

自然语言处理

第3讲：数学基础

刘洋



内容提要

微积分

概率论

线性代数

信息论

函数

设数集 $D \subset \mathbb{R}$ ，则称映射 $f: D \rightarrow \mathbb{R}$ 为定义在 D 上的函数，通常记为

$$y = f(x), x \in D$$

其中 x 称为自变量， y 称为因变量， D 称为定义域，记作 D_f ，即 $D_f = D$ 。

对于每个 $x \in D$ ，按对应法则 f ，总有唯一的值 y 与之相对应，这个值称为函数 f 在 x 处的函数值，记作 $f(x)$ ，即 $y = f(x)$ 。函数值 $f(x)$ 的全体所构成的集合称为函数 f 的值域，记作 R_f 或 $f(D)$ ，即

$$R_f = f(D) = \{y \mid y = f(x), x \in D\}$$

例如， $f(x) = 3x + 2$ 是一个函数，定义域是 \mathbb{R} ，值域是 \mathbb{R} ，自变量和因变量之间存在一一映射。表示函数的记号可以任意选取，除了常用的 f 以外，还可以用其他的英文字母或希腊字母，如 g 、 F 和 ϕ 。

复合函数

给定两个函数 f 和 g ，**复合函数**定义为

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

两个函数 f 和 g 能构成复合函数 $f \circ g$ 的条件是：函数 g 的值域 R_g 必须是函数 f 的定义域 D_f 的子集，即 $R_g \subseteq D_f$ 。

例如， $y = f(u) = 3u + 2$ 的定义域为 \mathbb{R} ，而 $u = g(x) = x^2 - 2$ 的定义域为 \mathbb{R} 。由于 $g(\mathbb{R}) \subseteq \mathbb{R}$ ，因此 f 和 g 可以构成复合函数

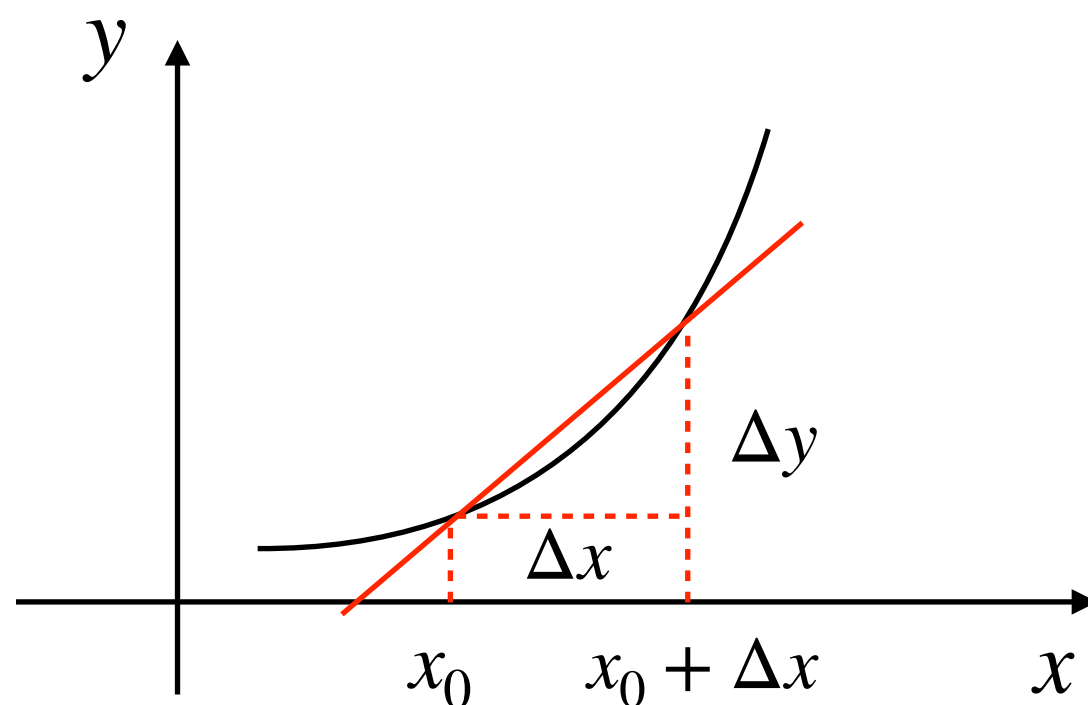
$$(f \circ g)(x) = 3(x^2 - 2) + 2 = 3x^2 - 4$$

另一个例子是 $y = f(u) = \sqrt{u}$ ，而 $u = g(x) = x - 2$ ，由于函数 g 的值域为 \mathbb{R} ，而函数 f 的定义域是 $\{u \mid u \geq 0\}$ ，不满足内层函数值域是外层函数定义域子集的约束条件，所以无法构成复合函数。

导数

设函数 $y = f(x)$ 在点 x_0 的某个邻域内有定义，当自变量 x 在 x_0 处有增量 Δx ，而且 $x_0 + \Delta x$ 也在该邻域内时，函数取得增量 $\Delta y = f(x_0 + \Delta x) - f(x_0)$ 。如果 Δy 与 Δx 之比当 $\Delta x \rightarrow 0$ 时极限存在，则称函数 $y = f(x)$ 在点 x_0 处可导，并称这个极限为函数 $y = f(x)$ 在点 x_0 处的导数，记作

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



导函数

如果函数 $y = f(x)$ 在开区间内每一点都可导，则称函数 $f(x)$ 在区间内可导。这时函数 $y = f(x)$ 对于区间内的每一个确定的 x 值，都对应着一个确定的导数值，这就构成一个新的函数。我们将该函数称之为原来函数的**导函数**，记作 y' 、 $f'(x)$ 或 $df(x)/dx$ ，简称**导数**。

名称	函数	导函数
常函数	$f(x) = C$	$f'(x) = 0$
幂函数	$f(x) = x^n$	$f'(x) = nx^{n-1}$
对数函数	$f(x) = a^x$	$f'(x) = a^x \ln a$
对数函数	$f(x) = \log_a x$	$f'(x) = 1/(x \ln a)$
正弦函数	$f(x) = \sin x$	$f'(x) = \cos x$
余弦函数	$f(x) = \cos x$	$f'(x) = -\sin x$

导数的四则运算

对于可导函数 f 和 g ，导数的四则运算规则如下：

$$\text{加法: } (f + g)' = f' + g'$$

$$\text{减法: } (f - g)' = f' - g'$$

$$\text{乘法: } (fg)' = f'g + fg'$$

$$\text{除法: } (f/g)' = (f'g - fg')/g^2$$

例如，令 $f(x) = x^2$ ， $g(x) = x$ ，则 $f'(x) = 2x$ ， $g'(x) = 1$ 。四则运算如下：

1. $f(x) + g(x) = x^2 + x$ 的导函数是 $2x + 1$ 。

2. $f(x) - g(x) = x^2 - x$ 的导函数是 $2x - 1$ 。

3. $f(x)g(x) = x^3$ 的导函数是 $2x \times x + x^2 \times 1 = 3x^2$ 。

4. $f(x)/g(x) = x$ 的导函数是 $(2x \times x - x^2 \times 1)/(x \times x) = 1$ 。

复合函数的导数

对于复合函数 $(f \circ g)(x)$ ，通常使用链式法则计算其导数：

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

令 $u = g(x)$ ，则链式法则的另一种表述方式为

$$\frac{df(g(x))}{dx} = \frac{df(u)}{du} \times \frac{du}{dx}$$

例如，令 $f(u) = u^2$ ， $u = g(x) = x + 1$ ，则 $f'(u) = 2u$ ， $g'(x) = 1$ 。直接复合函数可以得到 $(f \circ g)(x) = (x + 1)^2$ ，直接求导得到 $(f \circ g)'(x) = 2(x + 1)$ ，使用链式法则的结果是 $2 \times (x + 1) \times 1 = 2(x + 1)$ ，因此验证了链式法则的正确性。

