



统计方法与机器学习

第〇章：基础回顾

倪 蓓

DaSE@ECNU
(lni@dase.ecnu.edu.cn)



目录

① 向量与矩阵

向量

矩阵

求导法则

微商

② 概率论

随机向量

概率不等式

③ 优化理论

无约束优化问题

迭代算法

拉格朗日对偶

向量

在实数域 R^n 上的一个 n 维向量

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = (a_1, a_2, \cdots, a_n)'.$$

常用运算

- **加法** 设另有一个 n 维向量 $\mathbf{b} = (b_1, b_2, \cdots, b_n)'$, 则

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1, a_2 + b_2, \cdots, a_n + b_n)'$$

- **数乘** 设 c 为一个实数, 则

$$c\mathbf{a} = (ca_1, ca_2, \cdots, ca_n)'$$

向量

常用运算性质

- 若 $\mathbf{a}, \mathbf{b}, \mathbf{c} \in R^n$, 则

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}, \quad (\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$$

$$\mathbf{a} + \mathbf{0} = \mathbf{a}, \quad \mathbf{a} + (-\mathbf{a}) = \mathbf{0},$$

其中 $\mathbf{0} = (0, 0, \dots, 0)'$.

- 若 $\mathbf{a}, \mathbf{b} \in R^n, c, c_1, c_2 \in R$, 则

$$c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}, \quad (c_1 + c_2)\mathbf{a} = c_1\mathbf{a} + c_2\mathbf{a}$$

$$c_1(c_2\mathbf{a}) = c_1c_2\mathbf{a}, \quad 1 \cdot \mathbf{a} = \mathbf{a}.$$

向量

线性相关

- 对于向量 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$, 若存在一组数 c_1, c_2, \dots, c_k 使得

$$c_1 \mathbf{a}_1 + c_2 \mathbf{a}_2 + \dots + c_k \mathbf{a}_k = \mathbf{0},$$

则称向量 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ 是线性相关的; 否则, 称 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ 为线性无关。

- 给定向量 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$, 考虑由这些向量所有可能的线性组合 $\sum_{i=1}^k c_i \mathbf{a}_i$ 组成的集合

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k) = \left\{ \sum_{i=1}^k c_i \mathbf{a}_i : c_1, c_2, \dots, c_k \in R \right\}.$$

称其是由向量 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ 生成子空间。

向量

任意两个向量 $\mathbf{a} = (a_1, a_2, \dots, a_p)'$ 和 $\mathbf{b} = (b_1, b_2, \dots, b_p)'$ 。
 \mathbf{a}, \mathbf{b} 的内积为

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b} = \sum_{i=1}^p a_i b_i.$$

内积的性质

- $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \langle \mathbf{b}, \mathbf{a} \rangle$;
- $\langle \mathbf{a}, \mathbf{a} \rangle = \mathbf{a}'\mathbf{a} \geq 0$, 当且仅当 $\mathbf{a} = \mathbf{0}$; 记 $\langle \mathbf{a}, \mathbf{a} \rangle = \|\mathbf{a}\|^2$;
- $\langle c\mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{a}, c\mathbf{b} \rangle = c\langle \mathbf{a}, \mathbf{b} \rangle$ 对一切 $c \in R$ 成立;
- 对于三个向量 $\mathbf{a}, \mathbf{b}, \mathbf{c}$, 有

$$\langle \mathbf{a}, \mathbf{b} + \mathbf{c} \rangle = \langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{a}, \mathbf{c} \rangle.$$

向量

内积的不等式

- 柯西不等式: $\langle \mathbf{a}, \mathbf{b} \rangle^2 \leq \|\mathbf{a}\|^2 \cdot \|\mathbf{b}\|^2$;
- 三角不等式: $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$.

