f^{(3)}(x)}{6}\varepsilon^2 + \dots = f'(x) + O(\varepsilon^2) \tag{2.22}$$

trong đó $O()$ là *Big O notation*.



Hình 2.1: Giải thích cách xấp xỉ đạo hàm bằng hình học.

Từ đó, nếu xấp xỉ đạo hàm bằng công thức (2.21) (xấp xỉ đạo hàm phải), sai số sẽ là $O(\varepsilon)$. Trong khi đó, nếu xấp xỉ đạo hàm bằng công thức (2.22) (xấp xỉ đạo hàm hai phía), sai số sẽ là $O(\varepsilon^2)$. Khi ε rất nhỏ, $O(\varepsilon^2) \ll O(\varepsilon)$, tức cách đánh giá sử dụng công thức 2.22 có sai số nhỏ hơn, và vì vậy nó được sử dụng nhiều hơn.

Chúng ta cũng có thể giải thích điều này bằng hình học.

Bằng hình học

Quan sát Hình 2.1, vector màu đỏ là đạo hàm *chính xác* của hàm số tại điểm có hoành độ bằng x_0. Vector màu xanh lam và xanh lục lần lượt thể hiện cách xấp xỉ đạo hàm phía phải và phía trái. Vector màu nâu thể hiện cách xấp xỉ đạo hàm hai phía. Trong ba vector xấp xỉ đó, vector xấp xỉ hai phía màu nâu là gần với vector đỏ nhất nếu xét theo hướng.

Sự khác biệt giữa các cách xấp xỉ còn lớn hơn nữa nếu tại điểm x, hàm số bị *bẻ cong* mạnh hơn. Khi đó, xấp xỉ trái và phải sẽ khác nhau rất nhiều. Xấp xỉ hai bên sẽ *ổn định* hơn.

Từ đó ta thấy rằng xấp xỉ đạo hàm hai phía là xấp xỉ tốt hơn.

2.6.2 Xấp xỉ đạo hàm của hàm nhiều biến

Với hàm nhiều biến, công thức (2.22) được áp dụng cho từng biến khi các biến khác cố định. Cụ thể, ta sử dụng định nghĩa của hàm số nhận đầu vào là một ma trận như công thức (2.3). Mỗi thành phần của ma trận kết quả là đạo hàm của hàm số tại thành phần đó khi ta coi các thành phần còn lại cố định. Chúng ta sẽ thấy rõ điều này hơn ở cách lập trình so sánh hai cách tính đạo hàm ngay phía dưới.

Cách tính xấp xỉ đạo hàm theo phương pháp *numerical* thường cho giá trị khá chính xác. Tuy nhiên, cách này không được sử dụng để tính đạo hàm vì độ phức tạp quá cao so với cách tính trực tiếp. Tại mỗi thành phần, ta cần tính giá trị của hàm số tại phía trái và phía phải, như vậy sẽ không khả thi với các ma trận lớn. Khi so sánh đạo hàm *numerical* này với đạo hàm tính theo công thức, người ta thường giảm số chiều dữ liệu và giảm số điểm dữ liệu để thuận tiện cho tính toán. Nếu công thức đạo hàm ta tính được là chính xác, nó sẽ rất gần với đạo hàm *numerical*.

Đoạn Code 2.1 giúp kiểm tra đạo hàm của một hàm số khả vi $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, có kèm theo hai ví dụ. Để sử dụng hàm kiểm tra `check_grad` này, ta cần viết hai hàm. Hàm thứ nhất là hàm `fn(X)` tính giá trị của hàm số tại `X`. Hàm thứ hai là hàm `gr(X)` tính giá trị của đạo hàm mà ta cần kiểm tra.

```
from __future__ import print_function
import numpy as np

def check_grad(fn, gr, X):
    X_flat    = X.reshape(-1) # convert X to an 1d array -> 1 for loop needed
    shape_X   = X.shape               # original shape of X
    num_grad  = np.zeros_like(X)      # numerical grad, shape = shape of X
    grad_flat = np.zeros_like(X_flat) # 1d version of grad
    eps       = 1e-6            # a small number, 1e-10 -> 1e-6 is usually good
    numElems  = X_flat.shape[0] # number of elements in X
    # calculate numerical gradient
    for i in range(numElems):         # iterate over all elements of X
        Xp_flat      = X_flat.copy()
        Xn_flat      = X_flat.copy()
        Xp_flat[i]  += eps
        Xn_flat[i]  -= eps
        Xp           = Xp_flat.reshape(shape_X)
        Xn           = Xn_flat.reshape(shape_X)
        grad_flat[i] = (fn(Xp) - fn(Xn))/(2*eps)

    num_grad = grad_flat.reshape(shape_X)

    diff = np.linalg.norm(num_grad - gr(X))
    print('Difference between two methods should be small:', diff)

# ==== check if grad(trace(A*X)) == A^T ====
m, n = 10, 20
A = np.random.rand(m, n)
X = np.random.rand(n, m)

def fn1(X):
    return np.trace(A.dot(X))

def gr1(X):
    return A.T

check_grad(fn1, gr1, X)
# ==== check if grad(x^T*A*x) == (A + A^T)*x  ====
A = np.random.rand(m, m)
x = np.random.rand(m, 1)

def fn2(x):
    return x.T.dot(A).dot(x)

def gr2(x):
    return (A + A.T).dot(x)

check_grad(fn2, gr2, x)
```

Code 2.1: Kiểm tra đạo hàm bằng phương pháp numerical.

Kết quả:

```
Difference between two methods should be small: 2.02303323394e-08
Difference between two methods should be small: 2.10853872281e-09
```

Kết quả cho thấy sự khác nhau giữa Frobenious norm (mặc định của `np.linalg.norm`) của kết quả của hai cách tính là rất nhỏ. Sau khi chạy lại đoạn code với các giá trị `m, n` khác nhau và biến `X` khác nhau, nếu sự khác nhau vẫn là nhỏ, ta có thể tự tin rằng đạo hàm mà ta tính được là chính xác.

Bạn đọc có thể tự kiểm tra lại các công thức trong Bảng 2.1 theo phương pháp này.

Chương 3

Ôn tập Xác Suất

Chương này được viết dựa trên Chương 2 và 3 của cuốn Computer Vision: Models, Learning, and Inference–Simon J.D. Prince (http:// www.computervisionmodels.com).

3.1 Xác Suất

3.1.1 Random variables

Một *biến ngẫu nhiên* (*random variable*) x là một đại lượng dùng để đo những đại lượng không xác định. Biến này có thể được dùng để ký hiệu kết quả/đầu ra (*outcome*) của một thí nghiệm, ví dụ như tung đồng xu, hoặc một đại lượng biến đổi trong tự nhiên, ví dụ như nhiệt độ trong ngày. Nếu chúng ta quan sát rất nhiều đầu ra $\{x_i\}_{i=1}^{I}$ của các thí nghiệm này, ta có thể nhận được những giá trị khác nhau ở mỗi thí nghiệm. Tuy nhiên, sẽ có những giá trị xảy ra nhiều lần hơn những giá trị khác, hoặc xảy ra gần một giá trị này hơn những giá trị khác. Thông tin về đầu ra này được đo bởi một *phân phối xác suất* (*probaility distribution*) được biểu diễn bằng một hàm $p(x)$. Một biến ngẫu nhiên có thể là *rời rạc* (*discrete*) hoặc *liên tục* (*continuous*).

Một biến ngẫu nhiên rời rạc sẽ lấy giá trị trong một tập hợp các điểm rời rạc cho trước. Ví dụ tung đồng xu thì có hai khả năng là *head* và *tail*[1].Tập các giá trị này có thể là *có thứ tự* như khi tung xúc xắc hoặc *không có thứ tự*, ví dụ khi đầu ra là các giá trị *nắng*, *mưa*, *bão*. Mỗi đầu ra có một giá trị xác suất tương ứng với nó. Các giá trị xác suất này không âm và có tổng bằng một.

$$\text{Nếu } x \text{ là biến ngẫu nhiên rời rạc thì} \quad \sum_{x} p(x) = 1 \tag{3.1}$$

Biến ngẫu nhiên liên tục lấy các giá trị là các số thực. Những giá trị này có thể là hữu hạn, ví dụ thời gian làm bài của mỗi thí sinh trong một bài thi 180 phút, hoặc vô hạn, ví dụ thời

[1] đồng xu thường có một mặt có hình đầu người, được gọi là *head*, trái ngược với mặt này được gọi là mặt *tail*

gian phải chờ tới khách hàng tiếp theo. Không như biến ngẫu nhiên rời rạc, xác suất để đầu ra bằng *chính xác* một giá trị nào đó, theo lý thuyết, là bằng không. Thay vào đó, xác suất để đầu ra rời vào một khoảng giá trị nào đó là khác không. Việc này được mô tả bởi *hàm mật độ xác suất* (*probability density function - pdf*). Hàm mật độ xác suất luôn cho giá trị dương, và tích phân của nó trên toàn miền giá trị đầu ra *possible outcome* phải bằng một.

$$\text{Nếu } x \text{ là biến ngẫu nhiên liên tục thì} \quad \int p(x)dx = 1 \tag{3.2}$$

Nếu x là biến ngẫu nhiên rời rạc, thì $p(x) \leq 1, \forall x$. Trong khi đó, nếu x là biến ngẫu nhiên liên tục, $p(x)$ có thể nhận giá trị không âm bất kỳ, điều này vẫn đảm bảo là tích phân của hàm mật độ xác suất theo toàn bộ giá trị có thể có của x bằng một.