链式法则对于计算复合函数的导数而言非常重要，因此在神经网络等使用复合函数的计算模型中具有广泛的应用。

二阶导数

一般而言，函数 $y = f(x)$ 的导数 $y' = f'(x)$ 仍然是 x 的函数，可以进一步求导。**二阶导数**是原函数导数的导数，即对原函数进行二次求导，记作

$$y'' = (y')'$$

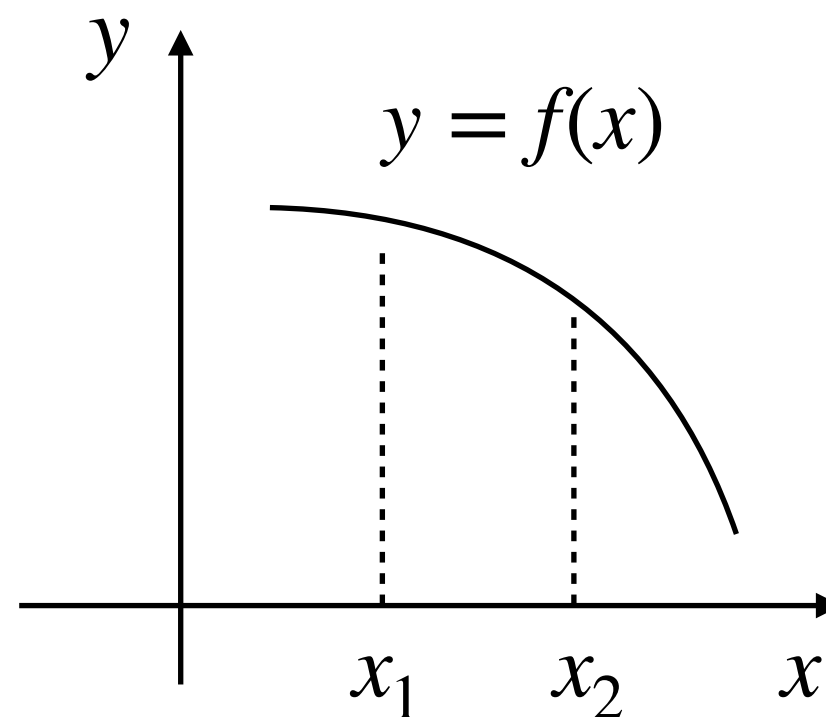
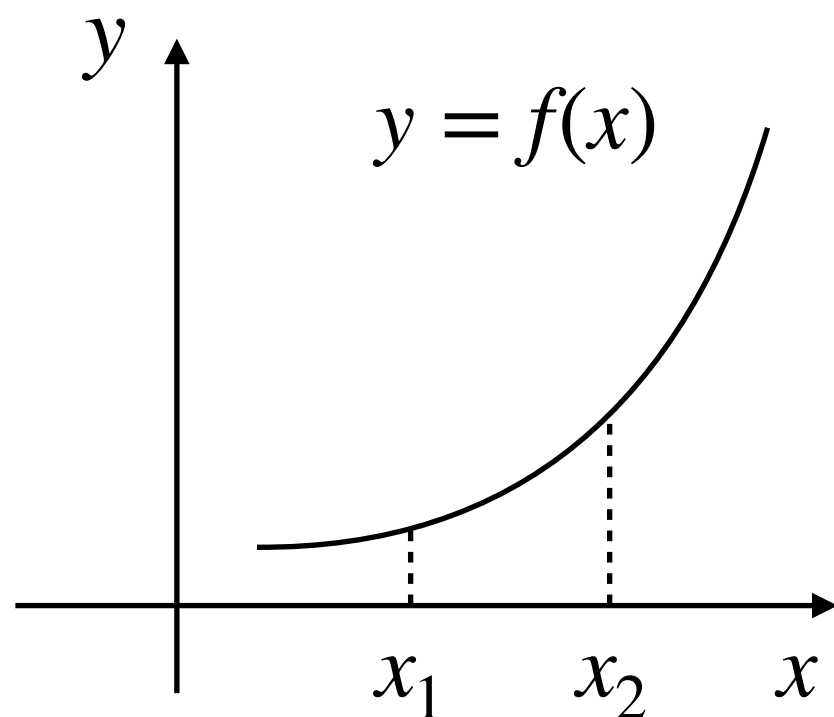
二阶导数的另一种常见的表示方法为

$$y'' = \frac{d^2y}{dx^2}$$

例如， $y = x^2$ 的一阶导数为 $y' = 2x$ ，而二阶导数则是一阶导数 $y' = 2x$ 的导数 $y'' = 2$ 。

二阶导数反映了一阶导数的变化率。我们通常使用二阶导数来判断函数的凹凸性并计算极值。类似地，在条件允许的情况下，还可以计算函数的三阶导数、四阶导数或高阶导数。

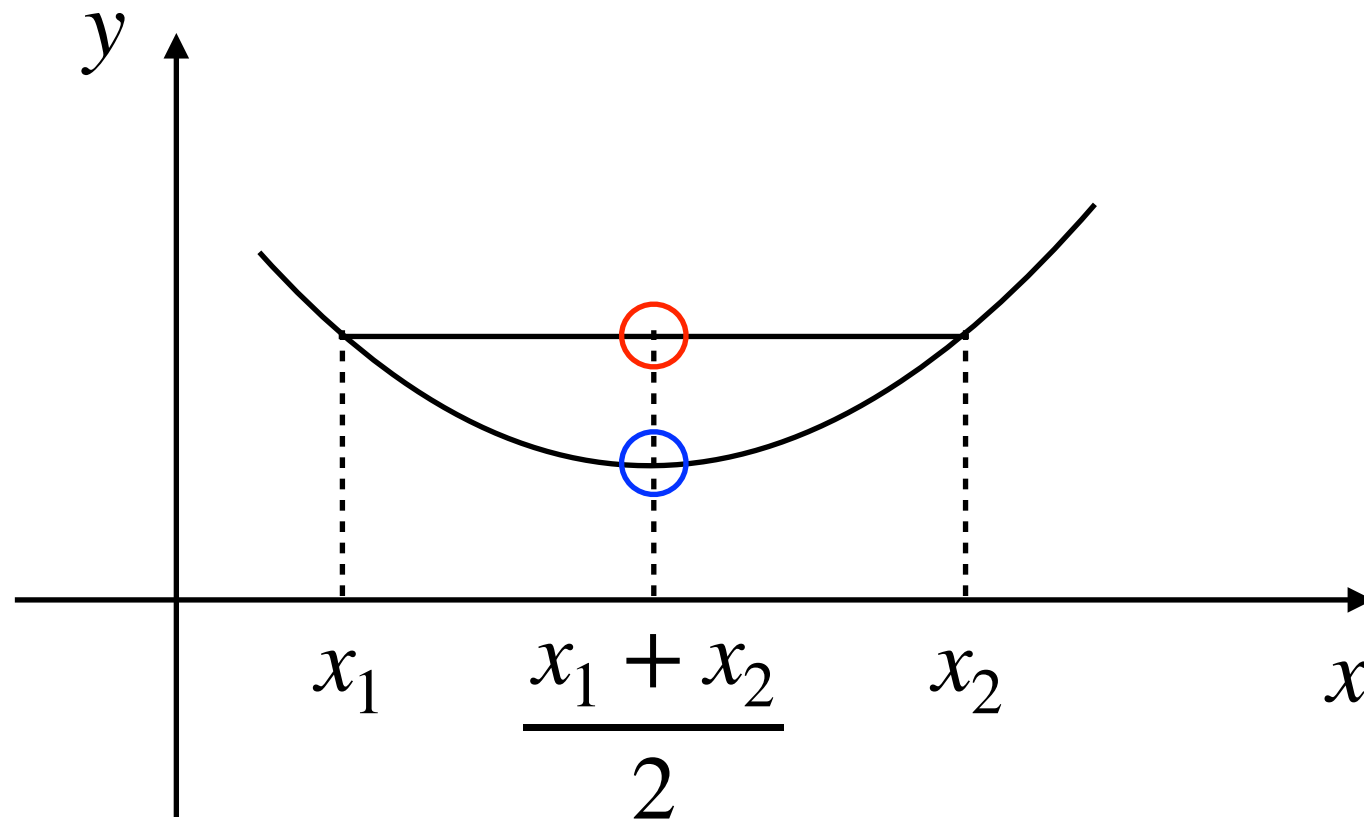
函数的单调性



设函数 $f(x)$ 的定义域为 D ，区间 $I \subset D$ 。如果对于区间 I 上任意两点 x_1 和 x_2 ，当 $x_1 < x_2$ 时，恒有 $f(x_1) < f(x_2)$ ，则称函数 $f(x)$ 在区间 I 上**单调递增**。

反之，如果对于区间 I 上任意两点 x_1 和 x_2 ，当 $x_1 < x_2$ 时，恒有 $f(x_1) > f(x_2)$ ，则称函数 $f(x)$ 在区间 I 上**单调递减**。

