尚硅谷大模型技术之数学基础

（作者：尚硅谷研究院）

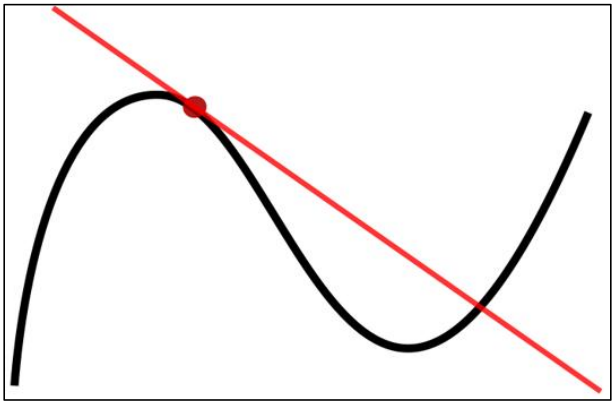
版本：V1.2.1

第 1 章 高等数学

1.1 导数

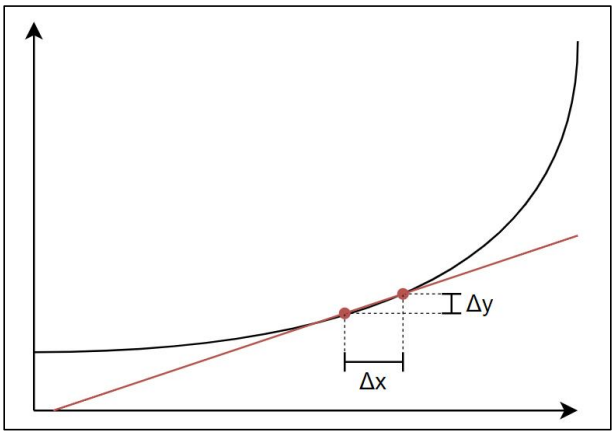
1.1.1 导数的概念

导数（derivative）是微积分中的一个概念。函数在某一点的导数是指这个函数在这一点附近的变化率（即函数在这一点的切线斜率）。导数的本质是通过极限的概念对函数进行局部的线性逼近。



当函数f的自变量在一点x_0上产生一个增量h时，函数输出值的增量Δy与自变量增量Δx的比值在Δx趋于 0 时的极限如果存在，即为f在x_0处的导数，记作$f'(x_0)$、$\frac{df}{dx}(x_0)$或$\frac{df}{dx}|_{x=x_0}$。

$$f^{'}(x_0)=\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$$



例如在运动学中，物体的位移对于时间的导数就是物体的瞬时速度：$v = \frac{dx}{dt}$。

1.1.2 基本函数的导数

说明	公式	例子
常数的导数	$(C)' = 0$	$(3)' = 0$
幂函数的导数	$(x^\alpha)' = \alpha x^{\alpha-1}$	$(x^3)' = 3x^2$
指数函数的导数	$(a^x)' = a^x \ln a$	$(3^x)' = 3^x \ln 3$
	$(e^x)' = e^x$	—
对数函数的导数	$(\log_a x)' = \frac{1}{x \ln a}$	$(\log_3 x)' = \frac{1}{x \ln 3}$
	$(\ln x)' = \frac{1}{x}$	—
三角函数的导数	$(\sin x)' = \cos x$	—
	$(\cos x)' = -\sin x$	—
	$(\tan x)' = \sec^2 x = \frac{1}{\cos^2 x}$	—
	$(\cot x)' = -\csc^2 x = \frac{-1}{\sin^2 x}$	—

1.1.3 导数的求导法则

说明	公式
两函数之和求导	$(f+g)' = f' + g'$
两函数之积求导	$(fg)' = f'g + fg'$
两函数之商求导	$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$

复合函数的导数	若$f(x)=h[g(x)]$, 则$f'(x)=h'[g(x)]\cdot g'(x)$

例如：求函数$f(x)=x^4+\sin(x^2)-\ln(x)\,e^x+7$在$x=3$处的导数。

$$f'(x)=4x^{4-1}+\cos(x^2)\cdot 2x-\left(\frac{e^x}{x}+\ln(x)\,e^x\right)+0$$

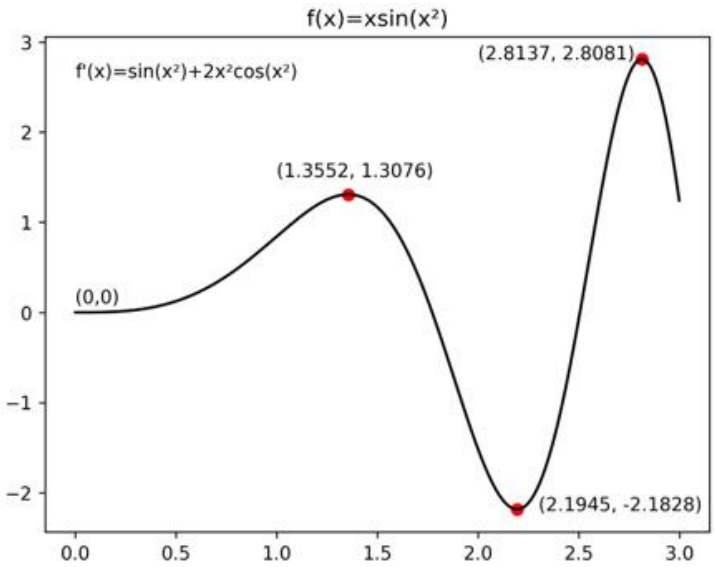
$$=4x^3+2xcos(x^2)-\frac{e^x}{x}-\ln(x)\,e^x$$

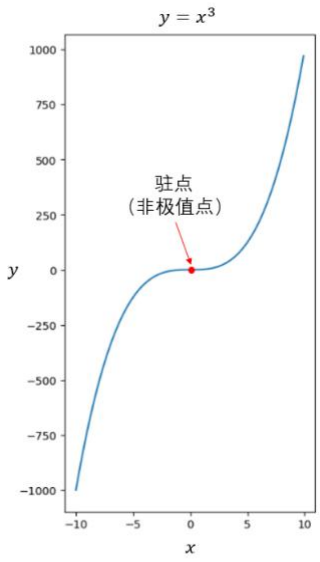
$$f'(3)=108+6\cos(9)-\frac{e^3}{3}-\ln(3)\,e^3$$