3.1.2 Xác suất đồng thời

Xét hai biến ngẫu nhiên x và y. Nếu ta quan sát rất nhiều cặp đầu ra của x và y, thì có những tổ hợp hai đầu ra xảy ra thường xuyên hơn những tổ hợp khác. Thông tin này được biểu diễn bằng một phân phối được gọi là *xác suất đồng thời* (*joint probability*) của x và y, được ký hiệu là $p(x, y)$, đọc là xác suất của x *và* y. Hai biến ngẫu nhiên x và y có thể đồng thời là biến ngẫu nhiên rời rạc, liên tục, hoặc một rời rạc, một liên tục. Luôn nhớ rằng tổng các xác suất trên mọi cặp giá trị có thể xảy ra (x, y) bằng một.

$$\text{Cả } x \text{ và } y \text{ là rời rạc:} \quad \sum_{x,y} p(x, y) = 1 \tag{3.3}$$

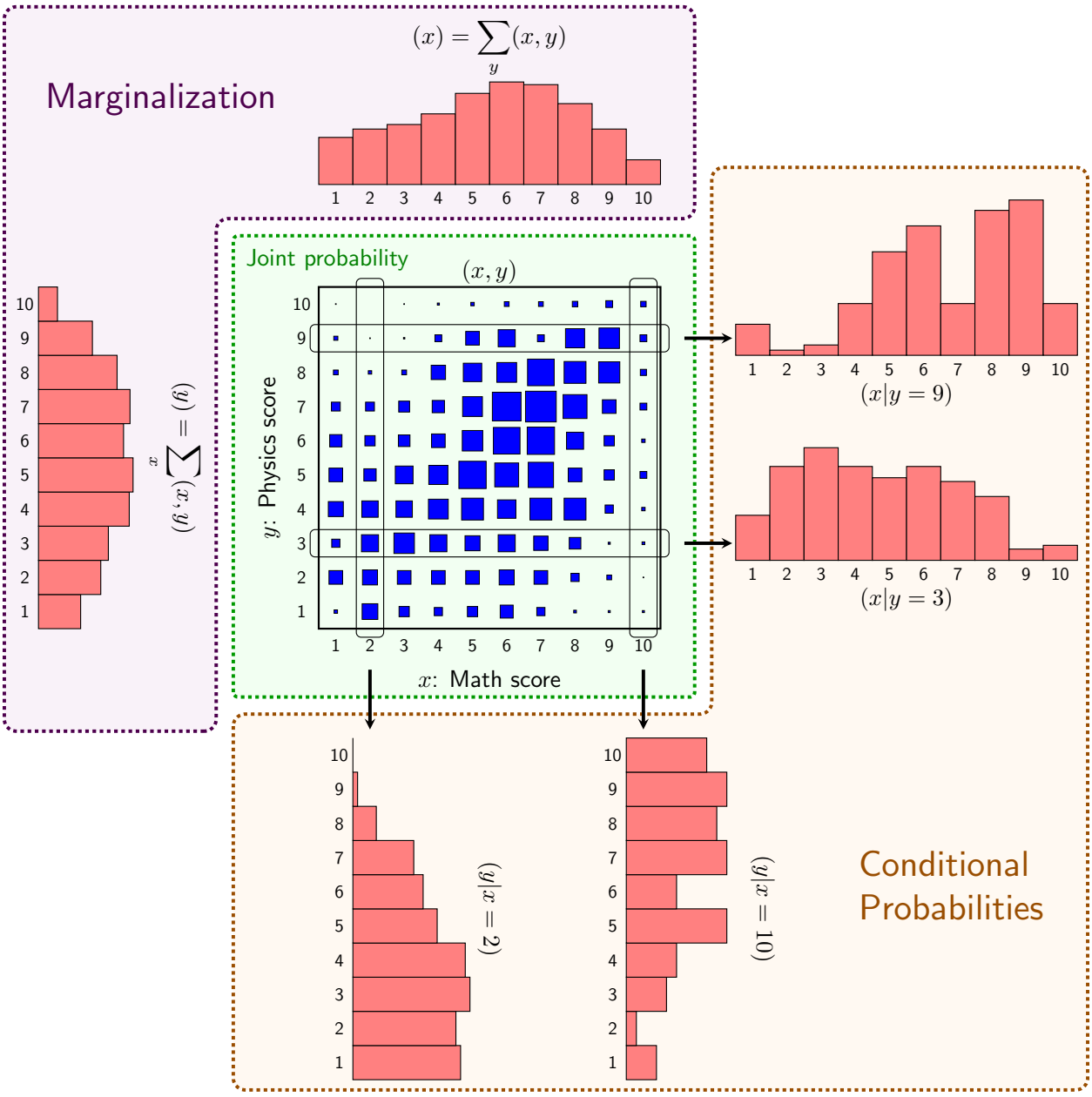
$$\text{Cả } x \text{ và } y \text{ là liên tục:} \quad \int p(x, y)dxdy = 1 \tag{3.4}$$

$$x \text{ rời rạc, } y \text{ liên tục:} \quad \sum_x \int p(x, y)dy = \int \left(\sum_x p(x, y) \right) dy = 1 \tag{3.5}$$

Xét ví dụ trong Hình 3.1, phần có nền màu lục nhạt. Biến ngẫu nhiên x thể hiện điểm thi môn Toán của học sinh ở một trường THPT trong một kỳ thi Quốc gia, biến ngẫu nhiên y thể hiện điểm thi môn Vật Lý cũng trong kỳ thi đó. Đại lượng $p(x = x^*, y = y^*)$ là tỉ lệ giữa tần suất số học sinh được *đồng thời* x^* điểm trong môn Toán và y^* điểm trong môn Vật Lý và toàn bộ số học sinh của trường đó. Tỉ lệ này có thể coi là xác suất khi số học sinh trong trường là lớn. Ở đây x^* và y^* là các số xác định. Thông thường, xác suất này được viết gọn lại thành $p(x^*, y^*)$, và $p(x, y)$ được dùng như một hàm tổng quát để mô tả các xác suất. Giả sử thêm rằng điểm các môn là các số tự nhiên từ 1 đến 10.

Các ô vuông màu lam thể hiện xác suất $p(x, y)$, với diện tích ô vuông càng to thể hiện xác suất đó càng lớn. Chú ý rằng tổng các xác suất này bằng một.

Các bạn có thể thấy rằng xác suất để một học sinh được 10 điểm một Toán và 1 điểm môn Lý rất thấp, điều tương tự xảy ra với 10 điểm môn Lý và 1 điểm môn Toán. Ngược lại, xác suất để một học sinh được khoảng 7 điểm cả hai môn là cao nhất.



Hình 3.1: Xác suất đồng thời (phần trung tâm có nền màu lục nhạt), Xác suất biên (phía trên và bên trái) và Xác suất có điền kiện (phía dưới và bên phải).

Thông thường, chúng ta sẽ làm việc với các bài toán ở đó xác suất có điều kiện được xác định trên nhiều hơn hai biến ngẫu nhiên. Chẳng hạn, $p(x, y, z)$ thể hiện joint probability của ba biến ngẫu nhiên x, y và z. Khi có nhiều biến ngẫu nhiên, ta có thể viết chúng dưới dạng vector. Cụ thể, ta có thể viết $p(\mathbf{x})$ để thể hiện xác suất có điều kiện của biến ngẫu nhiên nhiều chiều $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$. Khi có nhiều tập các biến ngẫu nhiên, ví dụ $\mathbf{x}$ và $\mathbf{y}$, ta có thể biết $p(\mathbf{x}, \mathbf{y})$ để thể hiện xác suất có điều kiện của tất cả các thành phần trong hai biến ngẫu nhiên nhiều chiều này.

3.1.3 Xác suất biên

Nếu biết xác suất đồng thời của nhiều biến ngẫu nhiên, ta cũng có thể xác định được phân phối xác suất của từng biến bằng cách lấy tổng với biến ngẫu nhiên rời rạc hoặc tích phân với biến ngẫu nhiên liên tục theo tất cả các biến còn lại:

$$\text{Nếu } x, y \text{ rời rạc}: \quad p(x) = \sum_y p(x, y) \tag{3.6}$$

$$p(y) = \sum_x p(x, y) \tag{3.7}$$

$$\text{Nếu } x, y \text{ liên tục}: \quad p(x) = \int p(x, y) dy \tag{3.8}$$

$$p(y) = \int p(x, y) dx \tag{3.9}$$

Với nhiều biến hơn, chẳng hạn bốn biến rời rạc x, y, z, w, cách tính được thực hiện tương tự:

$$p(x) = \sum_{y,z,w} p(x, y, z, w) \tag{3.10}$$

$$p(x, y) = \sum_{z,w} p(x, y, z, w) \tag{3.11}$$

Cách xác định xác suất của một biến dựa trên xác suất đồng thời của nó với các biến khác được gọi là *marginalization*. Phân phối đó được gọi là *xác suất biên* (*marginal probability*).

Từ đây trở đi, nếu không đề cập gì thêm, chúng ta sẽ dùng ký hiệu $\sum$ để chỉ chung cho cả hai loại biến. Nếu biến ngẫu nhiên là liên tục, bạn đọc ngầm hiểu rằng dấu $\sum$ cần được thay bằng dấu tích phân $\int$, biến lấy vi phân chính là biến được viết dưới dấu $\sum$. Chẳng hạn, trong (3.11), nếu z là liên tục, w là rời rạc, công thức đúng sẽ là

$$p(x, y) = \sum_w \left(\int p(x, y, z, w) dz \right) = \int \left(\sum_w p(x, y, z, w) \right) dz \tag{3.12}$$

Quay lại ví dụ trong Hình 3.1 với hai biến ngẫu nhiên rời rạc x, y. Lúc này, $p(x)$ được hiểu là xác suất để một học sinh đạt được x điểm môn Toán. Xác suất này được thể hiện ở khu vực có nền màu tím nhạt, phía trên. Nhắc lại rằng xác suất ở đây thực ra là tỉ lệ giữa số học sinh đạt x điểm môn Toán và toàn bộ số học sinh. Có hai cách tính xác suất này. Cách thứ nhất, dựa trên cách vừa định nghĩa, là đếm số học sinh được x điểm môn toán rồi chia cho tổng số học sinh. Cách tính thứ hai dựa trên xác suất đồng thời đã biết về xác suất để một học sinh được x điểm môn Toán và y điểm môn Lý. Số lượng học sinh đạt $x = x^*$ điểm môn Toán sẽ bằng tổng số lượng học sinh đạt $x = x^*$ điểm môn Toán *và* y điểm môn Lý, với y là một giá trị bất kỳ từ 1 đến 10. vì vậy, để tính xác suất $p(x)$, ta chỉ cần tính tổng của toàn bộ $p(x, y)$ với y chạy từ 1 đến 10. Tương tự nếu ta muốn tính $p(y)$ (xem phần bên trái của khu vực nền tím nhạt).