凹函数



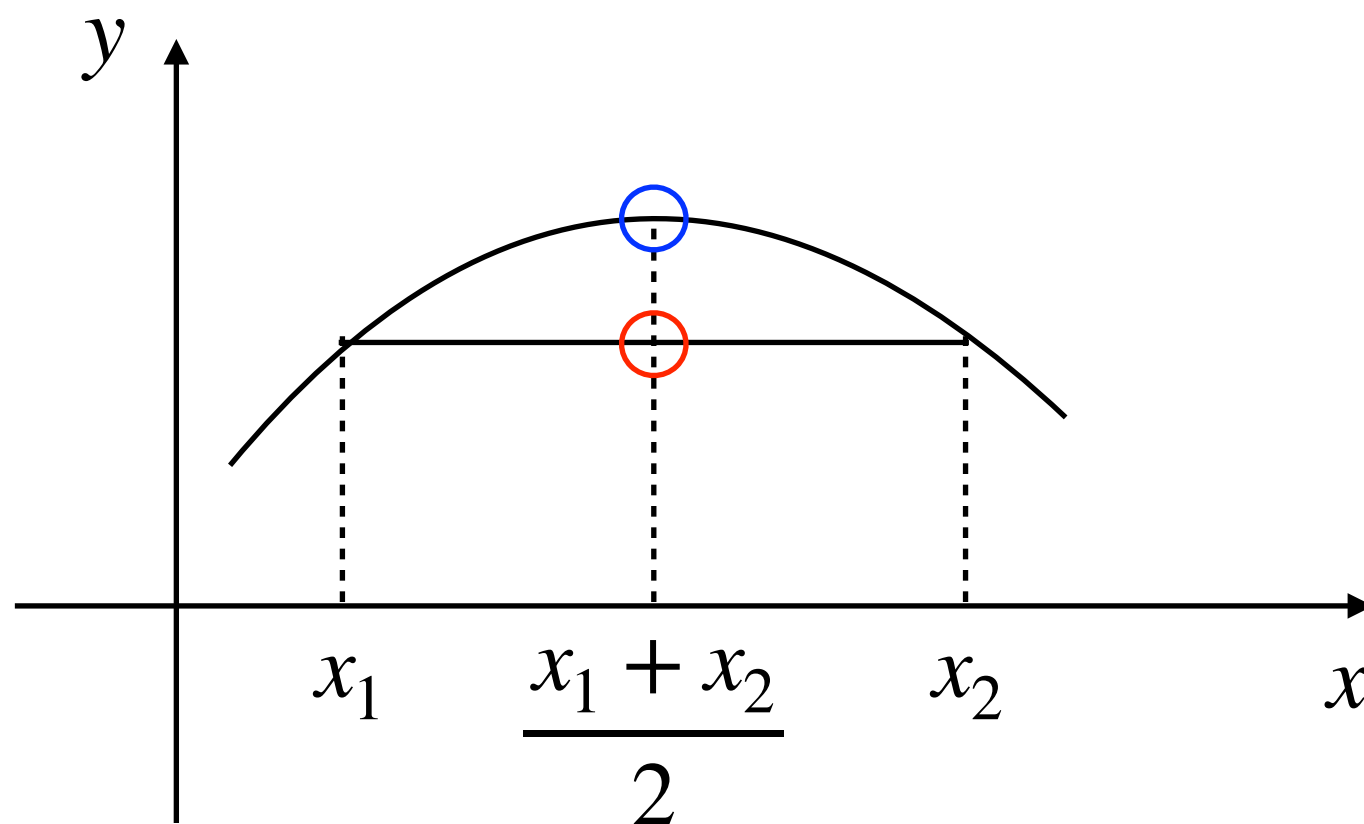
○	$\frac{f(x_1) + f(x_2)}{2}$
○	$f\left(\frac{x_1 + x_2}{2}\right)$

给定函数 $f: \mathbb{R} \rightarrow \mathbb{R}$ ，对于任意两个点 x_1 和 x_2 ，如果满足下列条件

$$f\left(\frac{x_1 + x_2}{2}\right) \leq \frac{f(x_1) + f(x_2)}{2}$$

则称 $f(x)$ 是一个凹函数。

凸函数



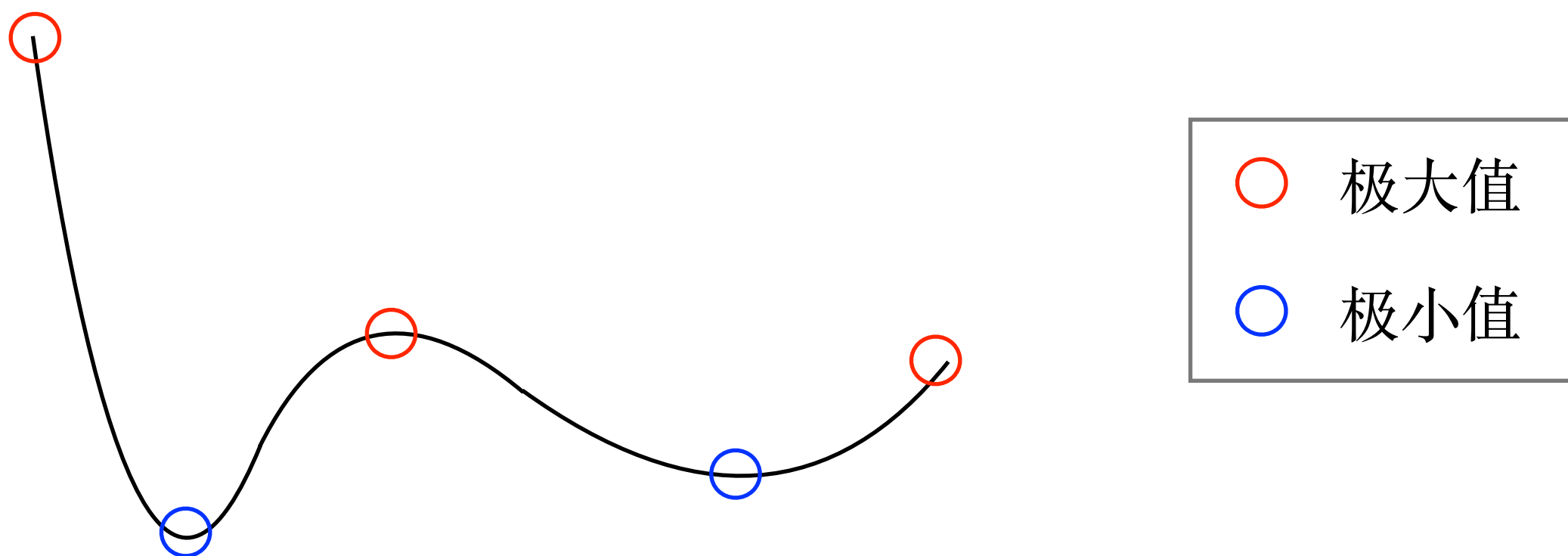
○	$\frac{f(x_1) + f(x_2)}{2}$
○	$f\left(\frac{x_1 + x_2}{2}\right)$

