Mục lục

Phần I

Lý thuyết

Chương 1

Học có giám sát

1.1 Cơ sở lý thuyết

1.1.1 Chuẩn của véctơ và ma trận

Trong học máy, các đối tượng thường được mô tả bởi véctơ hoặc ma trận thực. Vì vậy, chúng ta cần một thước đo để đánh giá độ lớn của từng đối tượng cũng như độ khác biệt (hay tương đồng) giữa các đối tượng.

Chuẩn của véctơ

Trong không gian véctơ thực $\mathbb{R}^n$, một véctơ $\boldsymbol{x}$, ký hiệu bằng chữ cái đậm, có các thành phần là số thực, ký hiệu bằng chữ cái thường. Véctơ luôn được viết dưới dạng cột

$$\boldsymbol{x} = (x_1, x_2, \dots, x_n)^\top \quad \text{hay} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1.1}$$

Xét chuẩn Minkowski, $\|\boldsymbol{x}\|_p$, với $p \geq 1$, đọc là *chuẩn ℓ_p của $\boldsymbol{x}$*, là số thực không âm xác định bởi.

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{1.2}$$

Một số chuẩn Minkowski hay dùng:

1) Với $p = 2$, ta có chuẩn Euclid, thường được dùng làm chuẩn mặc định. Đó là

căn bậc hai của tổng bình phương các tọa độ:

$$\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}. \tag{1.3}$$

2) Với $p = 1$, ta có chuẩn Manhattan, bằng tổng trị tuyệt đối các thành phần

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|. \tag{1.4}$$

3) Với $p = \infty$, chuẩn vô cùng của $\boldsymbol{x}$, được định nghĩa bởi

$$\|\boldsymbol{x}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{x}\|_p = \lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}. \tag{1.5}$$

Ta có thể chứng minh chuẩn vô cùng của $\boldsymbol{x}$ bằng giá trị lớn nhất của trị tuyệt đối các thành phần:

$$\|\boldsymbol{x}\|_\infty = \max_{1\le i\le n} |x_i| . \tag{1.6}$$

Chuẩn Minkowski có các tính chất chung của chuẩn, như sau:

1) Tính xác định dương:

$$\|\boldsymbol{x}\| \ge 0, \ \forall \boldsymbol{x} \in \mathbb{R}^n, \quad \text{và} \quad \|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}. \tag{1.7}$$

Lưu ý, khi viết $\|\boldsymbol{x}\| \ge 0$ hay $\|\boldsymbol{x}\| = 0$, ta hiểu 0 là số thực, nhưng khi viết $\boldsymbol{x} = \boldsymbol{0}$, ta phải hiểu $\boldsymbol{0}$ là véctơ không, $(0, 0, \dots, 0)$, gồm n số thực bằng 0. Như vậy cùng ký hiệu là 0 hoặc $\boldsymbol{0}$, nhưng ý nghĩa trong mỗi tình huống hoàn toàn khác nhau, người ta thường gọi là tính *tương thích* theo tình huống. Chúng ta cần nhớ điều này đặc biệt trong các bài trình bày bằng chữ viết tay, nơi người ta viết 0 và $\boldsymbol{0}$ là như nhau, và người đọc phải tự nhận diện đó là gì, cụ thể số thực không $0 \in \mathbb{R}$, hay là véctơ không $\boldsymbol{0} \in \mathbb{R}^n$, tức là đối tượng trong không gian nào. Cũng tương tự như vậy, ký hiệu đậm đặc trưng cho véctơ $\boldsymbol{x}$ cũng được viết là x trong trình bày viết tay, người đọc cũng cần nhận dạng nhanh kiểu dữ liệu nó.

2) Tính thuần nhất dương:

$$\|k\boldsymbol{x}\| = |k| \ \|\boldsymbol{x}\| \ \ \forall \boldsymbol{x} \in \mathbb{R}^n, k \in \mathbb{R}. \tag{1.8}$$

3) Bất đẳng thức tam giác:

$$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{1.9}$$

Ví dụ 1. Cho $\boldsymbol{x} = (6, -2, 3)^\top \in \mathbb{R}^3$. Tính các chuẩn Euclid, chuẩn Manhattan, và chuẩn vô cùng của $\boldsymbol{x}$.

Giải. Các chuẩn Euclid, chuẩn Manhattan, và chuẩn vô cùng của $\boldsymbol{x}$ lần lượt là

$$\begin{aligned} \|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2 &= \sqrt{6^2 + 2^2 + 3^2} = 7 \\ \|\boldsymbol{x}\|_1 &= 6 + 2 + 3 = 11 \\ \|\boldsymbol{x}\|_\infty &= \max\{6, 2, 3\} = 6. \end{aligned}$$

Mã lệnh Python để tính các chuẩn trên:

```
import numpy as np

x = np.array([6, -2, 3])
np.linalg.norm(x)           # hoặc np.linalg.norm(x, 2), → ||x||
np.linalg.norm(x, 1)        # ||x||_1
np.linalg.norm(x, np.inf)   # ||x||_∞
```

□

Chuẩn của ma trận

Cho ma trận thực $\boldsymbol{A}$ cỡ $m \times n$ gồm m hàng và n cột, trong đó phần tử ở hàng i cột j, $1 \leq i \leq m, 1 \leq j \leq n$, được ký hiệu là $a_{i,j}$, hoặc a_{ij} nếu không gây nhầm lẫn. Ta viết $\boldsymbol{A} = \left(a_{ij}\right)_{m\times n} \in \mathbb{R}^{m\times n}$, hoặc chi tiết

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & \dots\dots\dots & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}. \tag{1.10}$$

Với $p \geq 1$, chuẩn p của $\boldsymbol{A}$ được xác định bởi:

$$\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}. \tag{1.11}$$

trong đó $\boldsymbol{x}$ là véctơ có cỡ *tương thích*, tức là phép nhân $\boldsymbol{A}\boldsymbol{x}$ thực hiện được. Vì $\boldsymbol{A}$ có cỡ $m \times n$, nên $\boldsymbol{x}$ có cỡ $n \times 1$, tức là $\boldsymbol{x} \in \mathbb{R}^n$. Như vậy chuẩn $\|\boldsymbol{x}\|_p$ là chuẩn p trong $\mathbb{R}^n$, còn $\|\boldsymbol{A}\boldsymbol{x}\|_p$ là chuẩn p trong $\mathbb{R}^m$.

Chuẩn của ma trận được định nghĩa trên cơ sở chuẩn của véctơ, nên ta cũng có các chuẩn thông dụng tương ứng của ma trận tùy theo giá trị của p.

1) Với $p = \infty$, chuẩn vô cùng của $\boldsymbol{A}$ xác định bởi:

$$\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|, \tag{1.12}$$

tức là giá trị lớn nhất của các tổng trị tuyệt đối các phần tử trên mỗi hàng.

2) Với $p = 1$, chuẩn 1 của $\boldsymbol{A}$ có công thức khá tương tự (1.12), chỉ khác ở chỗ thay đổi thứ tự chỉ số hàng i và cột j trong tính toán:

$$\|\boldsymbol{A}\|_\infty = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \tag{1.13}$$

tức là giá trị lớn nhất của các tổng trị tuyệt đối các phần tử trên mỗi cột.