向量

基

- 设 \mathcal{L} 是 R^n 中的一个子空间, 如果存在 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ 使 $\mathcal{L} = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k)$ 且 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ 线性无关, 则称 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ 是 \mathcal{L} 的一组基。
- 在一个 R^n 的子空间 \mathcal{L} 的基 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ 具有性质

$$\begin{aligned}\langle \mathbf{a}_i, \mathbf{a}_i \rangle &= 1, i = 1, 2, \cdots, k, \\ \langle \mathbf{a}_i, \mathbf{a}_j \rangle &= 0, i \neq j.\end{aligned}$$

则称 $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ 是 \mathcal{L} 的一组标准正交基。

向量

在 R^n 中, 给定一个向量 a 及子空间 \mathcal{L} , 如果在 \mathcal{L} 中存在 b 使

$$\|a - b\| = \inf_{A \in \mathcal{L}} \|a - A\|$$

则称 b 是 a 在 \mathcal{L} 中的投影。

投影的性质

- 投影是存在且唯一的。

引理

在 R^n 中, 给定一个向量 a 及子空间 \mathcal{L} , b 是 a 在 \mathcal{L} 中的投影, 当且仅当

$$\langle a - b, A \rangle = 0, \quad \text{对一切 } A \in \mathcal{L} \text{ 成立。}$$

向量

格兰姆-施密特正交化方法

- 目的：任意一组线性无关的向量 $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$ ，可以找到子空间 $\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)$ 中的一组标准正交基。
- 具体步骤：
 - 令 $\mathbf{b}_1 = \mathbf{a}_1$;
 - 令 $\mathbf{b}_i = \mathbf{a}_i - h_{ii-1}\mathbf{b}_{i-1} - \dots - h_{i1}\mathbf{b}_1$ 使得

$$\langle \mathbf{b}_i, \mathbf{b}_{i-1} \rangle = \langle \mathbf{b}_i, \mathbf{b}_{i-2} \rangle = \dots = \langle \mathbf{b}_i, \mathbf{b}_1 \rangle = 0,$$

确定系数 h_{ii-1}, \dots, h_{i1}

- 以此类推，确定一组两两正交的向量 $\mathbf{b}_1, \dots, \mathbf{b}_k$;
- 令 $\mathbf{e}_i = \|\mathbf{b}_i\|^{-1}\mathbf{b}_i, i = 1, 2, \dots, k$, 则 $\mathbf{e}_1, \dots, \mathbf{b}_k$ 就是标准正交的向量组。

矩阵

称

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

为大小 $m \times n$ 的矩阵 $\mathbf{A} = \{a_{ij}\}_{m \times n}$ 。

常用符号

- \mathbf{I} : 单位矩阵, 主对角线的元素为一, 其他元素为零;
- $\mathbf{A}' = \{a_{ji}\}_{n \times m}$: 转置;
- $\text{rank}(\mathbf{X})$: 秩;
 - $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}')$;
 - $\text{rank}(\mathbf{AB}) \leq \min \{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$;

矩阵

当矩阵 A 的行数等于列数时，也称 A 为方阵。

常用符号

- A^{-1} : 逆矩阵;
- A^* : 伴随矩阵;
- $|A|$: 行列式;
 - $|A| = |A'|$;
 - 当 $|A| \neq 0$ 时, $|A|A^{-1} = A^*$;
 - $|AB| = |A||B|$;
 - $|A| \neq 0 \Leftrightarrow A_{n \times n}$ 非奇异 $\Leftrightarrow \text{rank}(A) = n$;
 - $|A|^{-1} = |A^{-1}|$ 。

矩阵

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

分块矩阵的逆

- 若 \mathbf{A}^{-1} 存在时, 且 $|\mathbf{A}_{11}| \neq 0$, 则

$$\begin{aligned} & \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} \\ = & \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{M}^{-1} \mathbf{A}_{21} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{M}^{-1} \\ -\mathbf{M}^{-1} \mathbf{A}_{21} \mathbf{A}_{11}^{-1} & \mathbf{M}^{-1} \end{pmatrix} \end{aligned}$$

其中 $\mathbf{M} = \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}$ 。

矩阵

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

分块矩阵的行列式

- 若 $|\mathbf{A}_{11}| \neq 0$, 则

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| |\mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}|.$$

- 若 $|\mathbf{A}_{22}| \neq 0$, 则

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}| |\mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}|.$$