Dựa trên nhận xét này, mỗi giá trị của $p(x)$ chính bằng tổng các giá trị trong cột thứ x của hình vuông trung tâm nền xanh lục. Mỗi giá trị của $p(y)$ sẽ bằng tổng các giá trị trong hàng thứ y tính từ đưới lên. Chú ý rằng tổng các xác suất luôn bằng một.

3.1.4 Xác suất có điều kiện.

Dựa vào phân phối điểm của các học sinh, liệu ta có thể tính được xác suất để một học sinh được điểm 10 môn Lý, biết rằng học sinh đó được điểm 1 môn Toán?

Xác suất để một biến ngẫu nhiên x nhận một giá trị nào đó biết rằng biến ngẫu nhiên y có giá trị y^* được gọi là *xác suất có điều kiện* (*conditional probability*), được ký hiệu là $p(x|y=y^*)$.

Xác suất có điều kiện $p(x|y=y^*)$ có thể được tính dựa trên xác suất đồng thời $p(x,y)$. Quay lại Hình 3.1 với vùng có nền màu nâu nhạt. Nếu biết rằng $y=9$, xác suất $p(x|y=9)$ có thể tính được dựa trên hàng thứ chín của hình vuông trung tâm, tức hàng $p(x,y=9)$. Trong hàng này, những ô vuông lớn hơn thể hiện xác suất lớn hơn. Tương ứng như thế, $p(x|y=9)$ cũng lớn nếu $p(x,y=9)$ lớn. Chú ý rằng tổng các xác suất $\sum_x p(x,y=9)$ nhỏ hơn một, và bằng tổng các xác suất trên hàng thứ chín này. Để thoả mãn điều kiện tổng các xác suất bằng một, ta cần chia mỗi đại lượng $p(x,y=9)$ cho tổng của toàn hàng này. Tức là

$$p(x|y=9) = \frac{p(x,y=9)}{\sum\limits_x p(x,y=9)} = \frac{p(x,y=9)}{p(y=9)} \tag{3.13}$$

Tổng quát,

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{\sum\limits_x p(x,y=y^*)} = \frac{p(x,y=y^*)}{p(y=y^*)} \tag{3.14}$$

ở đây ta đã sử dụng công thức tính xác suất biên trong (3.7) cho mẫu số. Thông thường, ta có thể viết xác suất có điều kiện mà không cần chỉ rõ giá trị $y=y^*$ và có công thức gọn hơn:

$$p(x|y) = \frac{p(x,y)}{p(y)}, \text{ và tương tự, } p(y|x) = \frac{p(y,x)}{p(x)} \tag{3.15}$$

Từ đó ta có quan hệ

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x) \tag{3.16}$$

Khi có nhiều hơn hai biến ngẫu nhiên, ta có các công thức

$$\begin{aligned} p(x,y,z,w) &= p(x,y,z|w)p(w) && (3.17)\\ &= p(x,y|z,w)p(z,w) = p(x,y|z,w)p(z|w)p(w) && (3.18)\\ &= p(x|y,z,w)p(y|z,w)p(z|w)p(w) && (3.19) \end{aligned}$$

Công thức (3.19) có dạng *chuỗi* (*chain*) và được sử dụng nhiều sau này.

3.1.5 Quy tắc Bayes

Công thức (3.16) biểu diễn xác suất đồng thời theo hai cách. Từ đó ta có thể suy ra:

$$p(y|x)p(x) = p(x|y)p(y) \tag{3.20}$$

Biến đổi một chút:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{3.21}$$

$$= \frac{p(x|y)p(y)}{\sum_y p(x,y)} \tag{3.22}$$

$$= \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} \tag{3.23}$$

ở đó dòng thứ hai và thứ ba các công thức về xác suất biên và xác suất đồng thời ở mẫu số đã được sử dụng. Từ (3.23) ta có thể thấy rằng $p(y|x)$ hoàn toàn có thể tính được nếu ta biết mọi $p(x|y)$ và $p(y)$. Tuy nhiên, việc tính trực tiếp xác suất này thường là phức tạp.

Ba công thức (3.21)-(3.23) thường được gọi là *Quy tắc Bayes* (Bayes' rule). Chúng được sử dụng rộng rãi trong Machine Learning

3.1.6 Biến ngẫu nhiên độc lập

Nếu biết giá trị của một biến ngẫu nhiên x không mang lại thông tin về việc suy ra giá trị của biến ngẫu nhiên y (và ngược lại), thì ta nói rằng hai biến ngẫu nhiên là *độc lập* (*independent*). Chẳng hạn, chiều cao của một học sinh và điểm thi môn Toán của học sinh đó có thể coi là hai biến ngẫu nhiên *độc lập*.

Khi hai biến ngẫu nhiên x và y là *độc lập*, ta sẽ có:

$$p(x|y) = p(x) \tag{3.24}$$

$$p(y|x) = p(y) \tag{3.25}$$

Thay vào biểu thức xác suất đồng thời trong (3.16), ta có:

$$p(x,y) = p(x|y)p(y) = p(x)p(y) \tag{3.26}$$

3.1.7 Kỳ vọng và ma trận hiệp phương sai

Kỳ vọng (*expectation*) của một biến ngẫu nhiên được định nghĩa là

$$\mathrm{E}[x] = \sum_x xp(x) \quad \text{nếu } x \text{ là rời rạc} \tag{3.27}$$

$$\mathrm{E}[x] = \int xp(x)dx \quad \text{nếu } x \text{ là liên tục} \tag{3.28}$$

Giả sử $f(.)$ là một hàm số trả về một số với mỗi giá trị x^* của biến ngẫu nhiên x. Khi đó, nếu x là biến ngẫu nhiên rời rạc, ta sẽ có

$$\mathrm{E}[f(x)] = \sum_x f(x)p(x) \tag{3.29}$$

Công thức cho biến ngẫu nhiên liên tục cũng được viết tương tự.

Với xác suất đồng thời

$$\mathrm{E}[f(x,y)] = \sum_{x,y} f(x,y)p(x,y)dxdy \tag{3.30}$$

Có ba tính chất cần nhớ về kỳ vọng:

1. Kỳ vọng của một hằng số theo một biến ngẫu nhiên x bất kỳ bằng chính hằng số đó:

$$\mathrm{E}[\alpha] = \alpha \tag{3.31}$$

2. Kỳ vọng có tính chất tuyến tính:

$$\mathrm{E}[\alpha x] = \alpha \mathrm{E}[x] \tag{3.32}$$
$$\mathrm{E}[f(x) + g(x)] = \mathrm{E}[f(x)] + \mathrm{E}[g(x)] \tag{3.33}$$

3. Kỳ vọng của tích hai biến ngẫu nhiên bằng tích kỳ vọng của hai biến đó **nếu hai biến ngẫu nhiên đó là độc lập**.

$$\mathrm{E}[f(x)g(y)] = \mathrm{E}[f(x)]\mathrm{E}[g(y)] \tag{3.34}$$

Khái niệm kỳ vọng thường đi kèm với khái niệm *phương sai*(*variance*) trong không gian một chiều, và *ma trận hiệp phương sai* (*covariance matrix*) trong không gian nhiều chiều.

Với dữ liệu một chiều

Cho N giá trị $x_1, x_2, \ldots, x_N$. *Kỳ vọng* và *phương sai* của bộ dữ liệu này được tính theo công thức:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\mathbf{x1} \tag{3.35}$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \tag{3.36}$$

với $\mathbf{x} = \begin{bmatrix} x_1, x_2, \ldots, x_N \end{bmatrix}$, và $\mathbf{1} \in \mathbb{R}^N$ là vector cột chứa toàn phần tử 1. Kỳ vọng đơn giản là trung bình cộng của toàn bộ các giá trị. Phương sai là trung bình cộng của bình phương khoảng cách từ mỗi điểm tới kỳ vọng. Phương sai càng nhỏ thì các điểm dữ liệu càng gần với kỳ vọng, tức các điểm dữ liệu càng giống nhau. Phương sai càng lớn thì ta nói dữ liệu càng có tính phân tán. Ví dụ về kỳ vọng và phương sai của dữ liệu một chiều có thể được thấy trong Hình 3.2a. Căn bậc hai của phương sai, σ còn được gọi là *độ lệch chuẩn* (*standard deviation*) của dữ liệu.