3) Với $p = 2$, chuẩn 2 của $\boldsymbol{A}$ được sử dụng nhiều, nhưng quy trình tính toán lại khá phức tạp. Đó là căn bậc hai của giá trị riêng lớn nhất của ma trận $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_2 = \sqrt{\max_{1\leq i\leq n} \lambda_i\left(\boldsymbol{A}^\top\boldsymbol{A}\right)}. \tag{1.14}$$

Ngoài ra, một chuẩn khá phổ biến của ma trận, được tính tương tự chuẩn Euclid cho véctơ, bằng cách xem ma trận như là một véctơ được "ngắt" thành nhiều đoạn và xếp chồng lên nhau. Chuẩn này được gọi là chuẩn Frobenius, gọi tắt là chuẩn F:

$$\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}. \tag{1.15}$$

Ví dụ 2. Tính các chuẩn vô cùng, chuẩn 1, chuẩn 2, và chuẩn Frobenius của

$$\boldsymbol{A} = \begin{bmatrix} 5 & 0 & 2 \\ 3 & -2 & -4 \end{bmatrix}.$$

Giải. 1) Chuẩn vô cùng của $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_\infty = \max\{5+0+2,\ 3+2+4\} = \max\{7,\ 9\} = 9.$$

2) Chuẩn 1 của $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_1 = \max\{5+3,\ 0+2,\ 2+4\} = \max\{8,\ 2,\ 6\} = 8.$$

3) Để tính chuẩn 2 của $\boldsymbol{A}$, ta tiến hành từng bước sau:

a) $\boldsymbol{A}^\top = \begin{bmatrix} 5 & 3 \\ 0 & -2 \\ 2 & -4 \end{bmatrix}$

b) $\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} 5 & 3 \\ 0 & -2 \\ 2 & -4 \end{bmatrix}\begin{bmatrix} 5 & 0 & 2 \\ 3 & -2 & -4 \end{bmatrix} = \begin{bmatrix} 34 & -6 & -2 \\ -6 & 4 & 8 \\ -2 & 8 & 20 \end{bmatrix}$

c) Đa thức đặc trưng của $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\begin{aligned} P(\lambda) = \left|\boldsymbol{A}^\top\boldsymbol{A} - \lambda\boldsymbol{I}\right| &= \begin{vmatrix} 34-\lambda & -6 & -2 \\ -6 & 4-\lambda & 8 \\ -2 & 8 & 20-\lambda \end{vmatrix} \\ &= (-\lambda)^3 + (34+4+20)(-\lambda)^2 \\ &\quad + \left(\begin{vmatrix} 34 & -6 \\ -6 & 4 \end{vmatrix} + \begin{vmatrix} 34 & -2 \\ -2 & 20 \end{vmatrix} + \begin{vmatrix} 4 & 8 \\ 8 & 20 \end{vmatrix}\right)(-\lambda) \\ &\quad + \begin{vmatrix} 34 & -6 & -2 \\ -6 & 4 & 8 \\ -2 & 8 & 20 \end{vmatrix} \\ &= -\lambda^3 + 58\lambda^2 - 792\lambda \end{aligned}$$

d) Các giá trị riêng của $\boldsymbol{A}^\top\boldsymbol{A}$ là nghiệm của đa thức đặc trưng $P(\lambda)$:

$$P(\lambda) = -\lambda(\lambda-36)(\lambda-22) = 0 \Leftrightarrow \begin{bmatrix} \lambda = 36 & = \lambda_1 \\ \lambda = 22 & = \lambda_2 \\ \lambda = 0 & = \lambda_3 \end{bmatrix}$$

e) Và cuối cùng, ta được chuẩn 2 của $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_2 = \sqrt{\max\{36,\ 22,\ 0\}} = \sqrt{36} = 6.$$

4) Chuẩn F của $\boldsymbol{A}$ bằng

$$\|\boldsymbol{A}\|_F = \sqrt{(5^2 + 0^2 + 2^2) + (3^2 + 2^2 + 4^2)} = \sqrt{58}.$$

Mã lệnh Python để tính chuẩn của ma trận cũng tương tự đối với chuẩn của véctơ:

```
import numpy as np

A = np.array([[5,  0,  2],
              [3, -2, -4]])

np.linalg.norm(A, np.inf)   # ||A||_∞
np.linalg.norm(A, 1)        # ||A||_1
np.linalg.norm(A, 2)        # ||A||_2
np.linalg.norm(A)           # hoặc np.linalg.norm(A, 'fro'), → ||A|| = ||A||_F
```

□

1.1.2 Bài toán bình phương tối thiểu

Cho ma trận $\boldsymbol{A}$ cỡ $m \times n$, và véctơ m chiều $\boldsymbol{b}$. Ta cần tìm véctơ n chiều $\boldsymbol{x}$ sao cho $\|\boldsymbol{Ax} - \boldsymbol{b}\|$ đạt giá trị nhỏ nhất. Bài toán được viết dưới dạng

$$\arg\min_{\boldsymbol{x}} \|\boldsymbol{Ax} - \boldsymbol{b}\|. \tag{1.16}$$

Phương trình chuẩn

Định lý 1.1. *Nếu $\boldsymbol{A}$ có hạng đầy đủ theo cột, tức là* $\text{rank}(\boldsymbol{A}) = n$*, thì nghiệm bình phương tối thiểu của bài toán (1.16) là*

$$\boldsymbol{x} = \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{b}. \tag{1.17}$$

Công thức (1.17) còn gọi là *phương trình chuẩn*.

1.1.3 Khai triển kỳ dị

Xét ma trận $\boldsymbol{A}$ cỡ $m \times n$ có hạng bằng r. Trường hợp phổ biến khi $m \geq n$, *khai triển kỳ dị* của $\boldsymbol{A}$ có dạng

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{1.18}$$

trong đó

- $\boldsymbol{\Sigma}$ là ma trận đường chéo cỡ $m \times n$, có các phần tử trên đường chéo chính là $\sigma_1, \sigma_2, ..., \sigma_r, 0, 0, ..., 0$:

$$\boldsymbol{\Sigma} = \text{diag}_{m\times n}(\sigma_1, \sigma_2, ..., \sigma_r, 0, ..., 0) = \left[\begin{array}{cccc|ccc} \sigma_1 & 0 & ... & 0 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 & 0 & ... & 0 \\ ... & ... & \ddots & ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_r & 0 & ... & 0 \\ \hline 0 & 0 & ... & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & ... & 0 \end{array}\right]_{m\times n} \tag{1.19}$$

  với các giá trị kỳ dị $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$ xác định bởi

$$\sigma_i = \sqrt{\lambda_i}, \quad i = 1, ..., n \tag{1.20}$$

  trong đó $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ là các giá trị riêng của $\boldsymbol{A}^\top\boldsymbol{A}$, kể cả các giá trị riêng bằng 0.