给定函数 $f: \mathbb{R} \rightarrow \mathbb{R}$ ，对于任意两个点 x_1 和 x_2 ，如果满足下列条件

$$f\left(\frac{x_1 + x_2}{2}\right) \geq \frac{f(x_1) + f(x_2)}{2}$$

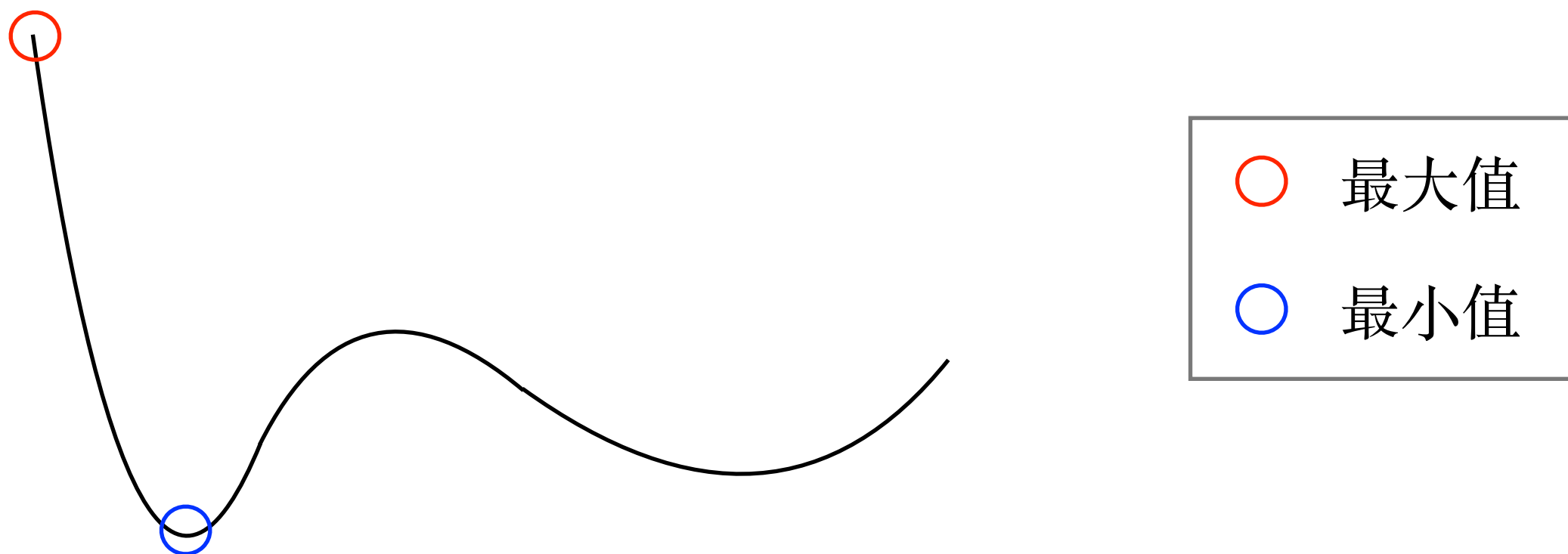
则称 $f(x)$ 是一个凸函数。

函数的极值



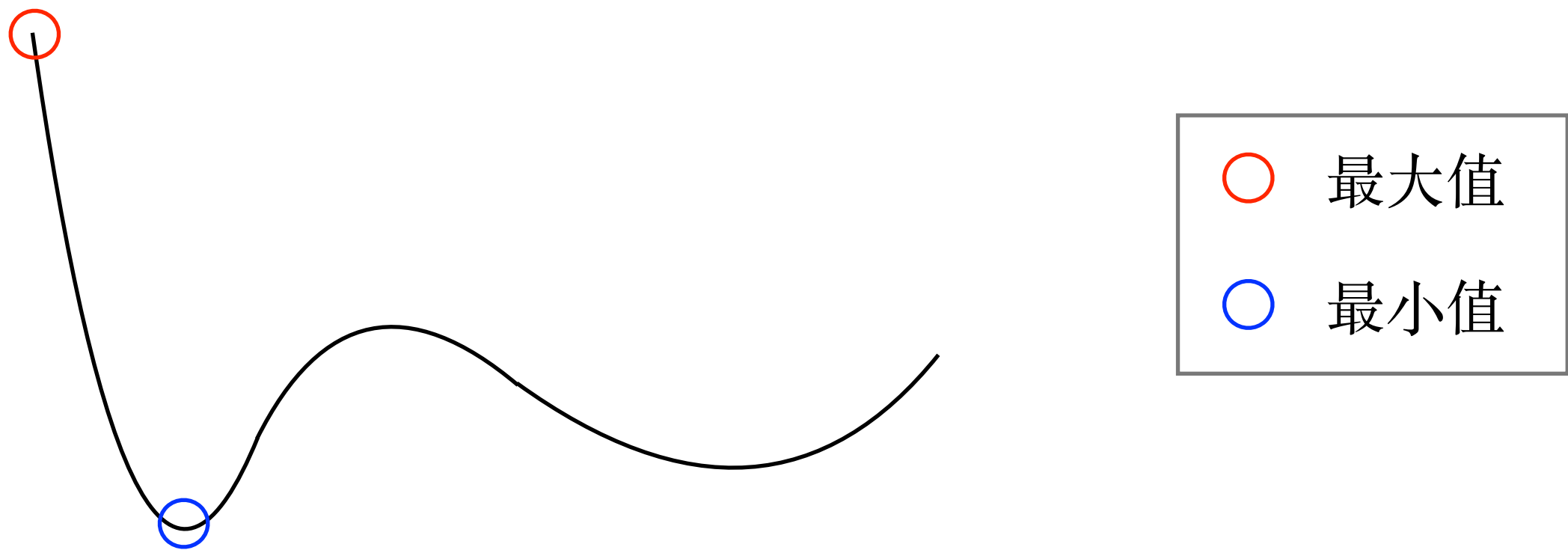
设函数 $f(x)$ 在点 $x = x_0$ 及其附近有定义。如果对于 x_0 附近的所有点都有 $f(x) < f(x_0)$ ，则 $f(x_0)$ 是函数 $f(x)$ 的一个极大值， x_0 是函数 $f(x)$ 的一个极大值点。如果对于 x_0 附近的所有点都有 $f(x) > f(x_0)$ ，则 $f(x_0)$ 是函数 $f(x)$ 的一个极小值， x_0 是函数 $f(x)$ 的一个极小值点。

函数的最值



函数在整个定义域内可能有许多极大值或极小值，而且某个极大值不一定大于某个极小值。函数 $f(x)$ 在整个定义域内的最小函数值 $f(x_0)$ 称为函数 $f(x)$ 的**最小值**， x_0 称为**最小值点**。类似地，函数 $f(x)$ 在整个定义域内的最大函数值 $f(x_0)$ 称为函数 $f(x)$ 的**最大值**， x_0 称为**最大值点**。

函数的最小值和最大值定理



如果函数 $f(x)$ 在闭区间 $[a, b]$ 上连续，则 $f(x)$ 在 $[a, b]$ 上必有最大值和最小值。在开区间 (a, b) 上连续的函数 $f(x)$ 不一定有最大值和最小值，如函数 $f(x) = 1/x$ 。函数的最值点必在函数的极值点或者区间的端点处获得。函数的极值可能有多个，但是最值最多只有一个。

计算函数的最值

如果函数 $f(x)$ 在闭区间 $[a, b]$ 上有定义，在开区间 (a, b) 内有导数，则求函数 $f(x)$ 在闭区间 $[a, b]$ 上的最大值和最小值的步骤如下：

- ① 求函数 $f(x)$ 在开区间 (a, b) 的导数 $f'(x)$ ；
- ② 求方程 $f'(x) = 0$ 在 (a, b) 内的解；
- ③ 求在 (a, b) 内使 $f'(x) = 0$ 的所有点的函数值和 $f(x)$ 在闭区间端点处的函数值 $f(a)$ 和 $f(b)$ ；
- ④ 比较上面所求的所有值，其中最大值为函数 $f(x)$ 在闭区间 $[a, b]$ 上的最大值，最小值为函数 $f(x)$ 在闭区间 $[a, b]$ 上的最小值。

例如，可以使用上述方法计算函数 $f(x) = x^2 - 2x + 1$ 在区间 $[-2, 2]$ 上的最大值和最小值，得到函数的最小值点是1，最大值点是-2。