矩阵

特征根和特征向量

给定一个 $n \times n$ 的方阵 A ,

- $\lambda I - A$ 是 λ 的 n 次多项式, 称为 A 的特征多项式。
- $|\lambda I - A| = 0$ 称为 A 的特征方程。
- 特征方程的解 (或特征多项式的根), 称为 A 的特征根。
- 因为 $|\lambda I - A| = 0$, 所以方程

$$(\lambda I - A)x = 0$$

一定有非零解, λ 所对应的非 0 解向量称为 λ 相应的特征向量, 也称为 A 的特征向量。

矩阵

迹

给定一个 $n \times n$ 的方阵 A , A 的迹 $\text{tr}(A)$ 是 A 的全年特征根之和。

- $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$;
- $\text{tr}(cA) = c\text{tr}(A)$;
- 当 AB 和 BA 均为方阵（但大小不要求相同）时，有

$$\text{tr}(AB) = \text{tr}(BA).$$

- 特例： $\text{tr}(A'A) = \text{tr}(AA')$ 。
- 特例： $\text{tr}(ab') = \text{tr}(b'a) = b'a$ 。

矩阵

特征根与行列式的关系

设 A 的全部的特征根为 $\lambda_1, \lambda_2, \dots, \lambda_n$, 有

$$|A| = \prod_{i=1}^n \lambda_i.$$

- A 非奇异, 当且仅当 A 的特征根均不为零;
- A 奇异, 当且仅当 A 至少有一个特征根为零;

矩阵

幂等阵

如果方阵 A 具有性质 $A^2 = A$, 则称 A 是幂等阵。

- 如果 $A^2 = A$, 则 A 的特征根非零即一。
- 如果 $A^2 = A$, 则 $\text{rank}(A) = \text{tr}(A)$ 。
- $R^n = \mathcal{L}(A') \oplus \mathcal{L}(I - A) = \mathcal{L}(A) \oplus \mathcal{L}(I - A')$
 $\Leftrightarrow A^2 = A$ 。

矩阵

对称阵

如果方阵 A 具有性质 $A' = A$ ，则称 A 是对称阵。

- 对每个对称阵 A ，任给一个向量 x ， $x'Ax$ 是 A 的一个齐次二次函数，称为 A 对应的二次型。
- 如果 A 的二次型 $x'Ax$ 恒不取负值，即 $x'Ax \geq 0$ 对一切 x 成立，则称 A 是非负定阵。
- 如果 A 是非负定阵，且 $x'Ax = 0$ 充要条件是 $x = 0$ ，则称 A 是正定的。

矩阵

对称阵的谱分解

- 对于对称阵 A , 设特征根分别为 $\lambda_1, \lambda_2, \dots, \lambda_n$, 可证明 λ_i 是实数, $i = 1, 2, \dots, n$ 。记

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

- 设 A 的 n 个单位正交特征向量为 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, 记

$$V = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$$

可证明 $V'V = VV' = I_n$ 。于是, $AV = V\Lambda$ 。

矩阵

对称阵的性质

设 A 是一个对称阵, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ 是 A 的特征根, 而 v_i 是 λ_i 对应的特征向量。于是, 有以下结论:

- 对于任意非零向量 x ,

$$\lambda_n x'x \leq x'Ax \leq \lambda_1 x'x.$$

- 推论: 对于任意非零向量 x ,

$$\sup_x \frac{x'Ax}{x'x} = \lambda_1, \quad \inf_x \frac{x'Ax}{x'x} = \lambda_n,$$

矩阵

证明：仅考虑 $\mathbf{x}'\mathbf{A}\mathbf{x} \leq \lambda_1 \mathbf{x}'\mathbf{x}$

不妨设特征向量 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ 是对称阵 \mathbf{A} 的标准正交基。
因此，

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \mathbf{v}_j' \mathbf{v}_i = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}.$$

对于任意非零向量 \mathbf{x} ， \mathbf{x} 可以写成 $\mathbf{x} = \sum_{i=1}^n k_i \mathbf{v}_i$ 。

矩阵

证明：仅考虑 $\mathbf{x}'\mathbf{A}\mathbf{x} \leq \lambda_1 \mathbf{x}'\mathbf{x}$ (续)