Với dữ liệu nhiều chiều

Cho N điểm dữ liệu được biểu diễn bởi các vector cột $\mathbf{x}_1, \dots, \mathbf{x}_N$, khi đó, *vector kỳ vọng* và *ma trận hiệp phương sai* của toàn bộ dữ liệu được định nghĩa là:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{3.37}$$

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T \tag{3.38}$$

Trong đó $\hat{\mathbf{X}}$ được tạo bằng cách trừ mỗi cột của $\mathbf{X}$ đi $\bar{\mathbf{x}}$:

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}} \tag{3.39}$$

Một vài tính chất của ma trận hiệp phương sai:

- Ma trận hiệp phương sai là một ma trận đối xứng, hơn nữa, nó là một ma trận nửa xác định dương.
- Mọi phần tử trên đường chéo của ma trận hiệp phương sai là các số không âm. Chúng cũng chính là phương sai của từng chiều của dữ liệu.
- Các phần tử ngoài đường chéo $s_{ij}, i \neq j$ thể hiện sự tương quan giữa thành phần thứ i và thứ j của dữ liệu, còn được gọi là hiệp phương sai. Giá trị này có thể dương, âm hoặc bằng không. Khi nó bằng không, ta nói rằng hai thành phần i, j trong dữ liệu là *không tương quan* (*uncorrelated*).
- Nếu ma trận hiệp phương sai là ma trận đường chéo, ta có dữ liệu hoàn toàn không tương quan giữa các chiều.

Ví dụ về dữ liệu không tương quan và tương quan được cho trong Hình 3.2b và 3.2c.

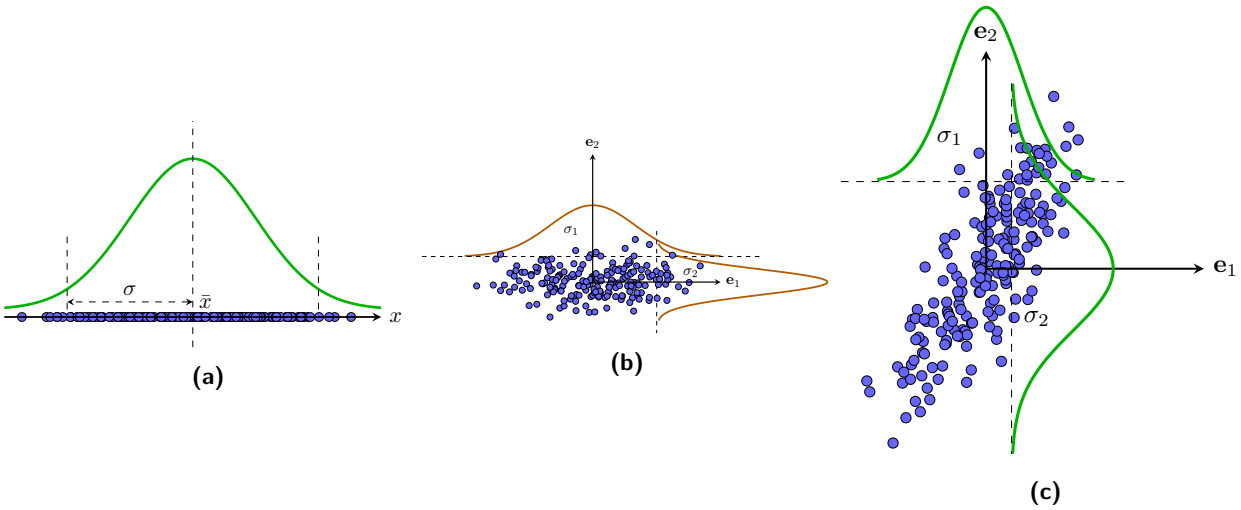
3.2 Một vài phân phối thường gặp

3.2.1 Phân phối Bernoulli

Phân phối Bernoulli là một phân phối rời rạc mô tả các biến ngẫu nhiên nhị phân: trường hợp đầu ra chỉ nhận một trong hai giá trị $x \in \{0, 1\}$. Hai giá trị này có thể là *head* và *tail* khi tung đồng xu; có thể là *giao dịch lừa đảo* và *giao dịch thông thường* trong bài toán xác định giao dịch lừa đảo trong tín dụng; có thể là *người* và *không phải người* trong bài toán tìm xem trong một bức ảnh có người hay không.

Bernoulli distribution được mô tả bằng một tham số $\lambda \in [0, 1]$ và là xác suất để biến ngẫu nhiên $x = 1$. Xác suất của mỗi đầu ra sẽ là

$$p(x = 1) = \lambda, \quad p(x = 0) = 1 - p(x = 1) = 1 - \lambda \tag{3.40}$$



Hình 3.2: Ví dụ về kỳ vọng và phương sai. (a) Trong không gian một chiều. (b) Trong không gian hai chiều mà hai chiều không tương quan. Trong trường hợp này, ma trận hiệp phương sai là ma trận đường chéo với hai phần tử trên đường chéo là σ_1, σ_2, đây cũng chính là hai trị riêng của ma trận hiệp phương sai và là phương sai của mỗi chiều dữ liệu. (c) Dữ liệu trong không gian hai chiều có tương quan. Theo mỗi chiều, ta có thể tính được kỳ vọng và phương sai. Phương sai càng lớn thì dữ liệu trong chiều đó càng phân tán. Trong ví dụ này, dữ liệu theo chiều thứ hai phân tán nhiều hơn so so với chiều thứ nhất.

Hai đẳng thức này thường được viết gọn lại:

$$p(x) = \lambda^x(1-\lambda)^{1-x} \tag{3.41}$$

với giả định rằng $0^0 = 1$. Thật vậy, $p(0) = \lambda^0(1-\lambda)^1 = 1-\lambda$, và $p(1) = \lambda^1(1-\lambda)^0 = \lambda$.

Phân phối Bernoulli thường được ký hiệu ngắn gọn dưới dạng

$$p(x) = \text{Bern}_x[\lambda] \tag{3.42}$$

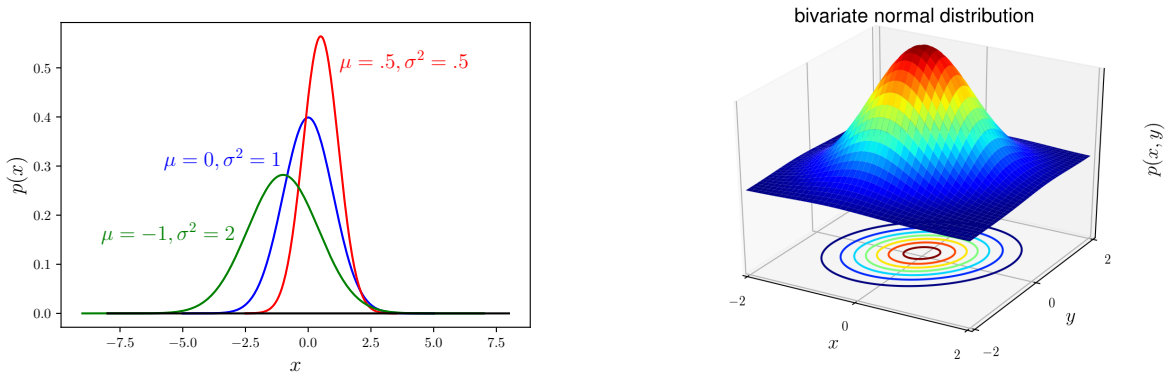
3.2.2 Phân phối Categorical

Trong nhiều trường hợp, đầu ra của biến ngẫu nhiên rời rạc có thể là một trong nhiều hơn hai giá trị khác nhau. Ví dụ, một bức ảnh có thể chứa một chiếc xe, một người, hoặc một con mèo. Khi đó, ta dùng một phân phối tổng quát của phân phối Bernoulli, được gọi là *phân phối Categorical*. Các đầu ra được mô tả bởi một phần tử trong tập hợp $\{1, 2, \ldots, K\}$.

Nếu có K đầu ra, phân phối Categorical sẽ được mô tả bởi K tham số, viết dưới dạng vector: $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ với các λ_k không âm và có tổng bằng một. Mỗi giá trị λ_k thể hiện xác suất để đầu ra nhận giá trị k: $p(x = k) = \lambda_k$.

Phân phối Categorical thường được ký hiệu dưới dạng:

$$p(x) = \text{Cat}_x[\lambda] \tag{3.43}$$



(a) (b)

Hình 3.3: Ví dụ về hàm mật độ xác suất của (a) phân phối chuẩn một chiều, và (b) phân phối chuẩn hai chiều.

Nếu thay vì biểu diễn đầu ra là một số k trong tập hợp $\{1, 2, \ldots, K\}$, ta biểu diễn đầu ra là một vector ở dạng *one-hot*, tức một vector K phần tử với chỉ phần tử thứ k bằng một, các phần tử còn lại bằng không. Nói cách khác, tập hợp các đầu ra là tập hợp các vector đơn vị bậc K: $\mathbf{x} \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ với $\mathbf{e}_k$ là vector đơn vị thứ k. Khi đó, ta sẽ có

$$p(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k \tag{3.44}$$

Khi $\mathbf{x} = \mathbf{e}_k, x_k = 1, x_j = 0, \ \forall j \neq k$. Thay vào (3.44) ta sẽ được $p(\mathbf{x} = \mathbf{e}_k) = \lambda_k = p(x = k)$.

3.2.3 Phân phối chuẩn một chiều

Phân phối chuẩn một chiều (*univariate normal* hoặc *Gaussian distribution*) được định nghĩa trên các biến liên tục nhận giá trị $x \in (-\infty, \infty)$. Đây là một phân phối được sử dụng nhiều nhất với các biến ngẫu nhiên liên tục. Phân phối này được mô tả bởi hai tham số: *kỳ vọng* μ và *phương sai* (*variance*) σ^2. Giá trị μ có thể là bất kỳ số thực nào, thể hiện vị trí của giá trị mà tại đó mà hàm mật độ xác suất đạt giá trị cao nhất. Giá trị σ^2 là một giá trị dương, với σ thể hiện *độ rộng* của phân phối này. σ lớn chứng tỏ khoảng giá trị đầu ra có khoảng biến đổi mạnh, và ngược lại.

Hàm mật độ xác suất của phân phối này được định nghĩa là

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{3.45}$$

Hoặc được viết gọn hơn dưới dạng $p(x) = \text{Norm}_x[\mu, \sigma^2]$, hoặc $\mathcal{N}(\mu, \sigma^2)$.