不定积分

函数 $f(x)$ 的**不定积分**是一个导数等于 $f(x)$ 的函数 F ，即 $F'(x) = f(x)$ 。相应地，函数 $F(x)$ 称为 $f(x)$ 的**原函数**。一个函数通常有多个原函数。例如，函数 $f(x) = 2x$ 的原函数可以是 $F(x) = x^2 + 1$ ，也可以是 $F(x) = x^2 + 2$ 。因此，我们通常将原函数写成以下的形式：

$$\int f(x)dx = F(x) + C$$

其中， C 表示任意常数。常见的积分公式如下：

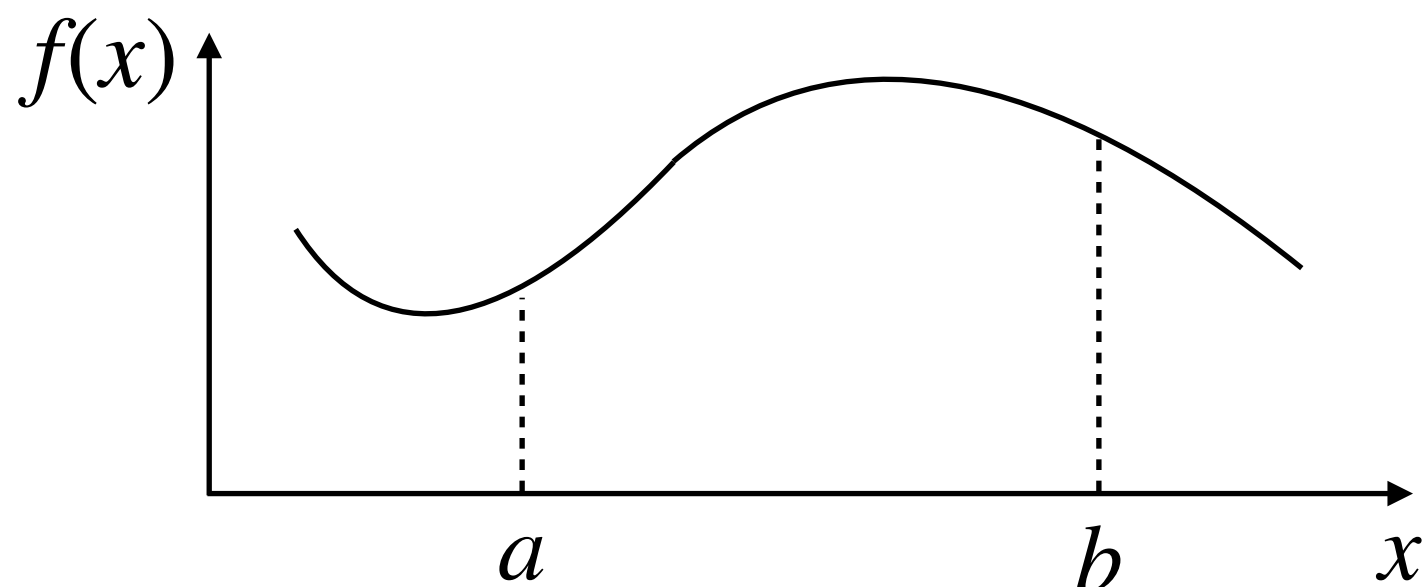
$$\int a dx = ax + C$$

$$\int x^a dx = x^{a+1}/(a+1) + C$$

$$\int e^x dx = e^x + C$$

$$\int \sin x dx = -\cos x + C$$

定积分

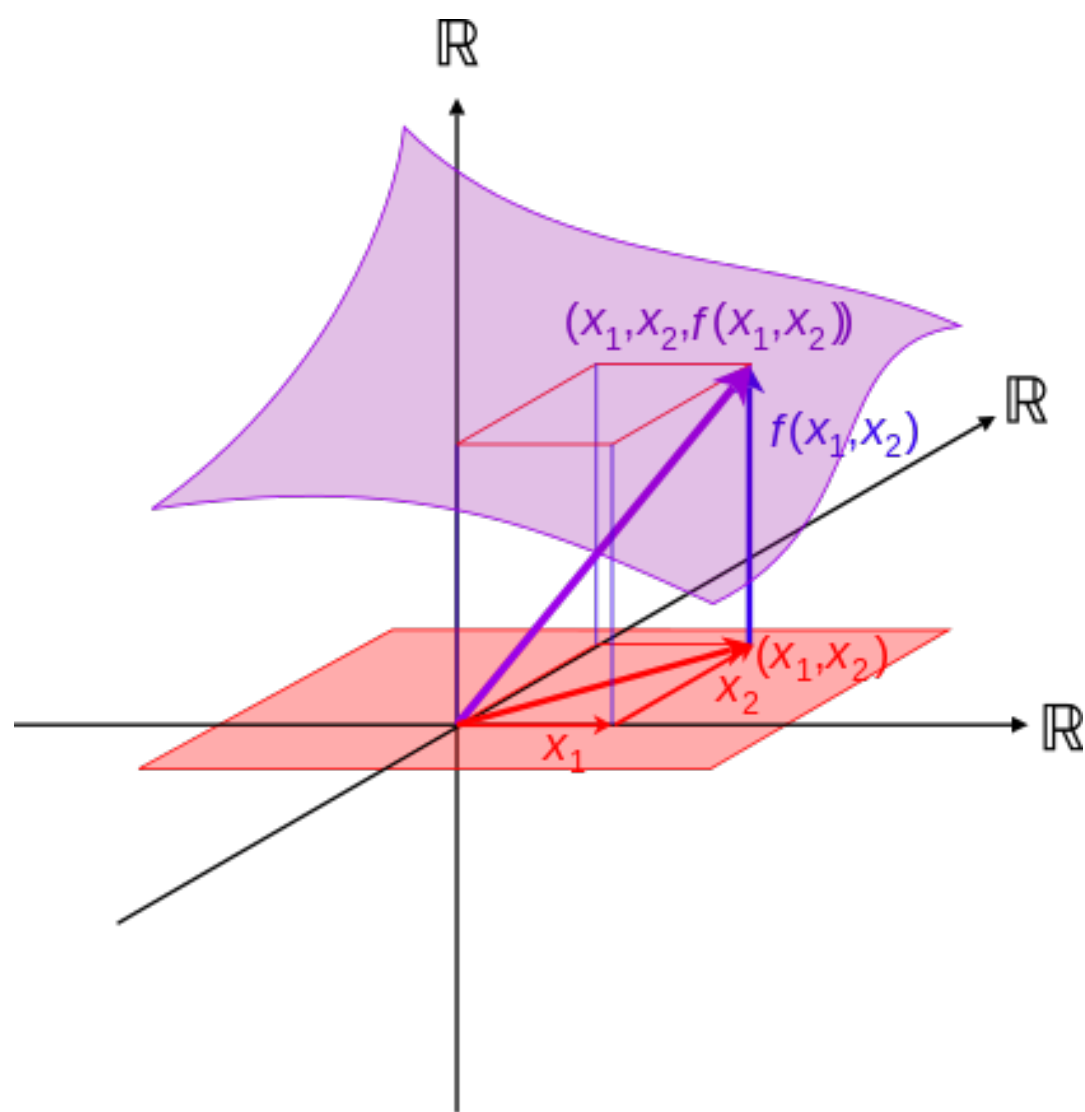


设函数 $f(x)$ 在区间 $[a, b]$ 上连续，将区间 $[a, b]$ 分成 n 个长度相等的子区间，则函数 $f(x)$ 在区间 $[a, b]$ 上的定积分定义为

$$\int_a^b f(x)dx = \lim_{n \rightarrow +\infty} \sum_{i=1}^n f\left(a + \frac{i}{n}(b-a)\right) \frac{b-a}{n}$$

其中， a 称为积分下限， b 称为积分上限， $[a, b]$ 称为积分区间， x 称为积分变量， $f(x)$ 称为被积函数。从直观上理解，定积分计算的是包围区域的面积。

多元函数



图片来源: https://en.wikipedia.org/wiki/Function_of_several_real_variables

设 D 是一个非空的 n 元有序数组的集合, f 为某一确定的对应法则, 如果对于每一个有限数组 $(x_1, x_2, \dots, x_n) \in D$, 通过对应法则 f , 都有唯一确定的实数 y 与之对应, 则称对应法则 f 为定义在 D 上的多元函数, 记为

$$y = f(x_1, x_2, \dots, x_n)$$

其中 x_1, x_2, \dots, x_n 称为自变量, y 称为因变量。

偏导数

设函数 $z = f(x, y)$ 在点 (x_0, y_0) 的某一邻域内有定义，当 y 固定在 y_0 而 x 在 x_0 处有增量 Δx 时，相应地函数值有增量 $f(x_0 + \Delta x, y_0) - f(x_0, y_0)$ 。如果极限

$$\lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}$$

存在，则称此极限为函数 $z = f(x, y)$ 在点 (x_0, y_0) 处对 x 的偏导数，记为

$$\left. \frac{\partial z}{\partial x} \right|_{x=x_0, y=y_0} = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}$$

另一种形式是 $f_x(x_0, y_0)$ 。同理可以定义函数在点 (x_0, y_0) 处对 y 的偏导数。如果函数 $z = f(x, y)$ 在区域 D 内任意一点 (x, y) 处对 x 的偏导数都存在，那么这个偏导数是 x 和 y 的函数，成为函数 $z = f(x, y)$ 对自变量 x 的偏导数，记为 $\partial z / \partial x$ 。