于是有

$$\begin{aligned}\mathbf{x}'\mathbf{A}\mathbf{x} &= \left(\sum_{j=1}^n k_j \mathbf{v}_j \right)' \mathbf{A} \left(\sum_{i=1}^n k_i \mathbf{v}_i \right) \\&= \left(\sum_{j=1}^n k_j \mathbf{v}_j \right)' \left(\sum_{i=1}^n \lambda_i k_i \mathbf{v}_i \right) \\&= \sum_{i,j} \lambda_i k_i k_j \mathbf{v}_j' \mathbf{v}_i \\&= \sum_i \lambda_i k_i^2 \\&\leq \lambda_1 \sum_i k_i^2 = \lambda_1 \mathbf{x}'\mathbf{x}.\end{aligned}$$

求导法则

- 列向量对标量求导：若

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

x 是一个标量，则其微商也是一个向量，即

$$\frac{\partial \mathbf{y}}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_n}{\partial x} \end{pmatrix}$$

求导法则

- 矩阵对标量求导：若

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix},$$

x 是一个标量，则其微商也是一个 $m \times n$ 矩阵，即

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{pmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \frac{\partial y_{21}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{2n}}{\partial x} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \frac{\partial y_{m2}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{pmatrix}$$

求导法则

- 标量对向量求导：若

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

y 是一个标量，则其微商也是一个向量，即

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{pmatrix}$$

求导法则

- 标量对矩阵求导：若

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix},$$

y 是一个标量，则其微商也是一个矩阵，即

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2n}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \frac{\partial y}{\partial x_{m2}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{pmatrix}.$$

求导法则

- **列向量对行向量求导**：若 $\mathbf{y} = (y_1, y_2, \dots, y_m)'$ 是一个 $m \times 1$ 列向量，而 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 是一个 $n \times 1$ 列向量，则

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

是一个 $m \times n$ 矩阵。

求导法则

- **行向量对列向量求导**：若 $\mathbf{y} = (y_1, y_2, \dots, y_m)'$ 是一个 $m \times 1$ 列向量，而 $\mathbf{x} = (x_1, x_2, \dots, x_n)$ 是一个 $n \times 1$ 列向量，则

$$\frac{\partial \mathbf{y}'}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

是一个 $n \times m$ 矩阵。

求导法则

- **列向量对矩阵求导**: 若 $\mathbf{y} = (y_1, y_2, \dots, y_m)'$ 是一个 $m \times 1$ 列向量, 而 \mathbf{X} 是一个 $p \times q$ 列向量, 则

$$\frac{\partial \mathbf{y}}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial y_1}{\partial \mathbf{X}} \\ \frac{\partial y_2}{\partial \mathbf{X}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{X}} \end{pmatrix}.$$

向量微商

常用符号及含义

设 x 是一个 $n \times 1$ 向量变量。

- 关于标量的微商：对于任意一个 $n \times 1$ 的非零常数向量 a ,

$$\frac{\partial}{\partial x}(a'x) = \frac{\partial}{\partial x}(x'a) = a.$$

- 关于二次型的微商：对于任意一个 $n \times n$ 的非零常数矩阵 A ,

$$\frac{\partial}{\partial x}(x'Ax) = (A + A')x.$$

特别地，若 $A = A'$ ，则 $\frac{\partial}{\partial x}(x'Ax) = 2Ax$.

矩阵微商

常用符号及含义

设 \mathbf{X} 是一个 $n \times n$ 矩阵变量。

- 关于标量的微商：对于任意两个 $n \times 1$ 的非零常数矩阵 \mathbf{a}, \mathbf{b} ,

$$\frac{\partial}{\partial \mathbf{X}}(\mathbf{a}'\mathbf{X}\mathbf{b}) = \mathbf{a}\mathbf{b}'.$$

特别地，若 $\mathbf{a} = \mathbf{b}$ ，则 $\frac{\partial}{\partial \mathbf{X}}(\mathbf{a}'\mathbf{X}\mathbf{a}) = \mathbf{a}\mathbf{a}'$.