Ví dụ về đồ thị hàm mật độ xác suất của phân phối chuẩn một chiều được cho trên Hình 3.3a.

3.2.4 Phân phối chuẩn nhiều chiều

Phân phối này là trường hợp tổng quát của phân phối chuẩn khi biến ngẫu nhiên là nhiều chiều, giả sử là D chiều. Có hai tham số mô tả phân phối này: *vector kỳ vọng* $\boldsymbol{\mu} \in \mathbb{R}^D$ và *ma trận hiệp phương sai* $\boldsymbol{\Sigma} \in \mathbb{S}^D$ là một ma trận *đối xứng xác định dương*.

Hàm mật độ xác suất có dạng

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{3.46}$$

với $|\boldsymbol{\Sigma}|$ là định thức của ma trận hiệp phương sai $\boldsymbol{\Sigma}$.

Phân phối này thường được viết gọn lại dưới dạng $p(\mathbf{x}) = \text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, hoặc $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Ví dụ về hàm mật độ xác suất của một phân phối chuẩn hai chiều (*bivariate normal distribution*) được mô tả bởi một mặt cong cho trên Hình 3.3b. Nếu cắt mặt này theo các mặt phẳng song song với mặt đáy, ta sẽ thu được các hình ellipse đồng tâm.

3.2.5 Phân phối Beta

Phân phối Beta (*Beta distribution*) là một phân phối liên tục được định nghĩa trên một biến ngẫu nhiên $\lambda \in [0, 1]$ Phân phối Beta distribution được dùng để mô tả *tham số* cho một distribution khác. Cụ thể, phân phối này phù hợp với việc mô tả sự *biến động* của tham số λ trong phân phối Bernoulli.

Phân phối Beta được mô tả bởi hai tham số *dương* α, β. Hàm mật độ xác suất của nó là

$$p(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \tag{3.47}$$

với $\Gamma(.)$ là hàm số gamma, được định nghĩa là

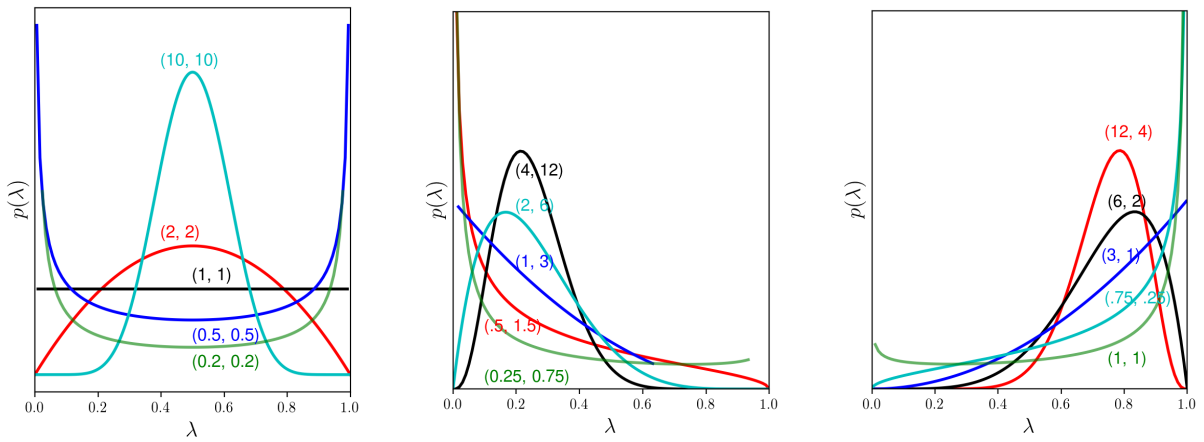
$$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)dt \tag{3.48}$$

Trên thực tế, việc tính giá trị của hàm số gamma không thực sự quan trọng vì nó chỉ mang tính chuẩn hoá để tổng xác suất bằng một.

Dạng gọn của phân phối Beta: $p(\lambda) = \text{Beta}_\lambda[\alpha, \beta]$

Hình 3.4 minh hoạ các hàm mật độ xác suất của phân phối Beta với các cặp giá trị (α, β) khác nhau.

- Trong Hình 3.4a, khi $\alpha = \beta$. Đồ thị của các hàm mật độ xác suất đối xứng qua đường thẳng $\lambda = 0.5$. Khi $\alpha = \beta = 1$, thay vào (3.47) ta thấy $p(\lambda) = 1$ với mọi λ. Trong trường



Hình 3.4: Ví dụ về hàm mật độ xác suất của phân phối Beta. (a) $\alpha = \beta$, đồ thị hàm số là đối xứng. (b) $\alpha < \beta$, đồ thị hàm số lệch sang trái, chứng tỏ xác suất λ nhỏ là lớn. (c) $\alpha > \beta$, đồ thị hàm số lệch sang phải, chứng tỏ xác suất λ lớn là lớn.

hợp này, phân phối Beta trở thành *phân phối đều* (*uniform distribution*). Khi $\alpha = \beta > 1$, các hàm số đạt giá trị cao tại gần trung tâm, tức là khả năng cao là λ sẽ nhận giá trị xung quanh điểm 0.5. Khi $\alpha = \beta < 1$, hàm số đạt giá trị cao tại các điểm gần 0 và 1.

- Trong Hình 3.4b, khi $\alpha < \beta$, ta thấy rằng đồ thị có xu hướng lệch sang bên trái. Các giá trị (α, β) này nên được sử dụng nếu ta dự đoán rằng λ là một số nhỏ hơn 0.5.

- Trong Hình 3.4c, khi $\alpha > \beta$, điều ngược lại xảy ra với các hàm số đạt giá trị cao tại các điểm gần 1.

3.2.6 Phân phối Dirichlet

Phân phối Dirichlet chính là trường hợp tổng quát của phân phối Beta khi được dùng để mô tả tham số của phân phối Categorical. Nhắc lại rằng phân phối Categorical là trường hợp tổng quát của phân phối Bernoulli.

Phân phối Dirichlet được định nghĩa trên K biến liên tục $\lambda_1, \ldots, \lambda_K$ trong đó các λ_k không âm và có tổng bằng một. Bởi vậy, nó phù hợp để mô tả tham số của phân phối Categorical. Có K tham số *dương* để mô tả một phân phối Dirichlet: $\alpha_1, \ldots, \alpha_K$.

Hàm mật độ xác suất của phân phối Dirichlet là

$$p(\lambda_1, \ldots, \lambda_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1} \tag{3.49}$$

Cách biểu diễn ngắn gọn: $p(\lambda_1, \ldots, \lambda_K) = \text{Dir}_{\lambda_1, \ldots, \lambda_K}[\alpha_1, \ldots, \alpha_K]$

Chương 4

Maximum Likelihood và Maximum A Posteriori

4.1 Giới thiệu

Có rất nhiều mô hình machine learning được xây dựng dựa trên các mô hình thống kê (*statistical models*). Các mô hình thống kê thường dựa trên các phân phối xác suất đã được đề cập trong Chương 3. Với phân phối Bernoulli, tham số là biến λ. Với phân phối chuẩn nhiều chiều, các tham số là mean vector $\boldsymbol{\mu}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}$. Với một mô hình thông kê bất kỳ, ký hiệu θ là tập hợp tất cả các tham số của mô hình đó. Learning chính là quá trình *ước lượng* (*estimate*) bộ tham số θ sao cho mô hình tìm được khớp với phân phối của dữ liệu nhất. Quá trình này còn được gọi là *ước lượng tham số* (*parameter estimation*).

Có hai cách ước lượng tham số thường được dùng trong các mô hình machine learning thống kê. Cách thứ nhất chỉ dựa trên dữ liệu đã biết trong tập huấn luyện, được gọi là *maximum likelihood estimation* hay *ML estimation* hoặc *MLE*. Cách thứ hai không những dựa trên tập huấn luyện mà còn dựa trên những thông tin biết trước của các tham số. Những thông tin này có thể có được bằng *cảm quan* của người xây dựng mô hình. *Cảm quan* càng rõ ràng, càng hợp lý thì khả năng thu được bộ tham số tốt là càng cao. Chẳng hạn, thông tin biết trước của λ trong Bernoulli distribution là việc nó là một số trong đoạn $[0, 1]$. Với bài toán tung đồng xu, với λ là xác suất có được mặt *head*, ta dự đoán được rằng giá trị này nên là một số gần với 0.5. Cách ước lượng tham số thứ hai này được gọi là *maximum a posteriori estimation* hay *MAP estimation*. Trong chương này, chúng ta cùng tìm hiểu ý tưởng và cách giải quyết bài toán ước lượng tham số mô hình theo *MLE* hoặc *MAP Estimation*.

4.2 Maximum likelihood estimation

4.2.1 Ý tưởng

Giả sử có các điểm dữ liệu $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Giả sử thêm rằng ta đã biết các điểm dữ liệu này tuân theo một phân phối nào đó được mô tả bởi bộ tham số θ.

Maximum likelihood estimation là việc đi tìm bộ tham số θ sao cho xác suất sau đây đạt giá trị lớn nhất:

$$\theta = \max_{\theta} p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta) \tag{4.1}$$

Biểu thức (4.1) có ý nghĩa như thế nào và vì sao việc này có lý?

Giả sử rằng ta đã biết dạng của mô hình, và mô hình này được mô tả bởi bộ tham số θ. Như vậy, $p(\mathbf{x}_1|\theta)$ chính là xác suất xảy ra *sự kiện* $\mathbf{x}_1$ biết rằng mô hình được mô tả bởi bộ tham số θ (đây là một xác suất có điều kiện). Và $p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)$ chính là xác suất để toàn bộ các *sự kiện* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ đồng thời xảy ra, xác suất đồng thời này còn được gọi là *likelihood*. Ở đây, *likelihood* chính là hàm mục tiêu.