1.1.4 利用导数求极值

导数等于零的点称为函数的**驻点**（或极值可疑点），在这类点上函数可能会取得**极大值**或**极小值**。进一步判断则需要知道导数在附近的符号。

例如，$f(x)=x^3$ 在 $x=0$ 处导数为 0，但并不会取得极大值或者极小值。





1.1.5 二阶导数

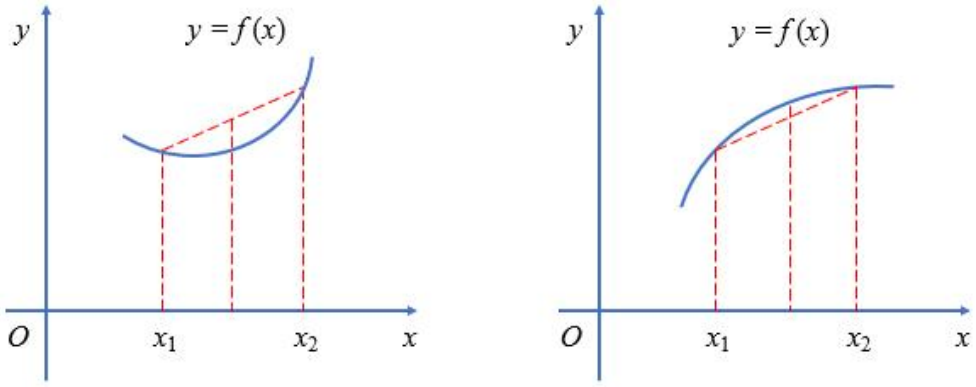
1）二阶导数的概念

在微积分中，函数的二阶导数是函数导数的导数。粗略来说，某个量的二阶导数描述该量变化率变化的快慢。例如物体位置对时间的二阶导数是物体的瞬时加速度，即该物体速度随时间的变化率：$a=\frac{dv}{dt}=\frac{d^2x}{dt^2}$。

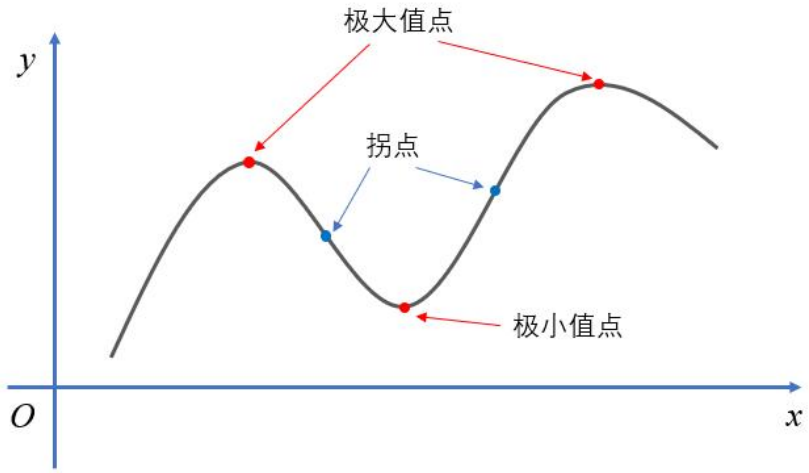
函数f的二阶导数通常记作f''、$\frac{d^2y}{dx^2}$或$\frac{d}{dx}\left(\frac{dy}{dx}\right)$。

2）二阶导数与函数凹凸的关系

函数的二阶导数描述了函数图像的凹凸方向和程度。若二阶导数在某区间恒为正，则函数在该区间向上弯（也称**下凸函数**）。反之，若二阶导数在某区间恒为负，则函数在该区间向下弯（也称**上凸函数**）。



若函数的二阶导数在某点左右异号，则图像由向上弯转为向下弯，或反之。这种点称之为**拐点**。若二阶导数连续，则在该点处二阶导数为 0。但反之二阶导数为 0 的点不一定是拐点。如$f(x) = x^4$，在$x = 0$ 处有$f''(0) = 0$，但$f(x)$在实数系上无拐点。



二阶导数与凹凸性的关系有助于判断函数的驻点是否为极大值点或极小值点：

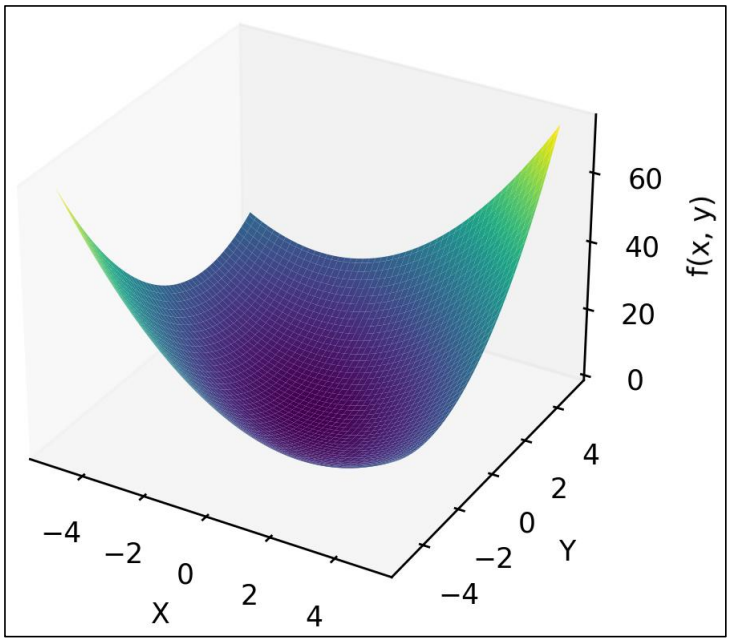
- 若$f'(x) = 0$，$f''(x) < 0$，则f在x取得极大值。
- 若$f'(x) = 0$，$f''(x) > 0$，则f在x取得极小值。
- 若$f'(x) = 0$，$f''(x) = 0$，则该点可能是拐点，也可能是极大值点或极小值点。