- Cột thứ i của $\boldsymbol{V}$, $\boldsymbol{v}_i = \boldsymbol{V}_{:,i}$, gọi là *các véctơ riêng phải*, là các véctơ riêng của $\boldsymbol{A}^\top\boldsymbol{A}$ ứng với mỗi giá trị riêng λ_i:

$$\boldsymbol{A}^\top A\boldsymbol{v}_i = \lambda_i\boldsymbol{v_i} \tag{1.21}$$

  $\boldsymbol{V}$ được thiết kế để thỏa mãn tính trực giao, tức là

$$\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I_n}, \quad \boldsymbol{I_n} \text{ là ma trận đơn vị cấp } n. \tag{1.22}$$

- Với $\sigma_i > 0$, dựng *các véctơ riêng trái*

$$\boldsymbol{u_i} = \frac{1}{\sigma_i}\boldsymbol{A}\boldsymbol{v_i}. \tag{1.23}$$

  tương ứng là cột thứ i của $\boldsymbol{U}$. Sau đó tiếp tục bổ sung hệ véctơ cơ sở trực chuẩn trong $\ker\left(\boldsymbol{A}^\top\right)$ vào $\boldsymbol{U}$ để đạt được ma trận trực giao cấp m:

$$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m. \tag{1.24}$$

Khai triển kỳ dị cho ta xác định ma trận *giả nghịch đảo Moore – Penrose* của $\boldsymbol{A}$:

$$\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{T}, \tag{1.25}$$

trong đó

$$\boldsymbol{\Sigma}^{+} = \text{diag}_{n\times m}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0\right) = \left[\begin{array}{cccc|ccc} \frac{1}{\sigma_1} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \ddots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_r} & 0 & \dots & 0 \\ \hline 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{array}\right]_{n\times m} \tag{1.26}$$

để từ đó tìm được nghiệm bình phương tối thiểu của bài toán (1.16).

Định lý 1.2. *Nghiệm bình phương tối thiểu của bài toán* (1.16) *là*

$$\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{b}. \tag{1.27}$$

Đặt

$$\boldsymbol{\Sigma}_r = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r) \quad \text{(ma trận vuông cấp } r\text{)} \tag{1.28}$$

$$\boldsymbol{\Sigma}_r^{+} = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \dots, \frac{1}{\sigma_r}\right) \tag{1.29}$$

$$\boldsymbol{U}_r = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_r] \quad (r \text{ cột đầu của } \boldsymbol{U},\ \text{cỡ } m \times r) \tag{1.30}$$

$$\boldsymbol{V}_r = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r] \quad (r \text{ cột đầu của } \boldsymbol{V},\ \text{cỡ } n \times r), \tag{1.31}$$

ta thu được khai triển kỳ dị rút gọn của $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^{\top}, \tag{1.32}$$

đồng thời

$$\boldsymbol{A}^{+} = \boldsymbol{V}_r\boldsymbol{\Sigma}_r^{+}\boldsymbol{U}_r^{\top}. \tag{1.33}$$

1.1.4 Bổ sung một số ký hiệu và phép toán ma trận

Đối với ma trận

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & \dots\dots\dots & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \tag{1.34}$$

hàng thứ i của $\boldsymbol{A}$ là véctơ hàng (đôi khi cũng gọi là ma trận hàng)

$$\boldsymbol{A}_{i,:} = (a_{i1}, a_{i2}, \dots, a_{in}) \tag{1.35}$$

và $\boldsymbol{A}$ được viết dưới dạng ma trận ô

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,:} \\ \boldsymbol{A}_{2,:} \\ \vdots \\ \boldsymbol{A}_{m,:} \end{bmatrix} \tag{1.36}$$

Tương tự, cột thứ j của $\boldsymbol{A}$ là véctơ cột (hay ma trận cột)

$$\boldsymbol{A}_{:,j} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix}, \tag{1.37}$$

và $\boldsymbol{A}$ có dạng ô

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{:,1}, \boldsymbol{A}_{:,2}, \dots, \boldsymbol{A}_{:,n} \end{bmatrix}. \tag{1.38}$$

Chuyển vị của ma trận ô được thực hiện tương tự ma trận thông thường, tức là nếu

$$\boldsymbol{A} = \left(\boldsymbol{A}_{ij}\right)_{i=\overline{1,m},\, j=\overline{1,n}} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \dots & \boldsymbol{A}_{1n} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \dots & \boldsymbol{A}_{2n} \\ & \dots\dots\dots & & \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \dots & \boldsymbol{A}_{mn} \end{bmatrix} \tag{1.39}$$

trong đó chỉ số chạy $i = \overline{1, m}$ được viết trước sẽ lập nên hàng thứ i, và chỉ số chạy sau $j = \overline{1, n}$ sẽ gọi ra ma trận ô thứ j trên hàng đó của ma trận, thì

$$\boldsymbol{A}^\top = \left(\boldsymbol{A}_{ij}^\top\right)_{j=\overline{1,n},\, i=\overline{1,m}} = \begin{bmatrix} \boldsymbol{A}_{11}^\top & \boldsymbol{A}_{21}^\top & \dots & \boldsymbol{A}_{m1}^\top \\ \boldsymbol{A}_{12}^\top & \boldsymbol{A}_{22}^\top & \dots & \boldsymbol{A}_{m2}^\top \\ & \dots\dots\dots & & \\ \boldsymbol{A}_{1n}^\top & \boldsymbol{A}_{2n}^\top & \dots & \boldsymbol{A}_{mn}^\top \end{bmatrix}. \tag{1.40}$$

Ngoài các phép toán trong đại số ma trận, trong học máy, chúng ta cũng hay gặp những phép toán được thực hiện một cách tự nhiên theo từng phần tử. Chẳng hạn, nếu a là một số, và $\boldsymbol{X} = \left(x_{ij}\right)_{m\times n}$, thì phép cộng một số với một ma trận được hiểu là cộng số đó vào từng phần tử của ma trận:

$$a + \boldsymbol{X} = a + \left(x_{ij}\right)_{m\times n} = \left(a + x_{ij}\right)_{m\times n}, \tag{1.41}$$

hay, một cách cụ thể

$$a + \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ & \dots\dots\dots & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} a + x_{11} & a + x_{12} & \dots & a + x_{1n} \\ a + x_{21} & a + x_{22} & \dots & a + x_{2n} \\ & \dots\dots\dots\dots\dots & & \\ a + x_{m1} & a + x_{m2} & \dots & a + x_{mn} \end{bmatrix}. \tag{1.42}$$

Cách viết đảo lại $\boldsymbol{X} + a$, và phép trừ, cũng được thực hiện tương tự.