- 关于行列式的微商：

$$\frac{\partial}{\partial \mathbf{X}}|\mathbf{X}| = |\mathbf{X}|(\mathbf{X}^{-1})' = |\mathbf{X}'|(\mathbf{X}')^{-1}.$$

随机向量

随机变量 vs 随机向量

- p 维随机向量 (random vector)

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix},$$

其中, 第 i 个分量 x_i 均是一个随机变量 (random variable);

- 当 $p = 1$ 时, $\mathbf{x} = x$ 是一个**标量**的随机变量 (scalar random variable)。

随机向量

高斯分布随机向量

- 如果随机向量 $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, 那么 \boldsymbol{x} 的 p.d.f. 为

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\},$$

- 该分布为多维高斯分布;
- $\boldsymbol{\mu}$ 是一个 p 维向量, 可证明 $\boldsymbol{\mu} = E(\boldsymbol{x})$;
- Σ 是一个 $p \times p$ 正定实对称矩阵, 可证明 $\Sigma = \text{Cov}(\boldsymbol{x})$ 。

随机向量

证明: $\boldsymbol{\mu} = E(\boldsymbol{x})$

$$\begin{aligned} E(\boldsymbol{x}) &= \int_{R^p} \boldsymbol{x} p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) d\boldsymbol{x} \\ &= \int_{R^p} \boldsymbol{x} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\} d\boldsymbol{x} \\ &= \int_{R^p} (\boldsymbol{y} + \boldsymbol{\mu}) (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \boldsymbol{y}' \Sigma^{-1} \boldsymbol{y} \right\} d\boldsymbol{y} \\ &= \int_{R^p} \boldsymbol{y} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \boldsymbol{y}' \Sigma^{-1} \boldsymbol{y} \right\} d\boldsymbol{y} \\ &\quad + \boldsymbol{\mu} \int_{R^p} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \boldsymbol{y}' \Sigma^{-1} \boldsymbol{y} \right\} d\boldsymbol{y} \\ &= \boldsymbol{\mu} \end{aligned}$$

随机向量

证明: $\Sigma = \text{Cov}(\mathbf{x})$

因为 $\text{Cov}(\mathbf{x}) = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'$, 所以,

$$\begin{aligned}\text{Cov}(\mathbf{x}) &= \int_{R^p} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) d\mathbf{x} \\&= \int_{R^p} (\mathbf{y} + \boldsymbol{\mu})(\mathbf{y} + \boldsymbol{\mu})' (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}' \Sigma^{-1} \mathbf{y} \right\} d\mathbf{y} \\&= \int_{R^p} \mathbf{y} \mathbf{y}' (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}' \Sigma^{-1} \mathbf{y} \right\} d\mathbf{y} \\&= \int_{R^p} \Sigma^{1/2} \mathbf{z} \mathbf{z}' \Sigma^{1/2} (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} \\&= \Sigma^{1/2} \cdot \int_{R^p} \mathbf{z} \mathbf{z}' (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} \cdot \Sigma^{1/2} \\&= \Sigma\end{aligned}$$

随机向量

证明: $\Sigma = \text{Cov}(\mathbf{x})$ (续)

只需要讨论 $\int_{R^p} \mathbf{z} \mathbf{z}' (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} = \mathbf{I}$ 的两种情况:

- 如果 $i \neq j$, 那么

$$\begin{aligned} & \int_{R^p} z_i z_j \cdot (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} \\ &= \int_{R^2} z_i z_j (2\pi)^{-1} \exp \left\{ -\frac{1}{2} (z_i^2 + z_j^2) \right\} dz_i dz_j \\ &= \int_R z_i (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} z_i^2 \right\} dz_i \\ & \quad \cdot \int_R z_j (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} z_j^2 \right\} dz_j \\ &= 0. \end{aligned}$$

随机向量

证明: $\Sigma = \text{Cov}(\mathbf{x})$ (续)

- 如果 $i = j$, 那么

$$\begin{aligned} & \int_{R^p} z_i^2 \cdot (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} \\ &= \int_R z_i^2 (2\pi)^{-1/2} \exp \left\{ -\frac{1}{2} z_i^2 \right\} dz_i \\ &= 1. \end{aligned}$$

综上,

$$\int_{R^p} \mathbf{z} \mathbf{z}' (2\pi)^{-p/2} \exp \left\{ -\frac{1}{2} \mathbf{z}' \mathbf{z} \right\} d\mathbf{z} = \mathbf{I}$$

随机向量

高斯分布随机向量

- 如果随机向量 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, 那么 \mathbf{x} 的 p.d.f. 为

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

- 特别地, 当 $\Sigma = \sigma^2 \mathbf{I}_p$ (\mathbf{I}_p 是单位矩阵) 时,

$$|\Sigma| = (\sigma^2)^p \quad \text{且} \quad \Sigma^{-1} = \sigma^{-2} \mathbf{I}_p,$$