1.2 偏导与梯度

1.2.1 偏导数

如果函数f的自变量并非单个元素，而是多个元素，例如：

$$f(x,y) = x^2 + xy + y^2$$



可将其中一个元素x看作常数，此时f可看作关于另一元素y的函数。

$$f_x(y) = x^2 + xy + y^2$$

在$x = a$固定的情况下，可计算f_x关于y的导数：

$$f_{x=a}'(y) = a + 2y$$

这种导数称为**偏导数**，一般记作：

$$\frac{\partial f}{\partial y}(x,y) = x + 2y$$

更一般地来说，一个多元函数$f(x_1, x_2, ..., x_n)$在点$(a_1, a_2, ..., a_n)$处对x_i的偏导数定义为：

$$\frac{\partial f}{\partial x_i}(a_1, a_2, ..., a_n) = \lim_{\Delta x_i \to 0} \frac{f(a_1, ...a_i + \Delta x_i, ..., a_n) - f(a_1, ...a_i, ..., a_n)}{\Delta x_i}$$

1.2.2 方向导数

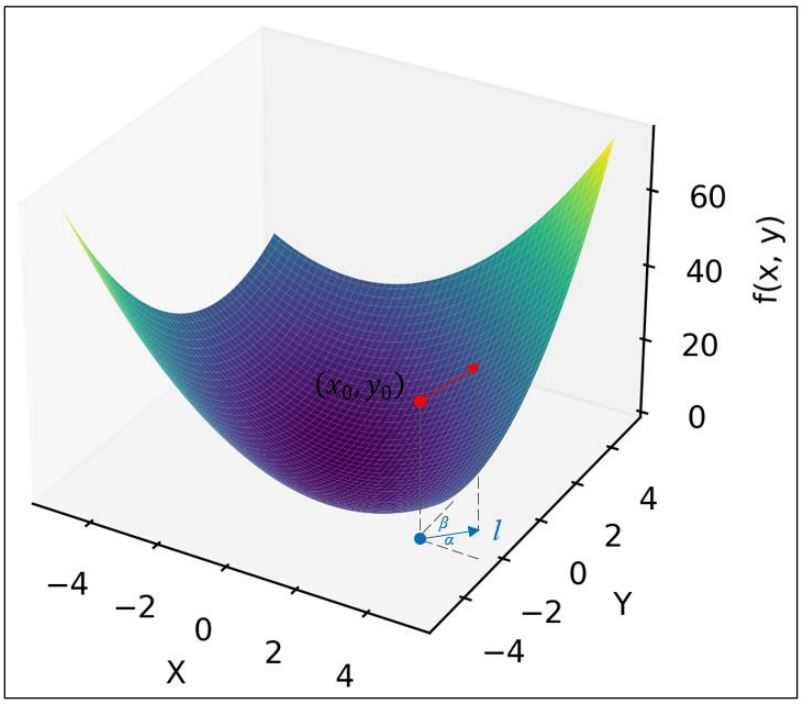
偏导数可以看作是多元函数 f 沿某个自变量轴方向的变化率。

如果我们任意选取一个方向 l，那么在某个点(x_0, y_0)处，二元函数$f(x,y)$ 沿着这个方向的变化率可以用极限定义为：

$$\frac{\partial f}{\partial l}(x_0, y_0) = \lim_{\Delta l \to 0} \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)}{\Delta l}$$

这里，Δl 就是沿方向 l 的微小改变量，Δx 和 Δy 与 Δl 的关系为：

$$\Delta x = \Delta l \cdot cos\alpha,\ \Delta y = \Delta l \cdot cos\beta$$



根据全微分公式，上式可以表示为：

$$\frac{\partial f}{\partial l}(x_0, y_0) = f_x(x_0, y_0)\cos\alpha + f_y(x_0, y_0)\cos\beta$$

其中$f_x(x_0, y_0)$、$f_y(x_0, y_0)$ 表示点(x_0, y_0)处f对 x、y 的偏导数；$\cos\alpha$、$\cos\beta$是方向 l 的方向余弦，即 l 方向的单位方向向量可以表示为$l_0 = (\cos\alpha,\ \cos\beta)$。

这个“沿某个方向的变化率”，就被称为$f(x, y)$ 沿方向 l 的**方向导数**。

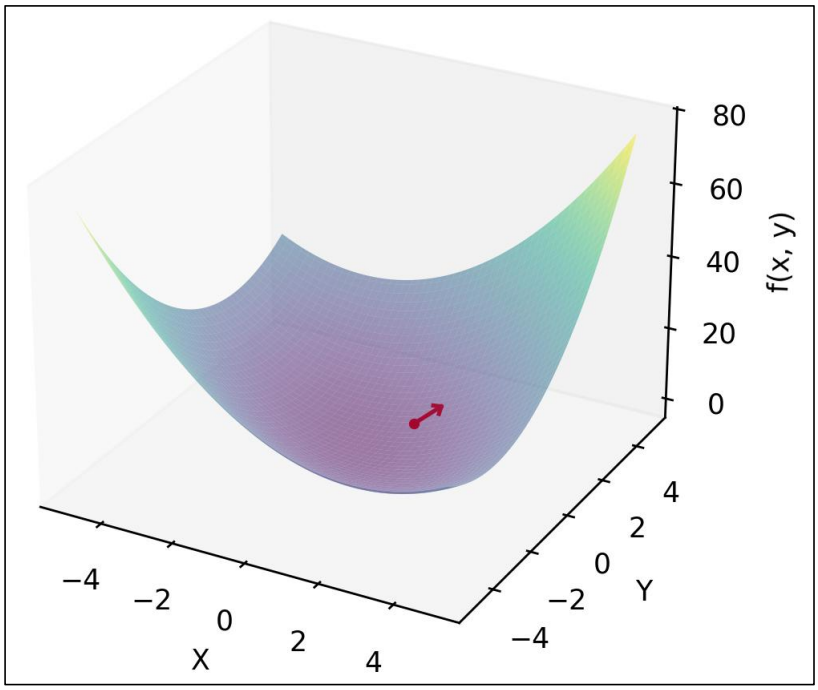
1.2.3 梯度

多元函数$f(x_1, ..., x_n)$关于每个变量x_i都有偏导数$\frac{\partial f}{\partial x_i}$，在点$a = (a_1, a_2, ..., a_n)$ 处，这些偏导数定义出一个向量：

$$\nabla f(a) = \left[\frac{\partial f}{\partial x_1}(a), \frac{\partial f}{\partial x_2}(a), ..., \frac{\partial f}{\partial x_n}(a)\right]$$