Đối với phép nhân, chúng ta đã có phép nhân ma trận theo nghĩa đại số. Vì vậy, để nhân theo vị trí tương ứng giữa hai ma trận cùng cấp $\boldsymbol{A} = \left(a_{ij}\right)_{m\times n}$ và $\boldsymbol{B} = \left(b_{ij}\right)_{m\times n}$, ta ký hiệu

$$\boldsymbol{A} \odot \boldsymbol{B} = \left(a_{ij}b_{ij}\right)_{m\times n} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ & \dots\dots\dots\dots\dots & & \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix}. \tag{1.43}$$

1.1.5 Ma trận Jacobi và gradient

Cho hai hàm véctơ m thành phần với n biến $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$ xác định bởi: với $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^\top$, $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_m(\boldsymbol{x})]^\top$ trong đó $f_i(\boldsymbol{x}) = f_i(x_1, x_2, \dots, x_n)$ là các hàm số n biến. Ta quy ước ký hiệu $\boldsymbol{f}(\boldsymbol{x})$ bởi $\boldsymbol{f}$, và tương tự đối với các hàm khác. Ký hiệu $\boldsymbol{J}(\boldsymbol{f}) = \left(\dfrac{\partial f_i}{\partial x_j}\right)$ là *ma trận Jacobi* của $\boldsymbol{f}$ tại $\boldsymbol{x}$, $i = \overline{1, m}$ là chỉ số hàng, $j = \overline{1, n}$ là chỉ số cột:

$$\boldsymbol{J}(\boldsymbol{f}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \dots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \dots & \dfrac{\partial f_2}{\partial x_n} \\ & \dots\dots\dots & & \\ \dfrac{\partial f_m}{\partial x_1} & \dfrac{\partial f_m}{\partial x_2} & \dots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}. \tag{1.44}$$

Dễ dàng chỉ ra rằng $\boldsymbol{J}$ là một ánh xạ tuyến tính, bao gồm tính cộng tính và tính thuần nhất, tức là

$$\boldsymbol{J}(\boldsymbol{f}+\boldsymbol{g}) = \boldsymbol{J}(\boldsymbol{f}) + \boldsymbol{J}(\boldsymbol{g}) \quad \text{và} \quad \boldsymbol{J}(\alpha\boldsymbol{f}) = \alpha\boldsymbol{J}(\boldsymbol{f}) \text{ với } \alpha \in \mathbb{R}. \tag{1.45}$$

Cho ma trận $\boldsymbol{A}$ cỡ $m \times n$, véctơ $\boldsymbol{b}$ có chiều m, và véctơ $\boldsymbol{x}$ có n biến. Khi đó

$$\boldsymbol{J}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}.$$

Thật vậy, đặt $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, với $\boldsymbol{A} = \left(a_{ij}\right)_{m\times n}$. Khi đó

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{n} a_{ij}x_j = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \quad \Rightarrow \quad \frac{\partial f_i}{\partial x_j} = a_{ij},$$

và dẫn đến khẳng định trên. Ngoài ra, rõ ràng $\boldsymbol{J}(\boldsymbol{b}) = \boldsymbol{0}$, nên theo tính cộng tính

$$\boldsymbol{J}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}) = \boldsymbol{A}. \tag{1.46}$$

Cho hàm số n biến $f : \mathbb{R}^n \to \mathbb{R}$. *Gradient* của f tại $\boldsymbol{x}$, ký hiệu $\nabla f(\boldsymbol{x})$, hay đơn giản ∇f, là véctơ mà mỗi thành phần lần lượt là đạo hàm riêng của hàm số đó theo mỗi biến:

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)^\top = \boldsymbol{J}(f)^\top. \tag{1.47}$$

Quay lại với các hàm véctơ m thành phần n biến $\boldsymbol{f}$, $\boldsymbol{g}$. Khi đó $\boldsymbol{f}^\top\boldsymbol{g} = \sum_{j=1}^{m} f_j g_j = \boldsymbol{g}^\top\boldsymbol{f}$ là một hàm số. Ta có thể chứng minh

$$\nabla\left(\boldsymbol{f}^\top\boldsymbol{g}\right) = \boldsymbol{J}(\boldsymbol{f})^\top\boldsymbol{g} + \boldsymbol{J}(\boldsymbol{g})^\top\boldsymbol{f}. \tag{1.48}$$

Chứng minh. Thành phần thứ i trong véctơ vế trái là

$$\frac{\partial\left(\boldsymbol{f}^\top\boldsymbol{g}\right)}{\partial x_i} = \sum_{j=1}^{m}\left(\frac{\partial f_j}{\partial x_i}g_i + \frac{\partial g_j}{\partial x_i}f_i\right).$$

Tiếp theo, ta viết cụ thể hạng tử $\boldsymbol{J}(\boldsymbol{f})^\top \boldsymbol{g}$ ở vế phải

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_2} \\ & \dots\dots\dots & & \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}$$

và thấy rằng phần tử thứ i trong kết quả bằng

$$\frac{\partial f_1}{\partial x_i} g_1 + \frac{\partial f_2}{\partial x_i} g_2 + \cdots + \frac{\partial f_m}{\partial x_i} g_m = \sum_{j=1}^{m} \frac{\partial f_j}{\partial x_i} g_j.$$

Tương tự với hạng tử thứ hai ở vế phải, sau đó ta sẽ thấy phần tử thứ i ở hai vế bằng nhau và do đó có điều phải chứng minh.

Một cách khác để chứng minh, dùng phương pháp ma trận ô (có cỡ tương thích, từ số chiều của ma trận tổng thể, đến cả số chiều của các ma trận ô). Phương pháp này mặc dù được diễn giải trong nhiều bước, nhưng lại được ưa thích sử dụng trong nhiều trường hợp sau này. Ta có

$$\begin{aligned} \frac{\partial \left(\boldsymbol{f}^\top \boldsymbol{g}\right)}{\partial x_i} &= \left(\frac{\partial f_1}{\partial x_i}, \frac{\partial f_2}{\partial x_i}, \dots, \frac{\partial f_m}{\partial x_i}\right)^\top \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} + \left(\frac{\partial g_1}{\partial x_i}, \frac{\partial g_2}{\partial x_i}, \dots, \frac{\partial g_m}{\partial x_i}\right)^\top \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \\ &= \left(\boldsymbol{J}(\boldsymbol{f})_{:,i}\right)^\top \boldsymbol{g} + \left(\boldsymbol{J}(\boldsymbol{g})_{:,i}\right)^\top \boldsymbol{f}. \end{aligned}$$

Suy ra

$$\begin{aligned} \nabla \left(\boldsymbol{f}^\top \boldsymbol{g}\right) &= \begin{bmatrix} \frac{\partial \left(\boldsymbol{f}^\top \boldsymbol{g}\right)}{\partial x_1} \\ \frac{\partial \left(\boldsymbol{f}^\top \boldsymbol{g}\right)}{\partial x_2} \\ \vdots \\ \frac{\partial \left(\boldsymbol{f}^\top \boldsymbol{g}\right)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \left(\boldsymbol{J}(\boldsymbol{f})_{:,1}\right)^\top \boldsymbol{g} + \left(\boldsymbol{J}(\boldsymbol{g})_{:,1}\right)^\top \boldsymbol{f} \\ \left(\boldsymbol{J}(\boldsymbol{f})_{:,2}\right)^\top \boldsymbol{g} + \left(\boldsymbol{J}(\boldsymbol{g})_{:,2}\right)^\top \boldsymbol{f} \\ \vdots \\ \left(\boldsymbol{J}(\boldsymbol{f})_{:,n}\right)^\top \boldsymbol{g} + \left(\boldsymbol{J}(\boldsymbol{g})_{:,n}\right)^\top \boldsymbol{f} \end{bmatrix} \\ &= \begin{bmatrix} \left(\boldsymbol{J}(\boldsymbol{f})_{:,1}\right)^\top \\ \left(\boldsymbol{J}(\boldsymbol{f})_{:,2}\right)^\top \\ \vdots \\ \left(\boldsymbol{J}(\boldsymbol{f})_{:,n}\right)^\top \end{bmatrix} \boldsymbol{g} + \begin{bmatrix} \left(\boldsymbol{J}(\boldsymbol{g})_{:,1}\right)^\top \\ \left(\boldsymbol{J}(\boldsymbol{g})_{:,2}\right)^\top \\ \vdots \\ \left(\boldsymbol{J}(\boldsymbol{g})_{:,n}\right)^\top \end{bmatrix} \boldsymbol{f} \end{aligned}$$

$$= \left(\boldsymbol{J}(\boldsymbol{f})_{:,1}, \boldsymbol{J}(\boldsymbol{f})_{:,2}, \ldots, \boldsymbol{J}(\boldsymbol{f})_{:,n}\right)^\top \boldsymbol{g} + \left(\boldsymbol{J}(\boldsymbol{g})_{:,1}, \boldsymbol{J}(\boldsymbol{g})_{:,2}, \ldots, \boldsymbol{J}(\boldsymbol{g})_{:,n}\right)^\top \boldsymbol{f}$$
$$= \boldsymbol{J}(\boldsymbol{f})^\top \boldsymbol{g} + \boldsymbol{J}(\boldsymbol{g})^\top \boldsymbol{f}.$$