Bởi vì sự việc đã xảy ra, tức dữ liệu huấn luyện bản thân chúng đã như thế, xác suất đồng thời này cần phải càng cao càng tốt. Việc này cũng giống như việc đã biết *kết quả*, và ta cần đi tìm *nguyên nhân* sao cho xác suất xảy ra kết quả càng cao càng tốt. MLE chính là việc đi tìm bộ tham số θ sao cho Likelihood là lớn nhất. Trong mô hình này ta cũng có một bài toán tối ưu với hàm mục tiêu là $p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)$. Lúc này ta không tối thiểu hàm mục tiêu mà cần tối đa nó, vì ta muốn rằng xác suất xảy ra việc này là lớn nhất.

4.2.2 Giả sử về sự độc lập và log-likelihood

Việc giải trực tiếp bài toán (4.1) thường là phức tạp vì việc đi tìm mô hình xác suất đồng thời cho toàn bộ dữ liệu là ít khi khả thi. Một cách tiếp cận phổ biến là giả sử đơn giản rằng các điểm dữ liệu $\mathbf{x}_n$ là độc lập với nhau. Nói cách khác, ta xấp xỉ likelihood trong (4.1) bởi

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta) \approx \prod_{n=1}^{N} p(\mathbf{x}_n|\theta) \tag{4.2}$$

(Nhắc lại rằng hai sự kiện x, y là độc lập nếu xác suất đồng thời của chúng bằng tích xác suất của từng sự kiện: $p(x, y) = p(x)p(y)$. Và khi là xác suất có điều kiện: $p(x, y|z) = p(x|z)p(y|z)$.) Lúc đó, bài toán (4.1) có thể được giải quyết bằng cách giải bài toán tối ưu sau:

$$\theta = \max_{\theta} \prod_{n=1}^{N} p(\mathbf{x}_n|\theta) \quad (3) \tag{4.3}$$

Việc tối ưu một tích thường phức tạp hơn việc tối ưu một tổng, vì vậy việc tối đa hàm mục tiêu thường được chuyển về việc tối đa log của hàm mục tiêu:

$$\theta = \max_{\theta} \sum_{n=1}^{N} \log\left(p(\mathbf{x}_n|\theta)\right) \tag{4.4}$$

Ôn lại một chút về hai tính chất của hàm logarit: (i) log của một tích bằng tổng của các log, và (ii) vì log là một hàm đồng biến, một biểu thức dương sẽ là lớn nhất nếu log của nó là lớn nhất, và ngược lại.

4.2.3 Ví dụ

Ví dụ 1: phân phối Bernoulli

Bài toán: giả sử tung một đồng xu N lần và nhận được n mặt *head*. Ước lượng xác suất khi tung đồng xu nhận được mặt *head*.

Lời giải:

Một cách trực quan, ta có thể ước lượng được rằng xác suất đó chính là $\lambda = \frac{n}{N}$. Chúng ta cùng ước lượng giá trị này sử dụng MLE.

Giả sử λ là xác suất để nhận được một mặt *head*. Đặt $x_1, x_2, \ldots, x_N$ là các đầu ra nhận được, trong đó có n giá trị bằng 1 tương ứng với mặt *head* và $m = N - n$ giá trị bằng 0 tương ứng với mặt *tail*. Ta có thể suy ra ngay rằng

$$\sum_{i=1}^{N} x_i = n, \quad N - \sum_{i=1}^{N} x_i = N - n = m \tag{4.5}$$

Vì đây là một xác suất của biến ngẫu nhiên nhị phân rời rạc, ta có thể nhận thấy việc nhận được mặt *head* hay *tail* khi tung đồng xu tuân theo phân phối Bernoulli:

$$p(x_i|\lambda) = \lambda^{x_i}(1-\lambda)^{1-x_i} \tag{4.6}$$

Khi đó tham số mô hình λ có thể được ước lượng bằng việc giải bài toán tối ưu sau đây, với giả sử rằng kết quả của các lần tung đồng xu là độc lập với nhau:

$$\lambda = \underset{\lambda}{\text{argmax}} \left[p(x_1, x_2, \ldots, x_N|\lambda)\right] = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} p(x_i|\lambda)\right] \tag{4.7}$$

$$= \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} \lambda^{x_i}(1-\lambda)^{1-x_i}\right] = \underset{\lambda}{\text{argmax}} \left[\lambda^{\sum_{i=1}^{N} x_i}(1-\lambda)^{N-\sum_{i=1}^{N} x_i}\right] \tag{4.8}$$

$$= \underset{\lambda}{\text{argmax}} \left[\lambda^n(1-\lambda)^m\right] = \underset{\lambda}{\text{argmax}} \left[n\log(\lambda) + m\log(1-\lambda)\right] \tag{4.9}$$

trong (4.9), ta đã lấy log của hàm mục tiêu. Tới đây, bài toán tối ưu (4.9) có thể được giải bằng cách lấy đạo hàm của hàm mục tiêu bằng 0. Tức λ là nghiệm của phương trình

$$\frac{n}{\lambda} - \frac{m}{1-\lambda} = 0 \Leftrightarrow \frac{n}{\lambda} = \frac{m}{1-\lambda} \Leftrightarrow \lambda = \frac{n}{n+m} = \frac{n}{N} \tag{4.10}$$

Vậy kết quả ta ước lượng ban đầu là có cơ sở.

Ví dụ 2: Categorical distribution

Một ví dụ khác phức tạp hơn một chút.

Bài toán: giả sử tung một viên xúc xắc sáu mặt có xác suất rơi vào các mặt có thể không đều nhau. Giả sử trong N lần tung, số lượng xuất hiện các mặt thứ nhất, thứ hai,..., thứ sáu lần lượt là $n_1, n_2, \ldots, n_6$ lần với $\sum_{i=1}^{6} n_i = N$. Tính xác suất rơi vào mỗi mặt ở lần tung tiếp theo. Giả sử thêm rằng $n_i > 0,\ \forall i = 1, \ldots, 6$.

Lời giải:

Bài toán này có vẻ phức tạp hơn bài toán trên một chút, nhưng ta cũng có thể dự đoán được ước lượng tốt nhất của xác suất rơi vào mặt thứ i là $\lambda_i = \frac{n_i}{N}$.

Mã hoá mỗi quan sát đầu ra thứ i bởi một vector 6 chiều $\mathbf{x}_i \in \{0, 1\}^6$ trong đó các phần tử của nó bằng 0 trừ phần tử tương ứng với mặt quan sát được là bằng 1. Nhận thấy rằng $\sum_{i=1}^{N} x_i^j = n_j,\ \forall j = 1, 2, \ldots, 6$, trong đó x_i^j là thành phần thứ j của vector $\mathbf{x}_i$.

Có thể thấy rằng xác suất rơi vào mỗi mặt tuân theo phân phối categorical với các tham số $\lambda_j > 0, j = 1, 2, \ldots, 6$. Ta dùng $\boldsymbol{\lambda}$ để thể hiện cho cả sáu tham số này.

Với các tham số $\boldsymbol{\lambda}$, xác suất để sự kiện $\mathbf{x}_i$ xảy ra là

$$p(\mathbf{x}_i|\lambda) = \prod_{j=1}^{6} \lambda_j^{x_i^j} \tag{4.11}$$

Khi đó, vẫn với giả sử về sự độc lập giữa các lần tung xúc xắc, ước lượng bộ tham số λ dựa trên việc tối đa log-likelihood ta có:

$$\boldsymbol{\lambda} = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\lambda) \right] = \underset{\lambda}{\text{argmax}} \left[\prod_{i=1}^{N} \prod_{j=1}^{6} \lambda_j^{x_i^j} \right] \tag{4.12}$$

$$= \underset{\lambda}{\text{argmax}} \left[\prod_{j=1}^{6} \lambda_j^{\sum_{i=1}^{N} x_i^j} \right] = \underset{\lambda}{\text{argmax}} \left[\prod_{j=1}^{6} \lambda_j^{n_j} \right] \tag{4.13}$$

$$= \underset{\lambda}{\text{argmax}} \left[\sum_{j=1}^{6} n_j \log(\lambda_j) \right] \tag{4.14}$$

Khác với bài toán (4.9) một chút, chúng ta không được quên điều kiện $\sum_{j=1}^{6} \lambda_j = 1$. Ta có bài toán tối ưu có ràng buộc sau đây

$$\max_{\lambda} \sum_{j=1}^{6} n_j \log(\lambda_j) \quad \text{thoả mãn:} \sum_{j=1}^{6} \lambda_j = 1 \tag{4.15}$$

Bài toán tối ưu này có thể được giải bằng phương pháp nhân tử Lagrange (xem Phụ lục **??**).

Lagrangian của bài toán này là

$$\mathcal{L}(\lambda, \mu) = \sum_{j=1}^{6} n_j \log(\lambda_j) + \mu(1 - \sum_{j=1}^{6} \lambda_j) \tag{4.16}$$

Nghiệm của bài toán là nghiệm của hệ đạo hàm của $\mathcal{L}(.)$ theo từng biến bằng 0

$$\frac{\partial \mathcal{L}(\lambda, \mu)}{\partial \lambda_j} = \frac{n_j}{\lambda_j} - \mu \quad = 0, \ \forall j = 1, 2, \ldots, 6 \tag{4.17}$$

$$\frac{\partial \mathcal{L}(\lambda, \mu)}{\partial \mu} = 1 - \sum_{j=1}^{6} \lambda_j = 0 \tag{4.18}$$

Từ (4.17) ta có $\lambda_j = \frac{n_j}{\mu}$. Thay vào (4.18),

$$\sum_{j=1}^{6} \frac{n_j}{\mu} = 1 \Rightarrow \mu = \sum_{j=1}^{6} n_j = N \tag{4.19}$$

Từ đó ta có ước lượng $\lambda_j = \frac{n_j}{N}, \quad \forall j = 1, 2, \ldots, 6.$

Qua hai ví dụ trên ta thấy MLE cho hết quả khá hợp lý.