这个向量称为f在点a的**梯度**，记作 $\nabla f(a)$ 或者 $\mathbf{grad}\ f(a)$。

例如：$f(x, y) = x^2 + xy + y^2$在$(1,1)$处的梯度为$[3,3]$。



梯度向量表示的方向，就是函数在这一点处，方向导数取最大值的方向。换句话说，梯度的方向，就是函数值变化最快的方向。

第 2 章 线性代数

2.1 标量与向量

2.1.1 标量与向量的概念

1）标量（scalar）

标量是一个单独的数，只有大小。

2）向量（vector）

向量由标量组成，有大小有方向。

行向量：(2　5　8)

列向量：$\begin{pmatrix}2\\5\\8\end{pmatrix}$

2.1.2 向量运算

1）向量转置：列向量转置结果为行向量

$$\boldsymbol{x} = \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}$$

$$\boldsymbol{x}^T = (2 \quad 5 \quad 8)$$

2）向量相加：对应元素相加

$$\begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} + \begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} = \begin{pmatrix} 3 \\ 8 \\ 15 \end{pmatrix}$$

3）向量与标量相乘：标量与向量每个元素相乘

$$3 \times \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} = \begin{pmatrix} 6 \\ 15 \\ 24 \end{pmatrix}$$

4）向量内积：又称向量点乘，两向量对应元素乘积之和，结果为标量

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} \rangle = 2 + 15 + 56 = 73$$

两向量之间夹角表示为

$$\cos\theta = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}\sqrt{\langle \boldsymbol{y}, \boldsymbol{y} \rangle}}$$

2.1.3 向量范数

范数（norm）是具有“长度”概念的函数。

1）L_0 范数（也称 0 范数）

$$\|\boldsymbol{x}\|_0 = \text{非零元素的个数}$$

例如：

$$\boldsymbol{x} = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|\boldsymbol{x}\|_0 = 2$$

2）L_1 范数（也称和范数或 1 范数）

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{m} |x_i| = |x_1| + \ldots + |x_m|$$

例如：

$$\boldsymbol{x} = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|\boldsymbol{x}\|_1 = 0 + 2 + 1 = 3$$

3）L_2 范数（也称欧几里得范数或 2 范数）

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{m} |x_i|^2\right)^{\frac{1}{2}} = \sqrt{(|x_1|^2 + \ldots + |x_m|^2)}$$

例如：

$$\boldsymbol{x} = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \|\boldsymbol{x}\|_2 = \sqrt{0 + 4 + 1} = \sqrt{5}$$

4）L_p 范数

$$\|\boldsymbol{x}\|_{\mathrm{p}} = \left(\sum_{i=1}^{m} |x_i|^p\right)^{\frac{1}{p}} = (|x_1|^p + \ldots + |x_m|^p)^{\frac{1}{p}}$$

在 numpy 中，可以利用 linalg.norm 函数方便地计算向量的范数。

2.2 矩阵与张量

2.2.1 矩阵的概念

一个$m \times n$的矩阵（matrix）是一个有 m 行 n 列元素的矩形阵列。用$R^{m \times n}$表示所有$m \times n$实数矩阵的向量空间。

$$\begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in R^{3 \times 2}$$

1）方阵：行数等于列数的矩阵

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in R^{2 \times 2}$$

2）对角矩阵：主对角线以外元素全为 0 的方阵

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

3）单位矩阵：主对角线元素全为 1 的对角矩阵

$$\boldsymbol{I}_{3 \times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2.2.2 矩阵乘法

1）矩阵乘法运算

两个矩阵的乘法仅当矩阵$\boldsymbol{A}$的列数和矩阵$\boldsymbol{B}$的行数相等时才能定义。如$\boldsymbol{A} \in R^{m \times n}$，$\boldsymbol{B} \in R^{n \times p}$，它们的乘积$\boldsymbol{AB} \in R^{m \times p}$

$$[\boldsymbol{AB}]_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj} = \sum_{r=1}^{n} a_{ik}b_{kj}$$

例如：

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{bmatrix} \times \begin{bmatrix} 3 & 1 \\ 2 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1\times3+0\times2+2\times1 & 1\times1+0\times1+2\times0 \\ (-1)\times3+3\times2+1\times1 & (-1)\times1+3\times1+1\times0 \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 4 & 2 \end{bmatrix}$$

特别地，矩阵与单位矩阵相乘等于矩阵本身：

$$\boldsymbol{AI} = \boldsymbol{A}(\boldsymbol{A} \in R^{m\times n}, \boldsymbol{I} \in R^{n\times n}) \text{ 或 } \boldsymbol{IA} = \boldsymbol{A}(\boldsymbol{I} \in R^{n\times n}, \boldsymbol{A} \in R^{n\times m})$$

例如：

$$\boldsymbol{AI} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1\times1+2\times0 & 1\times0+2\times1 \\ 3\times1+5\times0 & 3\times0+5\times1 \\ 4\times1+8\times0 & 4\times0+8\times1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} = \boldsymbol{A}$$