□

Một vài trường hợp đặc biệt hay dùng của các kết luận trên. Cho ma trận $\boldsymbol{A}$, $\boldsymbol{B}$ cỡ $m \times n$, véctơ $\boldsymbol{c}$, $\boldsymbol{d}$ có chiều m, và véctơ $\boldsymbol{x}$ có n biến. Khi đó

$$\nabla\left[(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d})\right] = \boldsymbol{J}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})^\top(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d}) + \boldsymbol{J}(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d})^\top(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$$
$$= \boldsymbol{A}^\top(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{d}) + \boldsymbol{C}^\top(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b})$$

và hệ quả

$$\nabla\left(\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2\right) = \nabla\left[(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})\right] = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})$$
$$= 2\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) \tag{1.49}$$

Một tính chất quan trọng nữa của ma trận Jacobi, đó là *tính chất chuỗi*. Cho hàm véctơ $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, và $\boldsymbol{g}: \mathbb{R}^m \to \mathbb{R}^p$. Hàm hợp thành $\boldsymbol{g} \circ \boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^p$ xác định bởi $(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{x}) = \boldsymbol{g}[\boldsymbol{f}(\boldsymbol{x})]$. Khi đó

$$\boldsymbol{J}(\boldsymbol{g} \circ \boldsymbol{f}) = \boldsymbol{J}(\boldsymbol{g})\,\boldsymbol{J}(\boldsymbol{f}). \tag{1.50}$$

1.2 Các mô hình tuyến tính

Cho hàm số $y = f(\boldsymbol{x})$, với $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top \in \mathbb{R}^d$ là tập các thuộc tính hay đặc trưng. Giả sử chúng ta biết giá trị của hàm số tại m điểm $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)}\right)^\top$, $i = 1, 2, \ldots, m$:

$$y_i = f\left(x^{(i)}\right), \quad i = 1, 2, \ldots, m. \tag{1.51}$$

Các thông tin trên thường được mô tả dưới dạng bảng dữ liệu

$\boldsymbol{x}^{(i)\top}$	x_1	x_2	...	x_d	y
$\boldsymbol{x}^{(1)\top}$	$x_1^{(1)}$	$x_2^{(1)}$	...	$x_d^{(1)}$	y_1
$\boldsymbol{x}^{(2)\top}$	$x_1^{(2)}$	$x_2^{(2)}$	...	$x_d^{(2)}$	y_2
...	...	...	...	...	...
$\boldsymbol{x}^{(m)\top}$	$x_1^{(m)}$	$x_2^{(m)}$	...	$x_d^{(m)}$	y_m

(1.52)

Ma trận đặc trưng

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \ldots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \ldots & x_d^{(2)} \\ & \ldots\ldots\ldots & & \\ x_1^{(m)} & x_2^{(m)} & \ldots & x_d^{(m)} \end{bmatrix} \tag{1.53}$$

có hàng thứ i chính là véctơ hàng $\boldsymbol{x}^{(i)\top}$, nên $\boldsymbol{X}$ có thể viết dưới dạng ma trận ô theo các dữ liệu

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{(1)\top} \\ \boldsymbol{x}^{(2)\top} \\ \vdots \\ \boldsymbol{x}^{(m)\top} \end{bmatrix} \tag{1.54}$$

Từ hai dạng trên của $\boldsymbol{X}$, ta có thể xác định được dạng ma trận ô của $\boldsymbol{X}^\top$. Cách thứ nhất, trực tiếp dựa trên dạng cụ thể

$$\boldsymbol{X}^\top = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ & \dots\dots\dots & & \\ x_d^{(1)} & x_d^{(2)} & \dots & x_d^{(m)} \end{bmatrix} = \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(m)}\right],$$

và cách thứ hai, dựa trên cách lấy chuyển vị của ma trận ô

$$\boldsymbol{X}^\top = \left[\left(\boldsymbol{x}^{(1)\top}\right)^\top, \left(\boldsymbol{x}^{(2)\top}\right)^\top, \dots, \left(\boldsymbol{x}^{(m)\top}\right)^\top\right] = \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(m)}\right],$$

ta đều được

$$\boldsymbol{X}^\top = \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(m)}\right]. \tag{1.55}$$

Tiếp theo, ta ký hiệu véctơ các nhãn dự báo

$$\boldsymbol{Y} = (y_1, y_2, \dots, y_m)^\top. \tag{1.56}$$

Tùy vào giá trị của nhãn dự báo, ta có các lớp bài toán tương ứng:

a) Nếu $y = f(\boldsymbol{x})$ nhận giá trị thực nói chung, ta có bài mô hình hồi quy.

b) Nếu $y = f(\boldsymbol{x}) \in \{0, 1\}$, ta có mô hình phân loại nhị phân, còn gọi là mô hình hồi quy logistic.

Ta cần tìm một hàm

$$\widehat{y} = \widehat{f}(\boldsymbol{x}) \tag{1.57}$$

để dự đoán $y = f(\boldsymbol{x})$.

Ký hiệu giá trị dự đoán tại điểm dữ liệu thứ i:

$$\widehat{y}_i = \widehat{f}\left(\boldsymbol{x}^{(i)}\right), \tag{1.58}$$

và véctơ chứa các giá trị dự đoán

$$\widehat{\boldsymbol{Y}} = \left(\widehat{y}_1, \widehat{y}_2, \dots, \widehat{y}_m\right)^\top. \tag{1.59}$$

1.2.1 Phương pháp bình phương tối thiểu

Trong bài toán bình phương tối thiểu, ta cần tìm một hàm $\widehat{y} = \widehat{f}(\boldsymbol{x})$ để dự đoán $y = f(\boldsymbol{x})$ sao cho sai số bình phương

$$\text{SE} = \sum_{i=1}^{m} \left[\widehat{f}\left(\boldsymbol{x}^{(i)}\right) - f\left(\boldsymbol{x}^{(i)}\right)\right]^2 = \sum_{i=1}^{m} \left(\widehat{y}_i - y_i\right)^2 = \left\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right\|^2 \tag{1.60}$$

đạt giá trị nhỏ nhất. Bài toán được ký hiệu bởi

$$\arg\min_{\widehat{f}} \text{SE}. \tag{1.61}$$

Ở đây, SE gọi là *hàm mục tiêu*, hàm $\widehat{f}$ làm cực tiểu hàm mục tiêu gọi là *nghiệm tối ưu* của bài toán.