Ví dụ 3: Univariate normal distribution

Bài toán: Khi thực hiện một phép đo, giả sử rằng rất khó để có thể đo *chính xác* độ dài của một vật. Thay vào đó, người ta thường đo vật đó nhiều lần rồi suy ra kết quả, với giả thiết rằng các phép đo là độc lập với nhau và kết quả mỗi phép đo là một phân phối chuẩn. Ước lượng chiều dài của vật đó dựa trên các kết quả đo được.

Lời giải:Vì biết rằng kết quả phép đo tuân theo phân phối chuẩn, ta sẽ cố gắng đi xây dựng phân phối chuẩn đó. Chiều dài của vật có thể được coi là giá trị mà hàm mật độ xác suất đạt giá trị cao nhất, tức khả năng rơi vào khoảng giá trị xung quanh nó là lớn nhất. Trong phân phối chuẩn, ta biết rằng hàm mật độ xác suất đạt giá trị lớn nhất tại chính kỳ vọng của phân phối đó. Chú ý rằng kỳ vọng của phân phối và kỳ vọng của dữ liệu quan sát được có thể không chính xác bằng nhau, nhưng rất gần nhau. Nếu ước lượng kỳ vọng của phân phối như cách làm dưới đây sử dụng MLE, ta sẽ thấy rằng kỳ vọng của dữ liệu chính là đánh giá tốt nhất cho kỳ vọng của phân phối.

Thật vậy, giả sử các kích thước quan sát được là $x_1, x_2, \ldots, x_N$. Ta cần đi tìm một phân phối chuẩn, được mô tả bởi một giá trị kỳ vọng μ và phương sai σ^2, sao cho các giá trị $x_1, x_2, \ldots, x_N$ là *likely nhất*. Ta đã biết rằng, hàm mật độ xác suất tại x_i của môt phân phối chuẩn có kỳ vọng μ và phương sai σ^2 là

$$p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{4.20}$$

Vậy, để đánh giá μ và σ, ta sử dụng MLE với giả thiết rằng kết quả các phép đo là độc lập:

$$\mu, \sigma = \underset{\mu,\sigma}{\text{argmax}} \left[\prod_{i=1}^{N} p(x_i|\mu, \sigma^2) \right] \tag{4.21}$$

$$= \underset{\mu,\sigma}{\text{argmax}} \left[\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2} \right) \right] \tag{4.22}$$

$$= \underset{\mu,\sigma}{\text{argmax}} \left[-N\log(\sigma) - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2} \triangleq J(\mu,\sigma) \right] \tag{4.23}$$

Ta đã lấy log của hàm bên trong dấu ngoặc vuông của (4.22) để được (4.23), phần hằng số có chứa 2π cũng đã được bỏ đi vì nó không ảnh hưởng tới kết quả.

Để tìm μ và σ, ta giải hệ phương trình đạo hàm của $J(\mu, \sigma)$ theo mỗi biến bằng không:

$$\frac{\partial J}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \tag{4.24}$$

$$\frac{\partial J}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \tag{4.25}$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{N} x_i}{N}, \quad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \tag{4.26}$$

Kết quả thu được không có gì bất ngờ.

Ví dụ 4: Multivariate normal distribution

Bài toán: Giả sử tập dữ liệu ta thu được là các giá trị nhiều chiều $\mathbf{x}_1, \ldots, \mathbf{x}_N$ tuân theo phân phối chuẩn. Hãy đánh giá các tham số, vector kỳ vọng $\boldsymbol{\mu}$ và ma trận hiệp phương sai $\boldsymbol{\Sigma}$ của phân phối này dựa trên MLE, giả sử rằng các $\mathbf{x}_1, \ldots, \mathbf{x}_N$ là độc lập.

Lời giải: Việc chứng minh các công thức

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{N}\mathbf{x}_i}{N} \tag{4.27}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T \tag{4.28}$$

xin được dành lại cho bạn đọc như một bài tập nhỏ. Dưới đây là một vài gợi ý:

- Hàm mật độ xác suất của phân phối chuẩn nhiều chiều là

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\|\boldsymbol{\Sigma}\|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right) \tag{4.29}$$

  Chú ý rằng ma trận hiệp phương sai $\boldsymbol{\Sigma}$ là xác định dương nên có nghịch đảo.

- Một vài đạo hàm theo ma trận:

$$\nabla_{\mathbf{\Sigma}} \log |\mathbf{\Sigma}| = (\mathbf{\Sigma}^{-1})^T \triangleq \mathbf{\Sigma}^{-T} \quad \text{(chuyển vị của nghịch đảo)} \tag{4.30}$$

$$\nabla_{\mathbf{\Sigma}} (\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = -\mathbf{\Sigma}^{-T} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-T} \tag{4.31}$$

(Xem thêm Matrix Calculus, mục D.2.1 và D.2.4 tại *https://goo.gl/JKg631*.)

4.3 Maximum a Posteriori

4.3.1 Ý tưởng

Quay lại với ví dụ 1 về tung đồng xu. Nếu tung đồng xu 5000 lần và nhận được 1000 lần *head*, ta có thể đánh giá xác suất của *head* là 1/5 và việc đánh giá này là đáng tin vì số mẫu là lớn. Nếu tung 5 lần và chỉ nhận được 1 mặt *head*, theo MLE, xác suất để có một mặt *head* được đánh giá là 1/5. Tuy nhiên với chỉ 5 kết quả, ước lượng này là không đáng tin, nhiều khả năng việc đánh giá đã bị overfitting. Khi tập huấn luyện quá nhỏ (*low-training*) chúng ta cần phải quan tâm tới một vài giả thiết của các tham số. Trong ví dụ này, giả thiết của chúng ta là xác suất nhận được mặt *head* phải gần 1/2.

Maximum A Posteriori (MAP) ra đời nhằm giải quyết vấn đề này. Trong MAP, chúng ta giới thiệu một giả thiết biết trước, được gọi là *prior*, của tham số θ. Từ giả thiết này, chúng ta có thể suy ra các khoảng giá trị và phân bố của tham số.

Ngược với MLE, trong MAP, chúng ta sẽ đánh giá tham số như là một xác suất có điều kiện của dữ liệu:

$$\theta = \arg\max_{\theta} \underbrace{p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N)}_{\text{posterior}} \tag{4.32}$$

Biểu thức $p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N)$ còn được gọi là *xác suất posterior* của θ. Chính vì vậy mà việc ước lượng θ theo (4.32) được gọi là *Maximum A Posteriori*.

Thông thường, hàm tối ưu trong (4.32) khó xác định dạng một cách trực tiếp. Chúng ta thường biết điều ngược lại, tức nếu biết tham số, ta có thể tính được hàm mật độ xác suất của dữ liệu. Vì vậy, để giải bài toán MAP, ta thường sử dụng quy tắc Bayes. Bài toán MAP thường được biến đổi thành

$$\theta = \arg\max_{\theta} p(\theta|\mathbf{x}_1, \dots, \mathbf{x}_N) = \arg\max_{\theta} \left[\frac{\overbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_N|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathbf{x}_1, \dots, \mathbf{x}_N)}_{\text{evidence}}} \right] \tag{4.33}$$

$$= \arg\max_{\theta} \left[p(\mathbf{x}_1, \dots, \mathbf{x}_N|\theta) p(\theta) \right] \tag{4.34}$$

$$= \arg\max_{\theta} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\theta) p(\theta) \right] \tag{4.35}$$

Đẳng thức (4.33) xảy ra theo quy tắc Bayes. Đẳng thức (4.34) xảy ra vì mẫu số của (4.33) không phụ thuộc vào tham số θ. Đẳng thức (4.35) xảy ra nếu chúng ta giả thiết về sự độc lập giữa các $\mathbf{x}_i$. Chú ý rằng giả thiết độc lập thường xuyên được sử dụng.

Như vậy, điểm khác biệt lớn nhất giữa hai bài toán tối ưu MLE và MAP là việc hàm mục tiêu của MAP có thêm $p(\theta)$, tức phân phối của θ. Phân phối này chính là những thông tin ta biết trước về θ và được gọi là *prior*. Ta kết luận rằng ***posterior* tỉ lệ thuận với tích của *likelihood* và *prior***.

Vậy chọn *prior* thế nào? chúng ta cùng làm quen với một khái niệm mới: *conjugate prior*.

4.3.2 Conjugate prior

Nếu phân phối xác suất posterior $p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *có cùng dạng* (*same family*) với phân phối xác suất $p(\theta)$, prior và posterior được gọi là *conjugate distributions*, và $p(\theta)$ được gọi là *conjugate prior* cho hàm likelihood $p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta)$. Nghiệm của bài toán MAP và MLE có cấu trúc giống nhau.

Một vài cặp các *conjugate distributions*[1]:

- Nếu likelihood function là một Gaussian (phân phối chuẩn), và prior cho vector kỳ vọng cũng là một Gaussian, thế thì phân phối posterior cũng là một Gaussian. Ta nói rằng Gaussian conjugate với chính nó (hay còn gọi là *self-conjugate*).
- Nếu likelihood function là một Gaussian và prior cho phương sai là một phân phối gamma[2], phân phối posterior cũng là một Gaussian. Ta nói rằng phân phối gamma là conjugate prior cho phương sai của Gassian. Chú ý rằng phương sai có thể được coi là một biến giúp đo *độ chính xác* của mô hình. Phương sai càng nhỏ thì độ chính xác càng cao.
- Phân phối Beta là conjuate của phân phối Bernoulli.
- Phân phối Dirichlet là conjugate của phân phối categorical.