2）矩阵乘法的性质

矩阵乘法满足结合律、左分配律和右分配律。不满足交换律即$\boldsymbol{AB} \neq \boldsymbol{BA}$。

结合律：若$\boldsymbol{A} \in R^{m\times n}, \boldsymbol{B} \in R^{n\times p}, \boldsymbol{C} \in R^{p\times q}$，则$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$

左分配律：若$\boldsymbol{A} \in R^{m\times n}, \boldsymbol{B} \in R^{m\times n}, \boldsymbol{C} \in R^{n\times p}$，则$(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$

右分配律：若$\boldsymbol{A} \in R^{m\times n}, \boldsymbol{B} \in R^{n\times p}, \boldsymbol{C} \in R^{n\times p}$，则$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$

2.2.3 矩阵转置

1）矩阵转置运算

矩阵$\boldsymbol{A} \in R^{m\times n}$的转置是一个$n \times m$的矩阵，记为$\boldsymbol{A}^T$。其中的第$i$个行向量是原矩阵的第$i$个列向量；或者说，转置矩阵$\boldsymbol{A}^T$第$i$行第$j$列的元素是原矩阵$\boldsymbol{A}$第$j$行第$i$列的元素。

$$[\boldsymbol{A}^T]_{ij} = a_{ji}$$

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \\ 4 & 8 \end{bmatrix} \in R^{3\times2}$$

$$\boldsymbol{A}^T = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 5 & 8 \end{bmatrix} \in R^{2\times3}$$

2）矩阵转置的性质

$(\boldsymbol{A}^T)^T = \boldsymbol{A}$

$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

$(\boldsymbol{kA})^T = \boldsymbol{kA}^T$

$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

2.2.4 矩阵的逆

对于方阵$\boldsymbol{A}$，如果存在另一个方阵$\boldsymbol{A}^{-1}$，使得$\boldsymbol{A}\boldsymbol{A}^{-1}=\boldsymbol{I}$成立，此时$\boldsymbol{A}^{-1}\boldsymbol{A}=\boldsymbol{I}$也同样成立。称$\boldsymbol{A}^{-1}$为$\boldsymbol{A}$的逆矩阵。例如：

$$\boldsymbol{A}\boldsymbol{A}^{-1}=\begin{bmatrix}1 & 2\\ 3 & 5\end{bmatrix}\times\begin{bmatrix}-5 & 2\\ 3 & -1\end{bmatrix}=\begin{bmatrix}1\times(-5)+2\times 3 & 1\times 2+2\times(-1)\\ 3\times(-5)+3\times 5 & 3\times 2+5\times(-1)\end{bmatrix}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=\boldsymbol{I}$$

2.2.5 其他矩阵运算

1）矩阵的向量化

矩阵$\boldsymbol{A}\in R^{m\times n}$的向量化 $\mathrm{vec}(\boldsymbol{A})$将矩阵$\boldsymbol{A}$的元素按列排列成一个$mn\times 1$ 的向量。

$$\mathrm{vec}(\boldsymbol{A})=[a_{11},\ldots a_{m1},\ldots a_{1n},\ldots a_{mn}]^T$$

矩阵也可以转化为行向量 $\mathrm{rvec}(\boldsymbol{A})$，称为矩阵的行向量化。

$$\mathrm{rvec}(\boldsymbol{A})=[a_{11},\ldots a_{1n},\ldots a_{m1},\ldots a_{mn}]$$

例如：$\boldsymbol{A}=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix},\mathrm{vec}(\boldsymbol{A})=\begin{pmatrix}1\\ 3\\ 2\\ 4\end{pmatrix},\mathrm{rvec}(\boldsymbol{A})=(1\quad 2\quad 3\quad 4)$。

2）矩阵的内积

矩阵$\boldsymbol{A}\in R^{m\times n}$和矩阵$\boldsymbol{B}\in R^{m\times n}$的内积记作$\langle\boldsymbol{A},\boldsymbol{B}\rangle$，它是两个矩阵对应元素乘积之和，是一个标量。

$$\langle\boldsymbol{A},\boldsymbol{B}\rangle=\langle\mathrm{vec}(\boldsymbol{A}),\mathrm{vec}(\boldsymbol{B})\rangle=\sum a_{i,j}b_{i,j}$$

3）矩阵的 Hadamard 积

矩阵$\boldsymbol{A}\in R^{m\times n}$和矩阵$\boldsymbol{B}\in R^{m\times n}$的 Hadamard 积记作$\boldsymbol{A}\odot\boldsymbol{B}$，它是两个矩阵对应元素的乘积，是一个$m\times n$的矩阵。

$$(\boldsymbol{A}\odot\boldsymbol{B})_{ij}=a_{ij}b_{ij}$$

4）矩阵的 Kronecker 积

矩阵$\boldsymbol{A}\in R^{m\times n}$和矩阵$\boldsymbol{B}\in R^{p\times q}$的 Kronecker 积记作$\boldsymbol{A}\otimes\boldsymbol{B}$，它是矩阵$\boldsymbol{A}$中每个元素与矩阵$\boldsymbol{B}$的乘积，是一个$mp\times nq$的矩阵。

$$\begin{aligned}(\boldsymbol{A}\otimes\boldsymbol{B})_{ij}&=[\boldsymbol{a}_1\boldsymbol{B}\quad\boldsymbol{a}_2\boldsymbol{B}\quad\cdots\quad\boldsymbol{a}_n\boldsymbol{B}]\\&=[a_{ij}\boldsymbol{B}]_{i=1,j=1}^{m,n}\\&=\begin{bmatrix}a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B}\\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B}\end{bmatrix}\end{aligned}$$

Kronecker 积也称为直积或张量积。

2.2.6 张量

张量（tensor）可视为多维数组，是标量，1 维向量和 2 维矩阵的 n 维推广。

例如：3 维张量

$$\begin{bmatrix}\begin{bmatrix}1 & 2\\ 3 & 5\\ 4 & 8\end{bmatrix} & \begin{bmatrix}3 & 2\\ 1 & 6\\ 7 & 3\end{bmatrix} & \begin{bmatrix}5 & 6\\ 9 & 1\\ 2 & 4\end{bmatrix}\end{bmatrix}$$