Xét không gian véctơ V gồm các hàm xác định tại các điểm $\boldsymbol{x}^{(i)}$, $i = 1, \dots, m$. Trong V, xét hệ n hàm độc lập tuyến tính

$$\boldsymbol{f}_{\text{base}} = (f_1, f_2, \dots, f_n)^\top. \tag{1.62}$$

Ta sẽ tìm hàm $\widehat{f}$ sinh bởi hệ hàm $\boldsymbol{f}_{\text{base}}$, tức là

$$\widehat{f} = \sum_{j=1}^{n} w_j f_j = \boldsymbol{w}^\top \boldsymbol{f}_{\text{base}}, \quad \text{trong đó } \boldsymbol{w} = (w_1, w_2, \dots, w_n)^\top. \tag{1.63}$$

Tại mỗi điểm dữ liệu $\boldsymbol{x}^{(i)}$, $i = 1, \dots, m$, ta có

$$\widehat{y}_i = \widehat{f}\left(\boldsymbol{x}^{(i)}\right) = \sum_{j=1}^{n} w_j f_j\left(\boldsymbol{x}^{(i)}\right). \tag{1.64}$$

Ma trận thiết kế chứa giá trị tại mỗi điểm dữ liệu của từng hàm trong hệ hàm:

$$\boldsymbol{A} = \left(a_{ij}\right)_{m\times n} \quad \text{với} \quad a_{ij} = f_j\left(\boldsymbol{x}^{(i)}\right), \tag{1.65}$$

và ký hiệu hàng thứ i của $\boldsymbol{A}$ là $\boldsymbol{A}_{i,:}$. Khi đó

$$\widehat{y}_i = \boldsymbol{A}_{i,:}\boldsymbol{w} \tag{1.66}$$

Suy ra

$$\widehat{\boldsymbol{Y}} = \begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{1,:}\boldsymbol{w} \\ \vdots \\ \boldsymbol{A}_{m,:}\boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{1,:} \\ \vdots \\ \boldsymbol{A}_{m,:} \end{bmatrix} \boldsymbol{w} = \boldsymbol{A}\boldsymbol{w} \tag{1.67}$$

Bài toán bình phương tối thiểu (1.61) trở thành

$$\arg\min_{\boldsymbol{w}} \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|^2, \tag{1.68}$$

có nghiệm được cho bởi phương trình chuẩn (1.17), tức là

$$\boldsymbol{w} = \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{Y}, \tag{1.69}$$

nếu đáp ứng được điều kiện $\text{rank}(\boldsymbol{A}) = n$, hoặc nếu không, sử dụng (1.27) theo khai triển kỳ dị:

$$\boldsymbol{w} = \boldsymbol{A}^+ \boldsymbol{Y}. \tag{1.70}$$

Lưu ý: $\widehat{f}$ làm cực tiểu sai số bình phương $\text{SE} = \left\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right\|^2$ tương đương với làm cực tiểu sai số bình phương trung bình

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{y}_i - y_i\right)^2 = \frac{1}{m}\left\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right\|^2 = \frac{1}{m}\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|^2, \tag{1.71}$$

cũng như căn bậc hai của sai số bình phương trung bình

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\widehat{y}_i - y_i\right)^2} = \frac{1}{\sqrt{m}}\left\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\right\| = \frac{1}{\sqrt{m}}\|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|, \tag{1.72}$$

tức là các bài toán tối ưu $\arg\min_{\widehat{f}} \text{SE}$, $\arg\min_{\widehat{f}} \text{MSE}$, và $\arg\min_{\widehat{f}} \text{RMSE}$ có chung nghiệm $\widehat{f}$ (nếu tồn tại), chỉ khác ở chỗ giá trị của hàm mục tiêu. Ta viết

$$\arg\min_{\widehat{f}} \text{SE} = \arg\min_{\widehat{f}} \text{MSE} = \arg\min_{\widehat{f}} \text{RMSE}. \tag{1.73}$$

Mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính có hàm dự báo

$$\widehat{f}(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x}, \tag{1.74}$$

trong đó $b \in \mathbb{R}$ là hệ số chặn, véctơ $\boldsymbol{w} = (w_1, w_2, \ldots, w_d)^\top \in \mathbb{R}^d$ gồm các hệ số (hay trọng số) hồi quy. Ta viết lại hàm dự báo dưới dạng

$$\widehat{f}(\boldsymbol{x}) = b \cdot 1 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d, \tag{1.75}$$

dẫn đến việc xét hệ hàm và véctơ tham số:

$$\boldsymbol{f}_{\text{base}} = \begin{bmatrix} f_0(\boldsymbol{x}) = 1 \\ f_1(\boldsymbol{x}) = x_1 \\ \vdots \\ f_d(\boldsymbol{x}) = x_d \end{bmatrix}, \quad \widetilde{\boldsymbol{w}} = \begin{bmatrix} 1 \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} 1 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \tag{1.76}$$

để quy về bài toán bình phương tối thiểu

$$\widehat{f}(\boldsymbol{x}) = \widetilde{\boldsymbol{w}}^\top \boldsymbol{f}_{\text{base}}. \tag{1.77}$$

Trong trường hợp này, ma trận thiết kế $\boldsymbol{A}$ các giá trị tại các điểm dữ liệu của hệ hàm theo công thức (1.65) được xác định bởi

$$\boldsymbol{A} = \begin{bmatrix} f_0\left(\boldsymbol{x}^{(1)}\right) & f_1\left(\boldsymbol{x}^{(1)}\right) & \dots & f_d\left(\boldsymbol{x}^{(1)}\right) \\ f_0\left(\boldsymbol{x}^{(2)}\right) & f_1\left(\boldsymbol{x}^{(2)}\right) & \dots & f_d\left(\boldsymbol{x}^{(2)}\right) \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots} \\ f_0\left(\boldsymbol{x}^{(m)}\right) & f_1\left(\boldsymbol{x}^{(m)}\right) & \dots & f_d\left(\boldsymbol{x}^{(m)}\right) \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & \dots & x_d^{(2)} \\ & & \dots\dots\dots & \\ 1 & x_1^{(m)} & \dots & x_d^{(m)} \end{bmatrix} = \left[\mathbf{1}_{m\times 1}, \boldsymbol{X}\right], \tag{1.78}$$

trong đó $\mathbf{1}_{m\times 1}$ là ma trận cỡ $m \times 1$ gồm toàn số 1, còn gọi là *cột chặn*. Trong nhiều trường hợp, nếu không sợ gây nhầm lẫn, ta đơn giản chỉ cần viết $\mathbf{1}$.

Một ví dụ ở dạng đơn giản nhất, khi hàm dự báo chỉ phụ thuộc vào một thuộc tính duy nhất. Tuy nhiên, ví dụ sẽ giúp chúng ta trực quan hóa dữ liệu, góp phần giải thích tính đúng đắn của các phương pháp.