4.3.3 Hyperparameters

Xét một ví dụ nhỏ với phân phối Bernoulli với hàm mật độ xác suất:

$$p(x|\lambda) = \lambda^x(1-\lambda)^{1-x} \tag{4.36}$$

và conjugate của nó, phân phối Beta, có hàm phân mật độ xác suất:

$$p(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1} \tag{4.37}$$

Bỏ qua thừa số hằng số chỉ mang mục đích chuẩn hoá cho tích phân của hàm mật độ xác suất bằng một, ta có thể nhận thấy rằng phần còn lại của phân phối Beta có cùng *họ* (*family*)

[1] Đọc thêm: *Conjugate prior–Wikipedia* (*https://goo.gl/E2SHbD*).
[2] *Gamma distribution–Wikipedia*, (*https://goo.gl/kdWd2R*.)

với phân phối Bernoulli. Cụ thể, nếu sử dụng phân phối Beta làm *prior* cho tham số λ, và bỏ qua phần thừa số hằng số, posterior sẽ có dạng

$$\begin{aligned} p(\lambda|x) &\propto p(x|\lambda)p(\lambda) \\ &\propto \lambda^{x+\alpha-1}(1-\lambda)^{1-x+\beta-1} \end{aligned} \tag{4.38}$$

trong đó, $\propto$ là ký hiệu của *tỉ lệ với*.

Nhận thấy rằng (4.38) *vẫn có dạng của một phân phối Bernoulli*. Chính vì vậy mà phân phối Beta được gọi là một *conjugate prior* cho phân phối Bernoulli.

Trong ví dụ này, tham số λ phụ thuộc vào hai tham số khác là α và β. Để tránh nhầm lẫn, hai tham số (α, β) được gọi là *siêu tham số* (*hyperparameters*).

Quay trở lại ví dụ về bài toán tung đồng xu N lần có n lần nhận được mặt *head* và $m = N-n$ lần nhận được mặt *tail*. Nếu sử dụng MLE, ta nhận được ước lượng $\lambda = n/M$. Nếu sử dụng MAP với prior là một Beta$[\alpha, \beta]$ thì kết quả sẽ thay đổi thế nào?

Bài toán tối ưu MAP:

$$\begin{aligned} \lambda &= \underset{\lambda}{\text{argmax}} \left[p(x_1, \ldots, x_N|\lambda)p(\lambda)\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\left(\prod_{i=1}^{N} \lambda^{x_i}(1-\lambda)^{1-x_i}\right)\lambda^{\alpha-1}(1-\lambda)^{\beta-1}\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\lambda^{\sum_{i=1}^{N} x_i+\alpha-1}(1-\lambda)^{N-\sum_{i=1}^{N} x_i+\beta-1}\right] \\ &= \underset{\lambda}{\text{argmax}} \left[\lambda^{n+\alpha-1}(1-\lambda)^{m+\beta-1}\right] \end{aligned} \tag{4.39}$$

Bài toán tối ưu (4.39) chính là bài toán tối ưu (4.38) với tham số thay đổi một chút. Tương tự như (4.38), nghiệm của (4.39) có thể được suy ra là

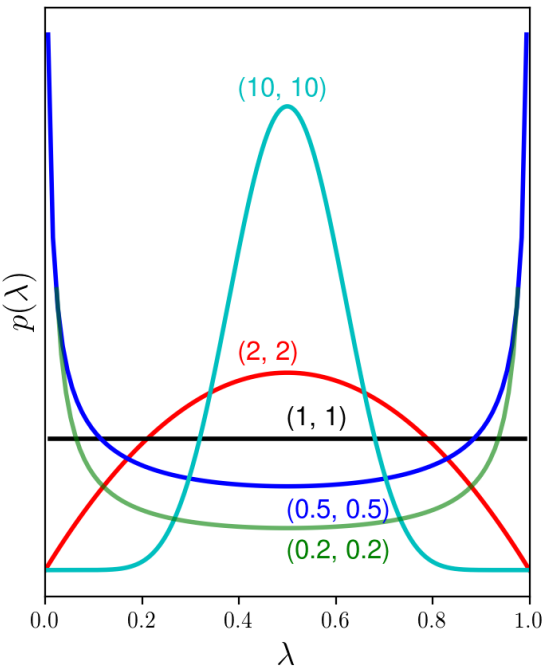
$$\lambda = \frac{n+\alpha-1}{N+\alpha+\beta-2} \tag{4.40}$$

Nhờ việc chọn prior phù hợp, ở đây là conjugate prior, posterior và likelihood có dạng giống nhau, khiến cho việc tối ưu bài toán MAP được thuận lợi.

Việc còn lại là chọn cặp *hyperparameters* α và β.

Chúng ta cùng xem lại hình dạng của phân phối Beta và nhận thấy rằng khi $\alpha = \beta > 1$, hàm mật độ xác suất của phân phối Beta đối xứng qua điểm 0.5 và đạt giá trị cao nhất tại 0.5. Xét Hình 4.1, ta nhận thấy rằng khi $\alpha = \beta > 1$, mật độ xác suất xung quanh điểm 0.5 nhận giá trị cao, điều này chứng tỏ λ có xu hướng gần với 0.5.

Nếu ta chọn $\alpha = \beta = 1$, ta nhận được phân phối đều vì đồ thị hàm mật độ xác suất là một đường thẳng. Lúc này, xác suất của λ tại mọi vị trí trong khoảng $[0, 1]$ là như nhau. Thực



Hình 4.1: Đồ thị hàm mật độ xác suất của phân phối Beta khi $\alpha = \beta$ và nhận các giá trị khác nhau. Khi cả hai giá trị này lớn, xác suất để λ gần 0.5 sẽ cao hơn.

chất, nếu ta thay $\alpha = \beta = 1$ vào (4.40) ta sẽ thu được $\lambda = n/N$, đây chính là ước lượng thu được bằng MLE. MLE là một trường hợp đặc biệt của MAP khi prior là một phân phối đều.

Nếu ta chọn $\alpha = \beta = 2$, ta sẽ thu được: $\lambda = \dfrac{n+1}{N+2}$. Chẳng hạn khi $N = 5, n = 1$ như trong ví dụ. MLE cho kết quả $\lambda = 1/5$, MAP sẽ cho kết quả $\lambda = 2/7$, gần với $1/2$ hơn.

Nếu chọn $\alpha = \beta = 10$ ta sẽ có $\lambda = (1+9)/(5+18) = 10/23$. Ta thấy rằng khi $\alpha = \beta$ và càng lớn thì ta sẽ thu được λ càng gần $1/2$. Điều này có thể dễ nhận thấy vì prior nhận giá trị rất cao tại 0.5 khi các siêu tham số $\alpha = \beta$ lớn.

4.3.4 MAP giúp tránh overfitting

Việc chọn các hyperparameter thường được dựa trên thực nghiệm, chẳng hạn bằng cross-validation. Việc thử nhiều bộ tham số rồi chọn ra bộ tốt nhất là việc mà các kỹ sư machine learning thường xuyên phải đối mặt. Cũng giống như việc chọn regularization parameter để tránh overfitting vậy.

Nếu viết lại bài toán MAP dưới dạng:

$$\theta = \underset{\theta}{\text{argmax}}\, p(\mathbf{X}|\theta)p(\theta) \tag{4.41}$$

$$= \underset{\lambda}{\text{argmax}} \left[\log \underbrace{p(\mathbf{X}|\theta)}_{\text{likelihood}} + \log \underbrace{p(\theta)}_{\text{prior}} \right] \tag{4.42}$$

ta có thể thấy rằng hàm mục tiêu có dạng $\mathcal{L}(\theta) + \lambda R(\theta)$ giống như trong regularization, với hàm log-likelihood đóng vai trò như hàm mất mát $\mathcal{L}(\theta)$, và log của prior đóng vai trò như hàm $R(\theta)$. Ta có thể nói rằng, MAP chính là một phương pháp giúp tránh overfitting trong các mô hình machine learning thống kê. MAP đặc biệt hữu ích khi tập huấn luyện là nhỏ.

4.4 Tóm tắt

- Khi sử dụng các mô hình thống kê machine learning, chúng ta thường xuyên phải ước lượng các tham số của mô hình θ, đại diện cho các tham số của các phân phối xác suất. Có hai phương pháp phổ biến được sử dụng để ước lượng θ là Maximum Likelihood Estimation (MLE) và Maximum A Posterior Estimation (MAP).

- Với MLE, việc xác định tham số θ được thực hiện bằng cách đi tìm các tham số sao cho xác suất của tập huấn luyện, hay còn gọi là *likelihood*, là lớn nhất:

$$\theta = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{x}_1, \ldots, \mathbf{x}_N|\theta) \tag{4.43}$$

- Để giải bài toán tối ưu này, giả thiết các dữ liệu $\mathbf{x}_i$ độc lập thường được sử dụng. Và bài toán MLP trở thành:

$$\theta = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}_i|\theta) \tag{4.44}$$

- Với MAP, các tham số được đánh giá bằng cách tối đa *posterior*:

$$\theta = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_N) \tag{4.45}$$

- Quy tắc Bayes và giả thiết về sự độc lập của dữ liệu thường được sử dụng:

$$\theta = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(\mathbf{x}_i|\theta)p(\theta)\right] \tag{4.46}$$

  Hàm mục tiêu ở đây chính là tích của *likelihood* và *prior*.

- *Prior* thường được chọn dựa trên các thông tin biết trước của tham số, và phân phối được chọn thường là các *conjugate distribution* với likelihood, tức các phân phối khiến việc nhân thêm *prior* vẫn giữ được cấu trúc giống như *likelihood*.

- MAP có thể được coi là một phương pháp giúp tránh overfitting. MAP thường mang lại hiệu quả cao hơn MLE với trường hợp có ít dữ liệu huấn luyện.

Tài liệu tham khảo

Index