2.3 矩阵求导

矩阵求导的本质就是函数对变元的每个元素逐个求导，只是写成了向量、矩阵的形式。

为方便理解，这里对变元（自变量）和函数作统一的符号规定：

$\boldsymbol{x} = [x_1, x_2, ..., x_m]^T \in R^m$为实向量变元。

$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m]^T \in R^{m\times n}$为实矩阵变元。

$f(\boldsymbol{x}) \in R$为实标量函数，其变元$\boldsymbol{x}$为实向量。

$f(\boldsymbol{X}) \in R$为实标量函数，其变元$\boldsymbol{X}$为实矩阵。

$\boldsymbol{f}(\boldsymbol{x}) \in R^p$为实向量函数，其变元$\boldsymbol{x}$为实向量。

$\boldsymbol{f}(\boldsymbol{X}) \in R^p$为实向量函数，其变元$\boldsymbol{X}$为实矩阵。

当然，函数也可以是矩阵形式 $\boldsymbol{F}$，可以看作是向量函数的扩展。

2.3.1 典型计算场景

（1）标量$f(\boldsymbol{x})$对向量$\boldsymbol{x}$求导

数学上，一般定义向量为列向量形式。由于$\boldsymbol{x} = [x_1, x_2, ..., x_m]^T$，所以

$$f(\boldsymbol{x}) = f(x_1, x_2, ..., x_m)$$

它对向量变元 $\boldsymbol{x}$ 求导，本质就是对 $\boldsymbol{x}$ 的每个元素求偏导：

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_m}\right]^T$$

例如：

函数 $f(x_1, x_2) = {x_1}^2 + x_1x_2 + 2{x_2}^2$，令$\boldsymbol{x} = [x_1, x_2]^T$，则

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right]^T = [2x_1 + x_2,\ x_1 + 4x_2]^T = \begin{bmatrix}2x_1 + x_2\\ x_1 + 4x_2\end{bmatrix}$$

（2）标量$f(\boldsymbol{X})$对矩阵$\boldsymbol{X}$求导

变元 $\boldsymbol{X}$ 是一个$m \times n$的矩阵，可以看成是 m 个行向量的组合，每个向量维度为 n。

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_m]^T = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix}^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \in R^{m \times n}$$

那么 f 对 $\boldsymbol{X}$ 求导，同样也是对其中的每个元素求导，结果形状与 $\boldsymbol{X}$ 相同：

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left[\frac{\partial f}{\partial \boldsymbol{x_1}}, \frac{\partial f}{\partial \boldsymbol{x_2}}, ..., \frac{\partial f}{\partial \boldsymbol{x_m}}\right]^T = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \frac{\partial f}{\partial x_{12}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \frac{\partial f}{\partial x_{2n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}$$

（3）向量$\boldsymbol{f}(x)$对标量x求导

对于向量函数 $\boldsymbol{f}$，可以写为：

$$\boldsymbol{f}(x) = [f_1(x), f_2(x), ..., f_p(x)]^T$$

显然，此时的 $\boldsymbol{f}$ 可以看作多个函数的组合，对 x 求导时，只要让其中的每个元素分别对 x 求导即可：

$$\frac{\partial \boldsymbol{f}(x)}{\partial x} = \left[\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}, ..., \frac{\partial f_p}{\partial x}\right]^T$$

例如：

函数 $\boldsymbol{f}(x) = [f_1(x), f_2(x)]^T = \begin{bmatrix} 2x^2 + 3x + 1 \\ \sin x \end{bmatrix}$，则

$$\frac{\partial \boldsymbol{f}(x)}{\partial x} = \left[\frac{\partial f_1}{\partial x}, \frac{\partial f_2}{\partial x}\right]^T = [4x + 3, \cos x]^T = \begin{bmatrix} 4x + 3 \\ \cos x \end{bmatrix}$$

（4）向量$\boldsymbol{f}(\boldsymbol{x})$对向量$\boldsymbol{x}$求导（了解）

类似（3）中的分析，向量函数 $\boldsymbol{f}$ 可以写为：

$$\boldsymbol{f}(x) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_p(\boldsymbol{x})]^T$$

这里 $\boldsymbol{f}$ 的每一个元素，都是变元 $\boldsymbol{x}$ 为向量的实标量函数。

接下来只要应用（1）中的结论，将 $\boldsymbol{f}$ 中的每个标量函数对 $\boldsymbol{x}$ 求导即可：

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\frac{\partial f_1}{\partial \boldsymbol{x}}, \frac{\partial f_2}{\partial \boldsymbol{x}}, ..., \frac{\partial f_p}{\partial \boldsymbol{x}}\right]^T$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_p}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_p}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_m} & \frac{\partial f_2}{\partial x_m} & \cdots & \frac{\partial f_p}{\partial x_m} \end{bmatrix}^T \in R^{p \times m}$$

2.3.2 常用求导结果

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^T$$

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A}}{\partial \boldsymbol{x}} = \boldsymbol{A}$$

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A}^T + \boldsymbol{A})\boldsymbol{x}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{b}}{\partial \boldsymbol{X}} = \frac{\partial (\boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{b})^T}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{b}^T \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{X}} = \boldsymbol{b}\boldsymbol{a}^T$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{X}^T \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T \boldsymbol{X} + \boldsymbol{b}\boldsymbol{a}^T \boldsymbol{X}$$

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{X}\boldsymbol{b}\boldsymbol{a}^T + \boldsymbol{X}\boldsymbol{a}\boldsymbol{b}^T$$

2.3.3 梯度矩阵

对于实向量变元 $\boldsymbol{x}$，实标量函数$f(\boldsymbol{x})$的梯度向量，为$m \times 1$ 的列向量（与 $\boldsymbol{x}$ 形状相同）：

$$\nabla_x f(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_m} \right]^T$$

对于矩阵变元 $\boldsymbol{X}$（m × n），可以类似地得到 $f(\boldsymbol{X})$ 的梯度矩阵：

$$\nabla_X f(\boldsymbol{X}) = \frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial x_{mn}} \end{bmatrix} \in R^{m \times n}$$