Ví dụ 3. Cho bảng dữ liệu của hàm số $f(x)$:

#	x	y
#1	1	0
#2	2	3
#3	3	2
#4	3	4
#5	5	5
#6	4	?

Giả sử dữ liệu tuân theo mô hình hồi quy tuyến tính, tức là có hàm dự báo $\widehat{f}(x) = b + wx$.

a) Thực thi mã lệnh với thư viện scikit–learn xác định các tham số b, w, và dự đoán $f(4)$. Từ đó

i) Xác định giá trị hàm mục tiêu MSE sau khi tối ưu mô hình.

ii) Trực quan hóa dữ liệu bao gồm: các điểm dữ liệu, điểm dự đoán, và đường hồi quy.

b) Xác định các tham số b, w bằng phương trình chuẩn. So sánh kết quả với ý (a).

c) Xác định các tham số b, w bằng phương pháp khai triển kỳ dị. So sánh kết quả với ý (a).

1.2.2 Hồi quy và phân loại Ridge

Hồi quy

$$\arg\min_{w} \|Aw - y\|^2 + \alpha\|w\|^2. \tag{1.79}$$

1.2.3 Lasso

$$\arg\min_{w} \frac{1}{2m}\|Aw - y\|^2 + \alpha\|w\|^2. \tag{1.80}$$

1.2.4 Elastic-Net

$$\arg\min_{w} \frac{1}{2m}\|Aw - y\|^2 + \alpha\left(\rho\|w\|_1 + \frac{1-\rho}{2}\|w\|^2\right). \tag{1.81}$$

1.2.5 Hồi quy logistic

Cho dữ liệu phân loại nhị phân

$\boldsymbol{x}^{(i)\top}$	x_1	x_2	y
$\boldsymbol{x}^{(1)\top}$	1	2	1
$\boldsymbol{x}^{(2)\top}$	2	3	1
$\boldsymbol{x}^{(3)\top}$	3	1	1
$\boldsymbol{x}^{(4)\top}$	2	5	0
$\boldsymbol{x}^{(5)\top}$	4	3	0
$\boldsymbol{x}^{\top}$	1	4	?

$$\widetilde{\boldsymbol{x}} = \begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}, \quad \widetilde{\boldsymbol{w}} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \tag{1.82}$$

$$\widetilde{\boldsymbol{X}} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_d^{(1)} \\ 1 & x_1^{(2)} & \dots & x_d^{(2)} \\ & \dots\dots\dots & & \\ 1 & x_1^{(m)} & \dots & x_d^{(m)} \end{bmatrix} = \left[\mathbf{1}_{m\times 1}, \boldsymbol{X} \right] \tag{1.83}$$

$$z = b + \boldsymbol{w}^\top \boldsymbol{x} = \widetilde{\boldsymbol{w}}^\top \widetilde{\boldsymbol{x}} \tag{1.84}$$

Để tránh phức tạp về ký hiệu, từ bây giờ, ta thay $\widetilde{\boldsymbol{x}}$ bởi $\boldsymbol{x}$, $\widetilde{\boldsymbol{w}}$ bởi $\boldsymbol{w}$, và $\widetilde{\boldsymbol{X}}$ bởi $\boldsymbol{X}$. Ta viết lại

$$z = \boldsymbol{w}^\top \boldsymbol{x} \tag{1.85}$$

Vì z là đại lượng vô hướng, nên

$$z = z^\top = \left(\boldsymbol{w}^\top \boldsymbol{x} \right) = \boldsymbol{x}^\top \left(\boldsymbol{w}^\top \right)^\top = \boldsymbol{x}^\top \boldsymbol{w}. \tag{1.86}$$

Lấy gradient của z theo $\boldsymbol{w}$, có thể dùng định nghĩa (bạn đọc tự thực hiện), hoặc thông qua ma trận Jacobi:

$$\nabla_{\boldsymbol{w}} z = \boldsymbol{J}_{\boldsymbol{w}}(z)^\top = \boldsymbol{J}_{\boldsymbol{w}} \left(\boldsymbol{x}^\top \boldsymbol{w} \right)^\top = \left(\boldsymbol{x}^\top \right)^\top = \boldsymbol{x}. \tag{1.87}$$

Xét hàm logistic sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{1.88}$$

Không khó để kiểm tra

$$p = \sigma(z) \quad \Rightarrow \quad \frac{dp}{dz} = \sigma(z)\left(1 - \sigma(z)\right) = p\,(1 - p) \tag{1.89}$$

Khi đó

$$\frac{d}{dz} \log p = \frac{1}{p} \frac{dp}{dz} = \frac{1}{p} p\,(1 - p) = 1 - p. \tag{1.90}$$

Thay p bởi $1 - p$ và áp dụng quy tắc đạo hàm của hàm hợp:

$$\frac{d}{dz} \log(1 - p) = [1 - (1 - p)]\, d(1 - p) = -p. \tag{1.91}$$

$$\ell = \ell(\boldsymbol{w}; \boldsymbol{x}, y) = -y \log p - (1 - y) \log(1 - p), \tag{1.92}$$

với $y \in \{0, 1\}$, $p = \sigma(z) = \sigma\left(\boldsymbol{w}^\top \boldsymbol{x} \right)$. Ta sẽ tính $\nabla_{\boldsymbol{w}} \ell$. Xét hai trường hợp

Trường hợp 1: $y = 1 \Rightarrow \ell = -\log p \Rightarrow \nabla_{\boldsymbol{w}} \ell = \dfrac{d\ell}{dz} \nabla_{\boldsymbol{w}} z = \dfrac{d(-\log p)}{dz} \nabla_{\boldsymbol{w}} z = -(1-p)\,\boldsymbol{x} = \left[\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}\right) - y\right] \boldsymbol{x}$.

Trường hợp 2: $y = 0 \Rightarrow \ell = -\log(1-p) \Rightarrow \nabla_{\boldsymbol{w}} \ell = \dfrac{d\ell}{dz} \nabla_{\boldsymbol{w}} z = -\dfrac{d\log(1-p)}{dz} \nabla_{\boldsymbol{w}} z = -(-p)\,\boldsymbol{x} = \left[\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}\right) - y\right] \boldsymbol{x}$.

Tóm lại

$$\nabla_{\boldsymbol{w}} \ell = \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}; \boldsymbol{x}, y) = \left[\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}\right) - y\right] \boldsymbol{x}. \tag{1.93}$$

Tại điểm dữ liệu $\left(\boldsymbol{x}^{(i)}, y_i\right)$

$$\nabla_{\boldsymbol{w}} \ell = \nabla_{\boldsymbol{w}} \ell\left(\boldsymbol{w}; \boldsymbol{x}^{(i)}, y_i\right) = \left[\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(i)}\right) - y_i\right] \boldsymbol{x}^{(i)}. \tag{1.94}$$

Hàm mất mát entropy chéo

$$\boxed{\begin{aligned} L = L(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{Y}) &= \frac{1}{m} \sum_{i=1}^{m} \ell\left(\boldsymbol{w}; \boldsymbol{x}^{(i)}, y_i\right) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left[-y_i \log p_i - (1-y_i) \log\left(1-p_i\right)\right]. \end{aligned}} \tag{1.95}$$

trong đó $p_i = \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(i)}\right)$.