于是, \mathbf{x} 的 p.d.f. 为

$$p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = (2\pi\sigma^2)^{-p/2} \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) \right\}.$$

随机向量

期望与协方差的性质

设随机变量 $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ 和 $\mathbf{y} = (y_1, y_2, \dots, y_p)'$ 。
设有 $m \times n$ 维常数矩阵 $\mathbf{A} = \{a_{ij}\}_{m \times n}$, $m' \times n$ 维常数矩阵 $\mathbf{B} = \{b_{ij}\}_{m' \times n}$ 及 m 维的常数向量 $\mathbf{c} = (c_1, c_2, \dots, c_m)'$, 可以证明以下结论:

- $E(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}E(\mathbf{x}) + \mathbf{c}$;
- $\text{Cov}(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}\text{Cov}(\mathbf{x})\mathbf{A}'$;
- $\text{Cov}(\mathbf{Ax}, \mathbf{By}) = \mathbf{A}\text{Cov}(\mathbf{x}, \mathbf{y})\mathbf{B}'$.

概率不等式

Jensen 不等式

给定一个概率空间 (Ω, \mathcal{F}, P) , x 是一个随机变量且 ϕ 是一个凸函数, 则

$$\phi(E(x)) \leq E(\phi(x)).$$

特殊形式

- 现有实数 x_1, x_2, \dots, x_n 且权重分别为 a_1, a_2, \dots, a_n , 有

$$\phi\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right) \leq \frac{\sum_{i=1}^n a_i \phi(x_i)}{\sum_{i=1}^n a_i}.$$

- 令 $\phi(x) = x^2$, 则 $(E(x))^2 \leq E(x^2)$.

概率不等式

Hoeffding 不等式

设 x_1, x_2, \dots, x_n 是相互独立的随机变量且 $P(a_i \leq x_i \leq b_i) = 1$ 。令

$$S_n = x_1 + x_2 + \dots + x_n.$$

对于任意 $t > 0$, 都

$$P(S_n - E(S_n) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$P(|S_n - E(S_n)| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

无约束优化问题

无约束的最优化问题本质上是

$$\min f(\boldsymbol{x}), \boldsymbol{x} \in R^n.$$

符号与定义

- 目标函数: $f(\boldsymbol{x}), \boldsymbol{x} \in R^n$.
- 最小值点: \boldsymbol{x}^* , 即

$$\boldsymbol{x}^* = \arg \min f(\boldsymbol{x}),$$

- 优化问题的根本目标: 得到 \boldsymbol{x}^* 解析解或近似解。

问题: 得到 \boldsymbol{x}^* 的解析解的通法是什么?

迭代算法

梯度下降中解的迭代式为

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 + \eta \cdot \left(- \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \bigg|_{\boldsymbol{x}=\boldsymbol{x}_0} \right)$$

其中,

- $-\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \bigg|_{\boldsymbol{x}=\boldsymbol{x}_0}$ 是梯度的反方向, 表示迭代时的方向;
- η 是学习率, 表示迭代时的步长;

迭代算法

梯度下降的原理

- 一阶泰勒展开公式为

$$f(\mathbf{x}_1) \approx f(\mathbf{x}_0) + \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \bigg|_{\mathbf{x}=\mathbf{x}_0} \right)' (\mathbf{x}_1 - \mathbf{x}_0)$$

- 令

$$\mathbf{x}_1 - \mathbf{x}_0 = l\mathbf{v}$$

其中,

- \mathbf{v} 是一个单位向量, 表示 $\mathbf{x}_1 - \mathbf{x}_0$ 的方向;
- $l > 0$ 是一个标量, 表示 $\mathbf{x}_1 - \mathbf{x}_0$ 的长度;

迭代算法

梯度下降的原理（续）

要使得 $f(\mathbf{x}_1) < f(\mathbf{x}_0)$ ，我们需要

$$\left(\left. \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} \right)' l \mathbf{v} < 0$$

令 $\left(\left. \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} \right) = \mathbf{u}$ 。因为 l 是一个正常数，不影响符号，所以，

$$\mathbf{u}' \mathbf{v} < 0$$

同时，我们知道

$$\mathbf{u}' \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \|\mathbf{v}\| \cos(\mathbf{u}, \mathbf{v}).$$