多元函数求导

设 $f(x, y) = x^2 + 3xy + y - 1$ ，求该函数对 x 和 y 的偏导在点 $(4, -5)$ 处的取值。求解方法如下。首先计算函数对 x 的偏导。在计算过程中，我们可以将 y 看作常量，然后对 x 求导：

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2 + 3xy + y - 1) = 2x + 3y$$

因此， $\partial f / \partial x$ 在 $(4, -5)$ 处的值为 $2 \times 4 + 3 \times (-5) = -7$ 。

接下来计算函数对 y 的偏导，将 x 看作常量：

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(x^2 + 3xy + y - 1) = 3x + 1$$

因此， $\partial f / \partial y$ 在 $(4, -5)$ 处的值为 $3 \times 4 + 1 = 13$ 。

多元复合函数求导

首先来考虑一元函数与多元函数复合的情况。若函数 $u = \phi(x)$ 和函数 $v = \psi(x)$ 都在点 x 可导，函数 $z = f(u, v)$ 在对应点 (u, v) 具有连续偏导数，那么复合函数 $z = f(\phi(x), \psi(x))$ 在点 x 可导，其导数为

$$\frac{dz}{dx} = \frac{\partial z}{\partial u} \frac{du}{dx} + \frac{\partial z}{\partial v} \frac{dv}{dx}$$

例如，令 $z = f(u, v) = u^2 - v^2$ ， $u = \phi(x) = x^2 - 1$ ， $v = \psi(x) = 3x + 2$ ，则复合函数 z 对 x 的导数可计算为

$$\begin{aligned} \frac{dz}{dx} &= \frac{\partial z}{\partial u} \frac{du}{dx} + \frac{\partial z}{\partial v} \frac{dv}{dx} \\ &= 2u \times 2x + (-2v) \times 3 \\ &= 4x^3 - 10x - 12 \end{aligned}$$

多元复合函数的求导

然后考虑多元函数与多元函数复合的情况。如果函数 $u = \phi(x, y)$ 与函数 $v = \psi(x, y)$ 具有对 x 和 y 的偏导数，函数 $z = f(u, v)$ 在对应点 (u, v) 具有连续偏导数，那么复合函数 $z = f(\phi(x, y), \psi(x, y))$ 在点 (x, y) 的两个偏导数存在：

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial x}$$

$$\frac{\partial z}{\partial y} = \frac{\partial z}{\partial u} \frac{\partial u}{\partial y} + \frac{\partial z}{\partial v} \frac{\partial v}{\partial y}$$

例如，令 $z = f(u, v) = u + v$ ， $u = \phi(x, y) = xy$ ， $v = \psi(x, y) = x + y$ ，则复合函数 z 对 x 和 y 的偏导数分别是

$$\frac{\partial z}{\partial x} = y + 1, \quad \frac{\partial z}{\partial y} = x + 1$$

梯度

设二元函数 $z = f(x, y)$ 在平面区域 D 上具有一阶连续偏导数，则对于每一个点 (x, y) 可以定义一个向量，称为函数 $z = f(x, y)$ 在点 (x, y) 的梯度，记作

$$\nabla f(x, y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

例如，令 $z = f(x, y) = x^2 - y^3$ ，则 x 和 y 的偏导函数为

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 3y^2$$

因此，函数 $f(x, y)$ 在点 $(2, 1)$ 处的梯度是一个二维向量 $(4, 3)$ 。多元函数的梯度可以类似地计算。梯度对于计算多元函数的极值而言非常重要，在深度学习的参数优化中被广泛使用。

多元函数极值

设函数 $z = f(x, y)$ 在点 (x_0, y_0) 的某个邻域内有定义，对于该邻域内异于 (x_0, y_0) 的点，如果不等式

$$f(x, y) < f(x_0, y_0)$$

成立，则称函数 $f(x, y)$ 在点 (x_0, y_0) 处有极大值。如果不等式

$$f(x, y) > f(x_0, y_0)$$

成立，则称函数 $f(x, y)$ 在点 (x_0, y_0) 处有极小值。

例如，函数 $z = 3x^2 + 4y^2$ 在点 $(0,0)$ 处有极小值，因为除了 $(0,0)$ 以外所有的点的函数值均为正，只有在点 $(0,0)$ 处的函数值为0。与之相反，函数 $z = -\sqrt{x^2 + y^2}$ 在点 $(0,0)$ 处有极大值，因为除了 $(0,0)$ 以外所有的点的函数值均为负，只有在点 $(0,0)$ 处的函数值为0。

多元函数极值条件

定理1（必要条件）： 设函数 $z = f(x, y)$ 在点 (x_0, y_0) 处具有偏导数，且在点 (x_0, y_0) 处有极值，则函数在该点的偏导数必然为0：

$$f_x(x_0, y_0) = 0, \quad f_y(x_0, y_0) = 0$$

定理2（充分条件）： 设函数 $z = f(x, y)$ 在点 (x_0, y_0) 的某邻域内连续且有一阶及二阶连续偏导数，并且 $f_x(x_0, y_0) = 0, f_y(x_0, y_0) = 0$ ，令

$$f_{xx}(x_0, y_0) = A, \quad f_{xy}(x_0, y_0) = B, \quad f_{yy}(x_0, y_0) = C$$

则 $f(x, y)$ 在 (x_0, y_0) 处是否取得极值的条件如下：

- ① 当 $AC - B^2 > 0$ 时有极值，当 $A < 0$ 时有极大值， $A > 0$ 时有极小值。
- ② 当 $AC - B^2 < 0$ 时没有极值。
- ③ 当 $AC - B^2 = 0$ 时可能有极值，也可能没有极值。

求多元函数极值

求二元函数 $f(x, y) = x^3 - y^3 + 3x^2 + 3y^2 - 9x$ 的极值。

首先求解一阶导数组成的方程组：

$$f_x(x, y) = 3x^2 + 6x - 9 = 0$$

$$f_y(x, y) = -3y^2 + 6y = 0$$

得到四组解：(1, 0)、(1, 2)、(-3, 0)和(-3, 2)。它们不一定是极值点，需要进一步考察二阶导数：

$$f_{xx}(x, y) = 6x + 6$$

$$f_{xy}(x, y) = 0$$

$$f_{yy}(x, y) = -6y + 6$$

求多元函数极值

对四个解分别计算 A 、 B 和 C ，考察定理2的条件。

- ① $(1, 0)$: $AC - B^2 = 12 \times 6 > 0$ 且 $A = 12 > 0$ ，因此 $(1, 0)$ 是函数 $f(x, y)$ 的一个极小值点，对应的极小值是 $f(1, 0) = -5$ 。
- ② $(1, 2)$: $AC - B^2 = 12 \times (-6) < 0$ ，因此 $(1, 2)$ 不是函数 $f(x, y)$ 的极值点。
- ③ $(-3, 0)$: $AC - B^2 = (-12) \times 6 < 0$ ，因此 $(-3, 0)$ 不是函数 $f(x, y)$ 的极值点。
- ④ $(-3, 2)$: $AC - B^2 = (-12) \times (-6) > 0$ 且 $A = -12 < 0$ ，因此 $(-3, 2)$ 是函数 $f(x, y)$ 的一个极大值点，对应的极大值是 $f(-3, 2) = -31$ 。