类似地，$f(\boldsymbol{x})$ 的二阶偏导构成的矩阵被称为“黑塞矩阵”（Hessian Matrix）：

$$H(\boldsymbol{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{n \times n}$$

第 3 章 概率论

3.1 概率

3.1.1 概率的概念

概率是对事件发生的可能性的度量。通常将事件 A 的概率写作$P(A)$。

3.1.2 概率的计算

事件	概率
A	$P(A) \in [0,1]$
非 A	$P(\overline{A}) = 1 - P(A)$
A 和 B **（联合概率）**	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ 当 A、B 相互独立时：$P(A \cap B) = P(A)P(B)$
A 或 B	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 当 A、B 互斥时：$P(A \cup B) = P(A) + P(B)$
B 的情况下 A 的概率 **（条件概率）**	$P(A\|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B\|A)P(A)}{P(B)}$

例如：现有一个装有 10 个球的袋子，其中有 6 个红球和 4 个蓝球。从中随机抽取两个球。我们定义以下事件：

事件 A：第一个抽到的是红球。

事件 B：两个抽到的球都是红球。

1）计算联合概率$P(A \cap B)$

第一个球是红球的概率：

$$P(A) = \frac{6}{10}$$

在第一个球是红球的情况下，两个球都是红球的概率：

$$P(B|A) = \frac{5}{9}$$

联合概率：

$$P(A \cap B) = P(B|A)P(A) = \frac{5}{9} \times \frac{6}{10} = \frac{1}{3}$$

2）计算条件概率$P(A|B)$

条件概率$P(A|B)$表示在已知两个球都是红球的情况下，第一个球是红球的概率。

两个球都是红球的概率：

$$P(B)=\frac{C_6^2}{C_{10}^2}=\frac{6\times5\div2}{10\times9\div2}=\frac{1}{3}$$

在两个球都是红球的情况下，第一个球是红球的概率：

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{1}{3}}{\frac{1}{3}}=1$$

3.2 概率分布

概率分布，是指用于表述随机变量取值的概率规律。事件的概率表示了一次试验中某一个结果发生的可能性大小。如果试验结果用变量 X 的取值来表示，则随机试验的概率分布就是随机变量的概率分布，即随机变量的可能取值及取得对应值的概率。

3.2.1 均匀分布

均匀分布也叫矩形分布，它表示在相同长度间隔的分布概率是等可能的。 均匀分布由两个参数 a 和 b 定义，它们是数轴上的最小值和最大值，通常缩写为 U（a，b）。

均匀分布的概率密度函数可写为：

$$P(x)=\frac{1}{b-a}，\ a<x<b$$

$$P(x)=0,\ \ else$$

3.2.2 正态分布

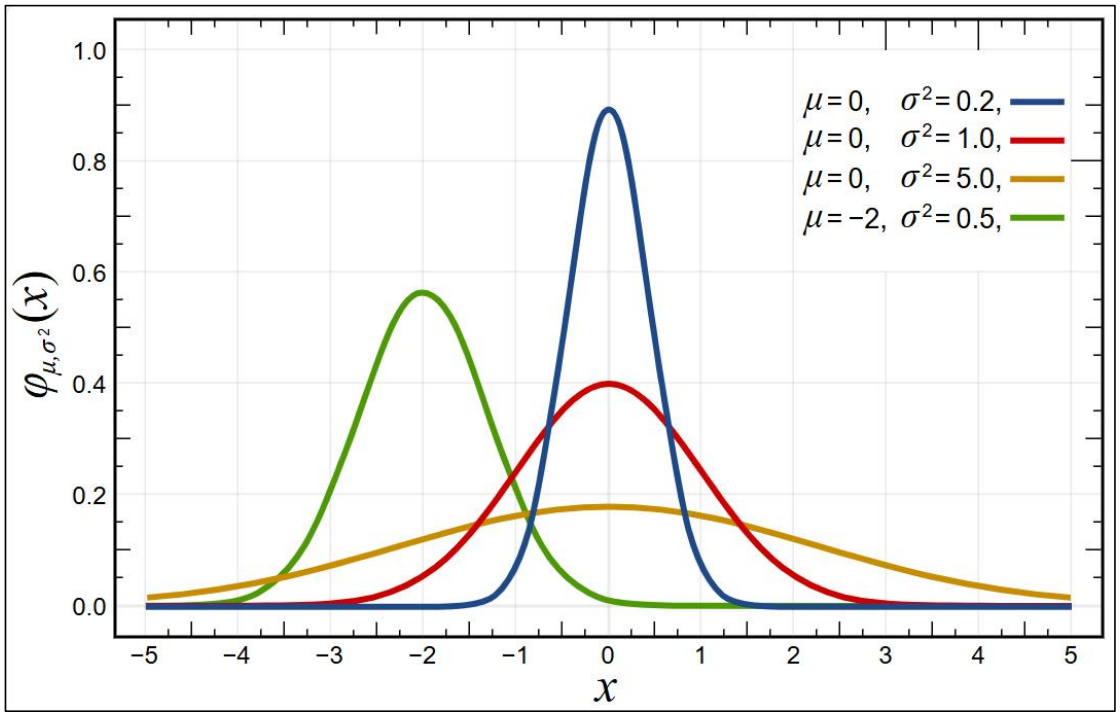
正态分布（Normal Distribution）也称高斯分布，是常见的连续概率分布。正态分布在统计学上十分重要，经常用在自然和社会科学来代表一个不明的随机变量。

若随机变量X服从一个平均数为μ、标准差为σ（$\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$）的正态分布，则记为$X\sim N(\mu,\sigma^2)$，其概率密度函数

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

正态分布的期望μ可解释为位置参数，决定了分布的位置；其方差σ^2可解释为尺度参数，决定了分布的幅度。

正态分布概率密度函数图：



正态分布的概率密度函数曲线呈钟形，因此人们又经常称之为钟形曲线。我们通常所说的标准正态分布是位置参数$\mu = 0$，尺度参数$\sigma^2 = 1$ 的正态分布。