Ta có

$$\nabla_{\boldsymbol{w}} L = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{w}} \ell\left(\boldsymbol{w}; \boldsymbol{x}^{(i)}, y_i\right) = \frac{1}{m} \sum_{i=1}^{m} \left[\sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(i)}\right) - y_i\right] \boldsymbol{x}^{(i)} \tag{1.96}$$

$$= \frac{1}{m} \left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\right] \begin{bmatrix} \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}\right) - y_1 \\ \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(2)}\right) - y_2 \\ \vdots \\ \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(m)}\right) - y_m \end{bmatrix} \tag{1.97}$$

Trong công thức (1.55), ta đã biết $\left[\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)}\right] = \boldsymbol{X}^\top$. Tiếp theo, vì $\boldsymbol{w}^\top \boldsymbol{x}^{(i)}$ là đại lượng vô hướng, nên $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} = \left(\boldsymbol{w}^\top \boldsymbol{x}^{(i)}\right)^\top = \boldsymbol{x}^{(i)\top} \left(\boldsymbol{w}^\top\right)^\top = \boldsymbol{x}^{(i)\top} \boldsymbol{w}$. Suy ra

$$\boldsymbol{X}\boldsymbol{w} = \begin{bmatrix} \boldsymbol{x}^{(1)\top} \\ \vdots \\ \boldsymbol{x}^{(m)\top} \end{bmatrix} \boldsymbol{w} = \begin{bmatrix} \boldsymbol{x}^{(1)\top} \boldsymbol{w} \\ \vdots \\ \boldsymbol{x}^{(m)\top} \boldsymbol{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^\top \boldsymbol{x}^{(1)} \\ \vdots \\ \boldsymbol{w}^\top \boldsymbol{x}^{(m)} \end{bmatrix} \tag{1.98}$$

Nếu quy ước hàm sigmoid có thể tác động lên từng phần tử của véctơ, thì

$$\sigma(\boldsymbol{X}\boldsymbol{w}) = \begin{bmatrix} \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}\right) \\ \vdots \\ \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(m)}\right) \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}\right) - y_1 \\ \vdots \\ \sigma\left(\boldsymbol{w}^\top \boldsymbol{x}^{(m)}\right) - y_m \end{bmatrix} = \sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}. \tag{1.99}$$

Công thức (1.97) trở thành

$$\boxed{\nabla_{\boldsymbol{w}} L(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{m} \boldsymbol{X}^\top \left[\sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}\right]} \tag{1.100}$$

Ma trận Hess của L theo $\boldsymbol{w}$:

$$\nabla^2_{\boldsymbol{w}} L = \begin{bmatrix} \dfrac{\partial^2 L}{\partial w_1^2} & \dfrac{\partial^2 L}{\partial w_1 \partial w_2} & \cdots & \dfrac{\partial^2 L}{\partial w_1 \partial w_d} \\ \dfrac{\partial^2 L}{\partial w_2 \partial w_1} & \dfrac{\partial^2 L}{\partial w_2^2} & \cdots & \dfrac{\partial^2 L}{\partial w_2 \partial w_d} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 L}{\partial w_d \partial w_1} & \dfrac{\partial^2 L}{\partial w_d \partial w_2} & \cdots & \dfrac{\partial^2 L}{\partial w_d^2} \end{bmatrix} \tag{1.101}$$

$$= \boldsymbol{J}(\nabla_{\boldsymbol{w}} L) = \frac{1}{m} \boldsymbol{X}^\top \boldsymbol{J}(\sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}) \tag{1.102}$$

$$= \left[\nabla_{\boldsymbol{w}} \frac{\partial L}{\partial w_1}, \ \nabla_{\boldsymbol{w}} \frac{\partial L}{\partial w_2}, \ldots, \nabla_{\boldsymbol{w}} \frac{\partial L}{\partial w_d}\right] \tag{1.103}$$

Từ (1.100), ta có

$$\frac{\partial L}{\partial w_j} = \frac{1}{m} \left(\boldsymbol{X}^\top\right)_{j,:} \left[\sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}\right], \tag{1.104}$$

trong đó $\left(\boldsymbol{X}^\top\right)_{j,:}$ là hàng thứ j của $\boldsymbol{X}^\top$, cũng chính là chuyển vị của cột thứ j của $\boldsymbol{X}$, nên

$$\frac{\partial L}{\partial w_j} = \frac{1}{m} \left(\boldsymbol{X}_{:,j}\right)^\top \left[\sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}\right]. \tag{1.105}$$

Như vậy

$$\nabla^2_{\boldsymbol{w}} L(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{m} \boldsymbol{X}^\top \operatorname{diag}[\sigma(\boldsymbol{X}\boldsymbol{w}) \odot (1 - \sigma(\boldsymbol{X}\boldsymbol{w}))] \boldsymbol{X}. \tag{1.106}$$

Phương pháp Newton – Raphson / IRLS (Iteratively Reweighted Least Squares)

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \left(\nabla^2_{\boldsymbol{w}} L\right)^{-1} \nabla_{\boldsymbol{w}} L = \tag{1.107}$$

$$= \boldsymbol{w} - \left\{\boldsymbol{X}^\top \operatorname{diag}[\sigma(\boldsymbol{X}\boldsymbol{w}) \odot (1 - \sigma(\boldsymbol{X}\boldsymbol{w}))] \boldsymbol{X}\right\}^{-1} \boldsymbol{X}^\top \left(\sigma(\boldsymbol{X}\boldsymbol{w}) - \boldsymbol{Y}\right). \tag{1.108}$$

1.2.6 Phân loại nhị phân

Đọc trực tiếp trên Internet tập dữ liệu chiều cao, cân nặng. Từ đó đưa ra dự báo giới tính cho người có chiều cao 1.68 m và nặng 66 kg.

1.2.7 Phân loại đa lớp

```
import numpy as np
from sklearn.linear_model import
   LogisticRegression

# 1. Dữ liệu: 10 mẫu, 2 đặc trưng, 3 lớp
X = np.array([
    [1.0, 2.0],    # lớp A
    [1.2, 1.8],
    [0.8, 2.2],
    [3.0, 3.5],    # lớp B
    [3.2, 3.0],
    [2.8, 3.2],
    [5.0, 1.0],    # lớp C
    [5.2, 1.3],
    [4.8, 0.8],
    [5.1, 1.1]
])

Y = np.array([
    0, 0, 0,    # 3 mẫu lớp A
    1, 1, 1,    # 3 mẫu lớp B
    2, 2, 2, 2 # 4 mẫu lớp C
])



# 2. Khởi tạo mô hình
model = LogisticRegression(
    solver="lbfgs",
    multi_class="multinomial",
    max_iter=500
)
```

```
# 3. Huấn luyện mô hình
model.fit(X, y)

# 4. Dự đoán
X_pred = np.array([
    [2.0, 2.5],
    [4.9, 1.2]
])
Y_pred = model.predict(X_pred)
```

1.3 Phương pháp véctơ hỗ trợ (SVM)

Sinh viên tập vận hành các mô hình sau (số liệu tự lấy, cỡ mẫu ≥ 20):

1. SVC
2. NuSVC
3. SVR
4. NuSVR
5. LinearSVC
6. LinearSVR
7. OneClassSVM

Tài liệu tham khảo

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., và Duchesnay, É. *scikit-learn Machine Learning in Python.* phiên bản 1.7.1 (2025). https://scikit-learn.org/.
2. Bonaccorso, G. *Machine Learning Algorithms.* In lần thứ 2. 514 trang (Packt, 2018).