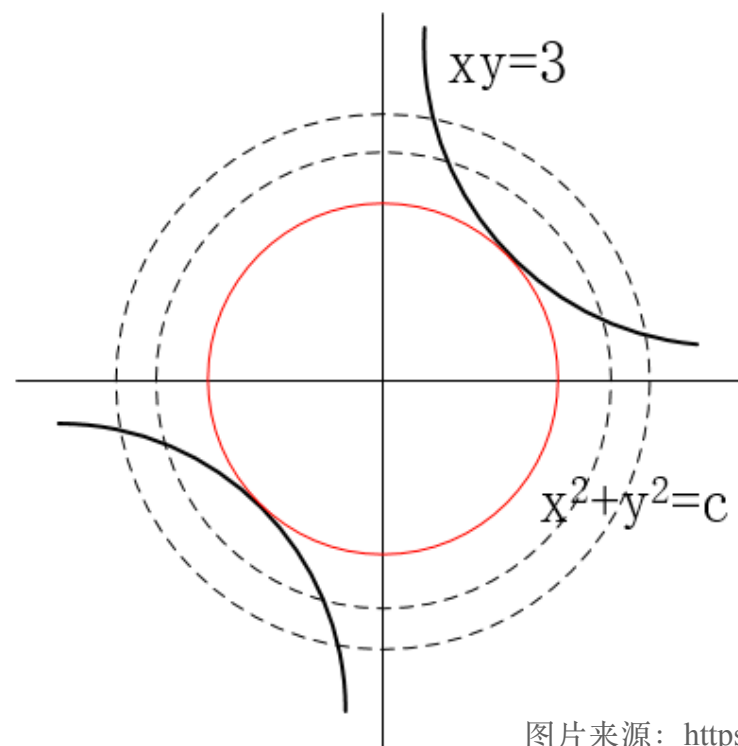
拉格朗日乘子法

求函数 $z = f(x, y)$ 在满足 $g(x, y) = 0$ 下的条件极值，可以转化为函数

$$F(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

的无约束条件极值问题。

例如，给定双曲线 $xy = 3$ 求该曲线上距离原点最近的点。这是一个典型的带约束的求极值问题。



图片来源: <https://www.cnblogs.com/maybe2030/p/4946256.html>

拉格朗日乘子法

原始问题可以转化为

$$F(x, y, \lambda) = x^2 + y^2 + \lambda(xy - 3)$$

计算函数 $F(x, y, \lambda)$ 的一阶偏导，得到方程组：

$$F_x(x, y, \lambda) = 2x + \lambda y = 0$$

$$F_y(x, y, \lambda) = 2y + \lambda x = 0$$

$$F_\lambda(x, y, \lambda) = xy - 3 = 0$$

求解该方程组，可以得到 $\lambda = 2$ 或 $\lambda = -2$ 。当 $\lambda = 2$ 时，无法求解 x 和 y ，因为势必有 $-x^2 = 3$ 。当 $\lambda = -2$ 时，有两组解： $(\sqrt{3}, \sqrt{3})$ 和 $(-\sqrt{3}, -\sqrt{3})$ 。

拉格朗日乘子法在自然语言处理中具有非常广泛的应用，必须熟练掌握。

内容提要

微积分

概率论

线性代数

信息论

随机试验

具备以下三个特点的试验称为随机试验：

- ① 可以在相同的条件下重复地运行；
- ② 每次试验的可能结果可能不止一个，并且能事先明确试验的所有可能结果；
- ③ 进行一次试验之前不能确定哪一个结果会出现。

以下是一些随机试验的例子：

- ① 抛一枚硬币，观察正面 H 、反面 T 出现的情况。
- ② 抛一颗骰子，观察出现的点数。
- ③ 在一批灯泡里任意抽取一只，测试它的寿命。

样本空间

对于随机试验，尽管在每次试验之前不能预知试验的结果，但试验的所有可能结果组成的集合是已知的。我们将随机试验 E 的所有可能结果组成的集合称为 E 的**样本空间**，记为 S 。样本空间中的元素，称为**样本点**。

例如，给定以下随机试验

- ① E_1 ：抛一枚硬币，观察正面 H 、反面 T 出现的情况。
- ② E_2 ：抛一颗骰子，观察出现的点数。
- ③ E_3 ：在一批灯泡里任意抽取一只，测试它的寿命。

对应的样本空间是：

- ① $S_1 : \{H, T\}$
- ② $S_2 : \{1, 2, 3, 4, 5, 6\}$
- ③ $S_3 : \{t | t \geq 0\}$

随机事件

试验 E 的样本空间 S 的子集称为 E 的随机事件，简称为事件。

例如，令“将一枚硬币抛掷两次，观察正面 H 、反面 T 出现的情况”是一个随机试验 E ，则其样本空间总共包含四个元素：

$$S = \{HH, HT, TT, TH\}$$

我们可以定义一个事件“第一次出现的是 H ”，即

$$A_1 = \{HH, HT\}$$

还可以定义另一个事件“两次出现的是同一面”，即

$$A_2 = \{HH, TT\}$$

显然， A_1 和 A_2 都是样本空间的子集。

概率

设 E 是随机试验， S 是样本空间。对于 E 的每一个事件 A 赋予一个实数，记为 $P(A)$ ，称为事件 A 的**概率**。概率必须满足以下条件：

- ① 非负性：对于每一个事件 A ，有 $P(A) \geq 0$ ；
- ② 规范性：对于必然发生的事件 S ，有 $P(S) = 1$ ；
- ③ 可列可加性：设 A_1, A_2, \dots 是两两互不相容的事件，即对于 $A_i \cap A_j = \emptyset$ ($i \neq j$)，有 $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$ 。

令 A 和 B 为任意两个事件， AB 表示两个事件同时发生，以下公式成立：

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

对于前面抛掷两次硬币的例子，如果 A 表示“第一次是 H ”， B 表示“两次结果都一样”，那么 AB 表示“两次都是 H ”。

等可能概型

等可能概型是指符合以下两个条件的随机试验：

- ① 试验的样本空间只能包含有限个元素；
- ② 试验中每个基本事件（即每个结果）发生的可能性基本相同。

例如，一个口袋里装有6只球，其中有4只白球和2只红球。从袋中取球两次，每次随机地取一只，假设每只球都有相等概率被抽中。第一次取一球不放回袋中，第二次从剩余的球中再取一球。计算：（1）取到的两只球都是白球的概率，（2）取到的两只球至少有一只是白球的概率。

首先计算两只球都是白球的概率： $(4/6) \times (3/5) = 2/5$ 。然后，先计算两只球都是红球的概率： $(2/6) \times (1/5) = 1/15$ ，然后可以得到取到的两只球至少有一只是白球的概率： $1 - (1/15) = 14/15$ 。

条件概率

设 A 和 B 是两个事件，且 $P(A) > 0$ ，称

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

为在事件 A 发生的条件下事件 B 发生的条件概率。

不难验证，条件概率符合概率定义中的三个条件：

- ① 非负性：对于每个事件 B ，有 $P(B|A) \geq 0$ ；
- ② 规范性：对于必然事件 S ，有 $P(S|A) = 1$ ；
- ③ 可列可加性：设 B_1 、 B_2 、...是两两不相容的事件，则有

$$P\left(\bigcup_{i=1}^{\infty} B_i | A\right) = \sum_{i=1}^{\infty} P(B_i | A)$$

条件概率

例如，一个口袋里装有6只球，其中有4只白球和2只红球。从袋中取球两次，每次随机地取一只，假设每只球都有相等的概率被抽中。第一次取一球不放回袋中，第二次从剩余的球中再取一球。设事件 A 为“第一次取到白球”，事件 B 为“第二次取到白球”，计算条件概率 $P(B|A)$ 。

首先计算 $P(A)$ 。由于开始口袋中有6只球，其中有4只白球，因此第一次取到白球的概率 $P(A) = 4/6$ 。然后计算 $P(AB)$ ，即事件“两次都抽到白球”的概率：

$$P(AB) = \frac{4}{6} \times \frac{3}{5} = \frac{2}{5}$$

因此，条件概率计算如下：

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{2}{5} \times \frac{6}{4} = \frac{3}{5}$$

全概率公式

设 S 为试验 E 的样本空间， B_1, B_2, \dots, B_n 为事件 E 的一组事件，如果以下两个条件成立

$$\textcircled{1} \quad B_i \cap B_j = \emptyset, i \neq j, i, j = 1, \dots, n,$$

$$\textcircled{2} \quad B_1 \cup B_2 \cup \dots \cup B_n = S,$$

则称 B_1, B_2, \dots, B_n 为样本空间 S 的一个划分。

例如，试验 E “掷一颗骰子观察其点数”样本空间为 $S = \{1, 2, 3, 4, 5, 6\}$ ，则 $B_1 = \{1, 2, 3\}$ ， $B_2 = \{4, 5\}$ 和 $B_3 = \{6\}$ 是 S 的一个划分。

设 A 是试验 E 的一个事件， B_1, B_2, \dots, B_n 是其样本空间的一个划分，则以下全概率公式成立：

$$P(A) = \sum_{i=1}^n P(A | B_i) P(B_i)$$

贝叶斯公式

设 A 和 B 是随机试验 E 的任意两个事件，以下贝叶斯公式成立：

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

可以进一步与全概率公式结合起来。令 B_1, B_2, \dots, B_n 是 S 的一个划分，而且 $P(B_i) > 0$ ($i = 1, 2, \dots, n$)，则有