中心极限定理指出，在特定条件下，一个具有有限均值和方差的随机变量的多个样本的平均值本身就是一个随机变量，其分布随着样本数量的增加而收敛于正态分布。因此，许多与独立过程总和有关的物理量，例如测量误差，通常可被近似为正态分布。

在 numpy 中，提供了各种随机函数，可以用来生成服从特定分布的数据。

3.3 贝叶斯定理

贝叶斯定理（Bayes' Theorem）是概率论中的一个核心定理，用于描述在已有条件概率信息的基础上，如何更新或计算事件的概率。它以英国数学家托马斯·贝叶斯的名字命名。贝叶斯定理特别适合处理“逆向概率”问题，即从结果反推原因的概率。

3.3.1 全概率公式

对于复杂事件 B，它可能有很多种具体情况，发生概率不容易直接求得。

这些不同的具体情况可以是一组简单事件，记作 A_1、A_2、…、A_n，发生的概率 $P(A_i)$ 已知；如果它们满足两两互不相容、且发生概率之和为 1，就称它们是一个完备事件组。

这样，如果知道了在每个简单事件发生的前提下、复杂事件发生的概率（条件概率 $P(B|A_i)$ ），就可以将它们全部合并起来，求出复杂事件的概率了。

$$P(B) = P(B|A_1)\cdot P(A_1) + P(B|A_2)\cdot P(A_2) + \ldots + P(B|A_n)\cdot P(A_n)$$

$$= \sum_{i}^{n} P(B|A_i)\cdot P(A_i)$$

这个公式就被称为“全概率公式”。

3.3.2 贝叶斯公式

贝叶斯定理建立在条件概率的基础上，假设有两事件A, B，贝叶斯定理描述了在已知B发生的情况下，A发生的概率：

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

- $P(A|B)$：后验概率，B发生的情况下A发生的概率。
- $P(B|A)$：似然概率，A发生的情况下B发生的概率。
- $P(A)$：先验概率，A发生的概率。
- $P(B)$：B发生的概率，通常通过全概率公式计算。

在实际问题中$P(B)$通常不是直接给出，而是需要通过全概率公式计算。假设样本空间被一组互斥且完备的事件$A_1, A_2, \ldots, A_n$划分，则：

$$P(B) = \sum_{i=1}^{n} P(B|A_i)\cdot P(A_i)$$

例如：某疾病发病率为 1%，如果一个人有疾病，检测呈阳性的概率为 95%；如果一个人没有疾病，检测呈阳性的概率为 5%，现有一人检测结果呈阳性，问他真正患病的概率是多少？

$$P(\text{疾病}|\text{阳性}) = \frac{P(\text{阳性}|\text{疾病})\cdot P(\text{疾病})}{P(\text{阳性})}$$

$$P(\text{阳性}) = P(\text{阳性}|\text{疾病})\cdot P(\text{疾病}) + P(\text{阳性}|\text{无疾病})\cdot P(\text{无疾病})$$

$$= 0.95 \times 0.01 + 0.05 \times 0.99$$

$$= 0.0095 + 0.0495$$

$$= 0.059$$

$$P(\text{疾病}|\text{阳性}) = \frac{0.0095}{0.059} \approx 0.161$$

检测呈阳性的人真正患病的概率为 16.1%。

3.4 似然函数

3.4.1 似然函数的概念

概率用于在已知一些参数的情况下，预测接下来在观测上所得到的结果。而似然性则是用于在已知某些观测所得到的结果时，对有关事物之性质的参数进行估值。

似然函数是对参数的函数，其定义为在给定参数值的条件下，观察到某个特定数据的概率。换句话说，似然函数是一个关于参数的函数，而不是关于数据的函数。

如果我们有一个参数化的概率模型$P(X|\theta)$，其中X是观测数据，θ是模型参数，似然函数$L(\theta|X)$定义为：

$$L(\theta|X) = P(X|\theta)$$

这里，$P(X|\theta)$ 表示在参数为θ的情况下，观察到数据X的概率。

设有一组独立同分布的观测数据$X = (x_1, x_2, ..., x_n)$，并且这些数据服从某个分布（例如正态分布、二项分布等），比如服从参数为θ的某个分布，那么似然函数可以写作：

$$L(\theta|X) = P(X|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$$

针对其中存在的乘法，可以使对数函数将其转化为加法：

$$\log L(\theta|X) = \log \prod_{i=1}^{n} P(x_i|\theta) = \sum_{i=1}^{n} \log P(x_i|\theta)$$

3.4.2 极大似然估计

似然函数常用于极大似然估计。我们希望找到使似然函数最大化的参数θ。这意味着在给定观测数据的情况下，选择最可能生成这些数据的参数值。

例如，掷硬币 3 次，2 次正面 1 次背面，能否依据此结果逆推出正面的概率；正面概率为 0.5 的概率为多少、正面概率为 0.6 的概率为多少；最有可能的正面概率是多少？

我们用θ代表硬币正面朝上的概率，用X代表 2 次正面 1 次背面的结果

$$L(\theta|X) = P(X|\theta) = C_3^2\theta^2(1-\theta)$$

当正面概率为 0.5 时：$P(X|\theta = 0.5) = C_3^2 0.5^2(1-0.5) = 0.375$

当正面概率为 0.6 时：$P(X|\theta = 0.6) = C_3^2 0.6^2(1-0.6) = 0.432$

为了找出极大似然估计，对似然函数取对数并求导，使其等于 0

$$\log L(\theta|X) = \log C_3^2 \cdot \theta^2(1-\theta)$$
$$= \log 3 + 2\log\theta + \log(1-\theta)$$
$$\frac{d\log L(\theta|X)}{d\theta} = \frac{2}{\theta} - \frac{1}{1-\theta} = 0$$

解得$\theta = \frac{2}{3}$，意味着当掷硬币 3 次，出现 2 次正面 1 次背面的结果时，硬币正面朝上的概率最有可能为$\frac{2}{3}$。