迭代算法

梯度下降的原理（续）

当 $\cos(\mathbf{u}, \mathbf{v}) = -1$ 时, $\mathbf{u}'\mathbf{v}$ 最小, 于是取

$$\mathbf{v} = -\frac{\mathbf{u}}{\|\mathbf{u}\|} = -\frac{\left. \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0}}{\left\| \left. \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} \right\|}.$$

因此,

$$\mathbf{x}_1 = \mathbf{x}_0 + \eta \cdot \left(-\left. \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_0} \right).$$

带约束的优化问题

带约束的最优化问题本质上是

$$\begin{aligned} \min \quad & f(\mathbf{x}), \mathbf{x} \in R^n, \\ \text{s.t.} \quad & h_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m; \\ & l_i(\mathbf{x}) = 0, i = 1, 2, \dots, r. \end{aligned}$$

符号与定义

- 拉格朗日函数为

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^m u_i h_i(\mathbf{x}) + \sum_{j=1}^r v_j l_j(\mathbf{x})$$

其中, $\mathbf{u} = (u_1, \dots, u_m)' \in R^m, \mathbf{v} = (v_1, \dots, v_r)' \in R^r$
且 $u_i \geq 0$.

带约束的优化问题

性质

对于 $u_i \geq 0, i = 1, 2, \dots, m$ 且 $v_i, i = 1, 2, \dots, r$, 有

$$f(\mathbf{x}) \geq L(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

原因如下:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^m u_i \underbrace{h_i(\mathbf{x})}_{\leq 0} + \sum_{j=1}^r v_j \underbrace{l_j(\mathbf{x})}_{=0} \leq f(\mathbf{x})$$

带约束的优化问题

令 \mathcal{X} 是原始可行集。假设 f^* 是带约束的优化问题

$$\begin{aligned} \min \quad & f(\mathbf{x}), \mathbf{x} \in R^n, \\ \text{s.t.} \quad & h_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m; \\ & l_i(\mathbf{x}) = 0, i = 1, 2, \dots, r \end{aligned}$$

的原始最优值。于是有

$$f^* \geq \min_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) \geq \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{u}, \mathbf{v}) := g(\mathbf{u}, \mathbf{v})$$

我们称 $g(\mathbf{u}, \mathbf{v})$ 为拉格朗日对偶函数。

带约束的优化问题

给定原始优化问题

$$\begin{array}{ll}\min & f(\mathbf{x}), \mathbf{x} \in R^n, \\ \text{s.t.} & h_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m; \\ & l_i(\mathbf{x}) = 0, i = 1, 2, \dots, r.\end{array}$$

对偶函数 $g(\mathbf{u}, \mathbf{v})$ 满足：对于任意 $\mathbf{u} \geq 0$ 和 \mathbf{v} ，有

$$f^* \geq g(\mathbf{u}, \mathbf{v}).$$

这样可以有**最优下界**： $\max_{\mathbf{u}, \mathbf{v}} g(\mathbf{u}, \mathbf{v})$ 。
这自然导出了拉格朗日对偶问题为

$$\begin{array}{ll}\max_{\mathbf{u}, \mathbf{v}} & g(\mathbf{u}, \mathbf{v}) \\ \text{s.t.} & \mathbf{u} \geq 0.\end{array}$$

带约束的优化问题

重要性质

- **弱对偶性**：如果对偶最优值为 g^* ，那么

$$f^* \geq g^*.$$

注意到：这个性质总是成立的（即使原始问题是非凸的）。

- 对偶问题是一个凸优化问题（即使原始问题是非凸的）。
- **强对偶性**：如果对偶最优值为 g^* ，那么

$$f^* = g^*.$$

注意到：这个性质成立时是**有条件的**。（Slater's 条件，KKT 条件）

带约束的优化问题

对偶问题的用途

- 在强对偶条件下，给定对偶最优值 u^*, v^* ，原始最优解 x^* 也是

$$\min_x L(x, u^*, v^*)$$

的解。

- 这表明：可以通过对偶问题的解来计算原始问题的解。
- 获取对偶问题的解，**可能更简单！**