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^n P(A|B_j)P(B_j)}$$

贝叶斯公式在人工智能中非常重要，产生了重要的贝叶斯学派。贝叶斯公式对于揭示信息认知加工过程与规律、实现有效的学习和判断决策都具有十分重要的理论意义和实践价值。

独立性

设 A 和 B 是两个随机事件，如果满足等式

$$P(AB) = P(A)P(B)$$

则称事件 A 和 B 相互独立。

两个事件相互独立的含义是其中一个事件已发生，不影响另一个事件发生的概率。在实际应用中，对于事件的独立性通常是根据事件的实际意义去判断。如果根据实际情况分析，两个事件之间没有关联或者关联很弱，那么就认为它们之间是相互独立的。例如，如果甲、乙两人同一天感冒，甲在中国，乙在美国，双方并未接触，则可以认为两个事件是独立的。如果甲、乙是住在同一个宿舍的舍友，那么就不能认为是相互独立的。

在实际应用中，为了简化概率模型，通常会做很多独立性假设。

随机变量

将一枚硬币抛掷两次，观察出现正面 H 和反面 T 的情况，样本空间是

$$S = \{HH, HT, TT, TH\}$$

以 X 表示两次投掷得到正面 H 的总数，则 X 的取值是一个随机变量：

- ① $X = 0$ ：当投掷结果是 $\{TT\}$ 时；
- ② $X = 1$ ：当投掷结果是 $\{HT\}$ 或 $\{TH\}$ 时；
- ③ $X = 2$ ：当投掷结果是 $\{HH\}$ 时。

随机变量的取值随试验的结果而定，在试验之前不能预知取什么值，并且其取值有一定的概率。随机变量的引入，使我们能够描述各种随机现象，并能利用数学方法对随机试验的结果进行深入分析。

离散型随机变量

取值是有限个或可列举无限个的随机变量称为离散型随机变量。例如，抛掷一枚硬币，只可能取正面和反面两个取值，因此是离散型随机变量。

设离散型随机变量 X 可能的取值为 x_k ($k = 1, 2, \dots$)， X 取各个可能值的概率，即事件 $\{X = x_k\}$ 的概率，为

$$P(X = x_k) = p_k, k = 1, 2, \dots$$

上式称为离散型随机变量 X 的分布律。

注意，根据概率的定义， p_k 满足以下两个条件：

① $p_k \geq 0, k = 1, 2, \dots;$

② $\sum_{k=1}^{\infty} p_k = 1。$

离散型随机变量分布

以下两种离散型随机变量经常被使用。

第一个是(0 – 1)分布。设随机变量 X 只能取0和1两个值，其分布律为

$$P(X = k) = p^k(1 - p)^{1-k}$$

其中， k 的取值是0或1， $0 < p < 1$ 。

第二个是二项分布。设 n 是一个正整数， k 是一个不大于 n 的非负整数，即 $0 \leq k \leq n$ ，某个随机事件 A 发生的概率为 p ，则在 n 次试验中事件 A 发生 k 次的概率为

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

显然，当 $n = 1$ 时，二项分布等价于(0 – 1)分布。

随机变量的分布函数

对于非离散型随机变量，其取值不能一一列举，因此需要采用新的形式对离散型和非离散型随机变量进行统一描述。

设 X 是一个随机变量， x 是任意实数，函数

$$F(x) = P(X \leq x)$$

称为 X 的分布函数。

对于任意两个实数 x_1 和 x_2 且满足 $x_1 < x_2$ ，均有

$$\begin{aligned} P(x_1 < X \leq x_2) &= P(X \leq x_2) - P(X \leq x_1) \\ &= F(x_2) - F(x_1) \end{aligned}$$

因此，如果已知 X 的分布函数，我们就知道 X 落在任意区间 $(x_1, x_2]$ 的概率。从这个意义上说，分布函数完整地描述了随机变量的统计规律性。

分布律与分布函数

X	-1	2	3
p_k	0.25	0.50	0.25

给定上表所示的分布律，相应的分布函数定义如下：

$$F(x) = \begin{cases} 0.00 & x < -1 \\ 0.25 & -1 \leq x < 2 \\ 0.75 & 2 \leq x < 3 \\ 1.00 & x \geq 3 \end{cases}$$

由此可见，分布函数可以全面地描述离散型随机变量。

连续型随机变量

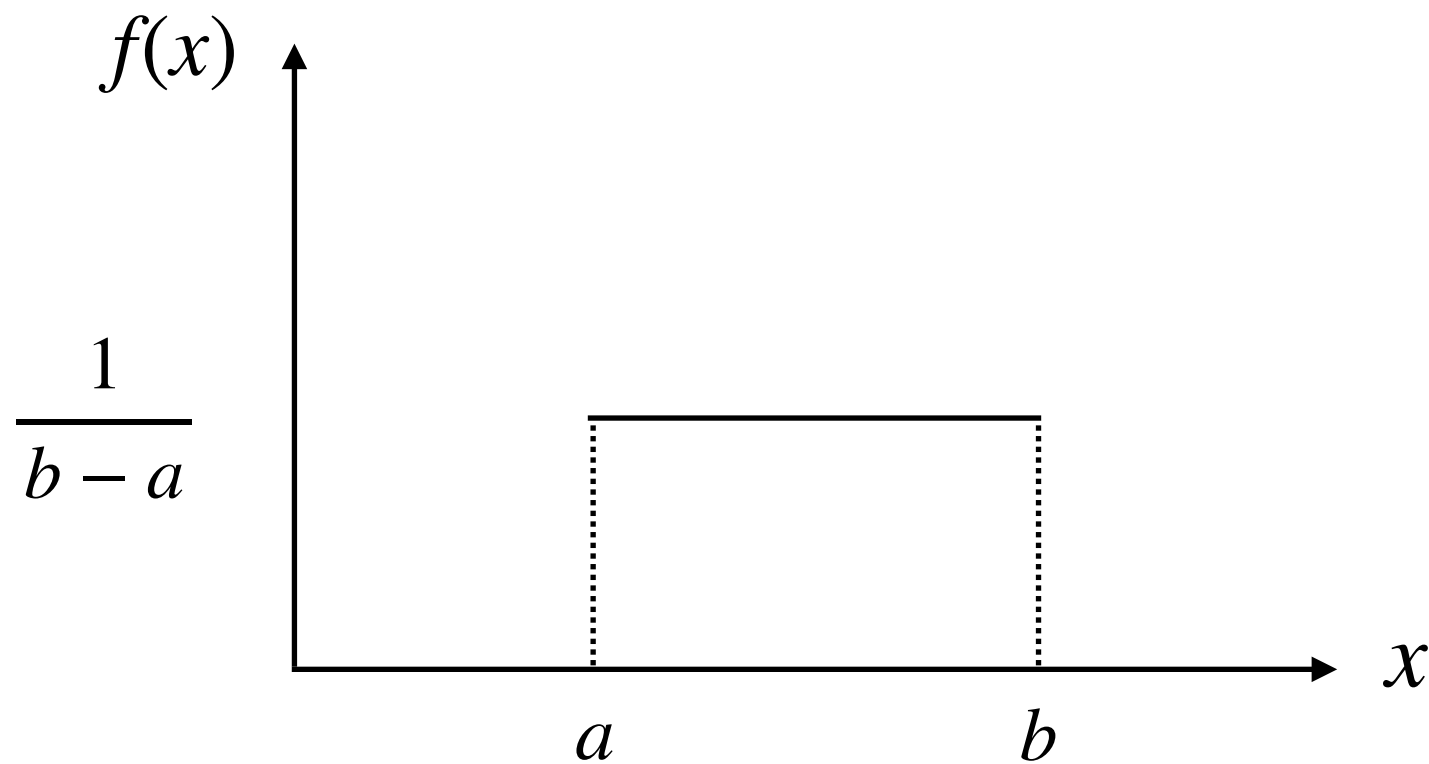
如果对于随机变量 X 的分布函数 $F(x)$ ，存在非负函数 $f(x)$ ，使对于任意实数 x 有

$$F(x) = \int_{-\infty}^x f(t)dt$$

则称 X 为连续型随机变量。 $f(x)$ 称为 X 的概率密度函数，具有以下性质：

- ① $f(x) \geq 0$;
- ② $\int_{-\infty}^{\infty} f(x)dx = 1$;
- ③ 对于任意实数 x_1 和 x_2 ($x_1 \leq x_2$) , $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$;
- ④ 若 $f(x)$ 在点 x 处连续，则有 $F'(x) = f(x)$ 。

均匀分布

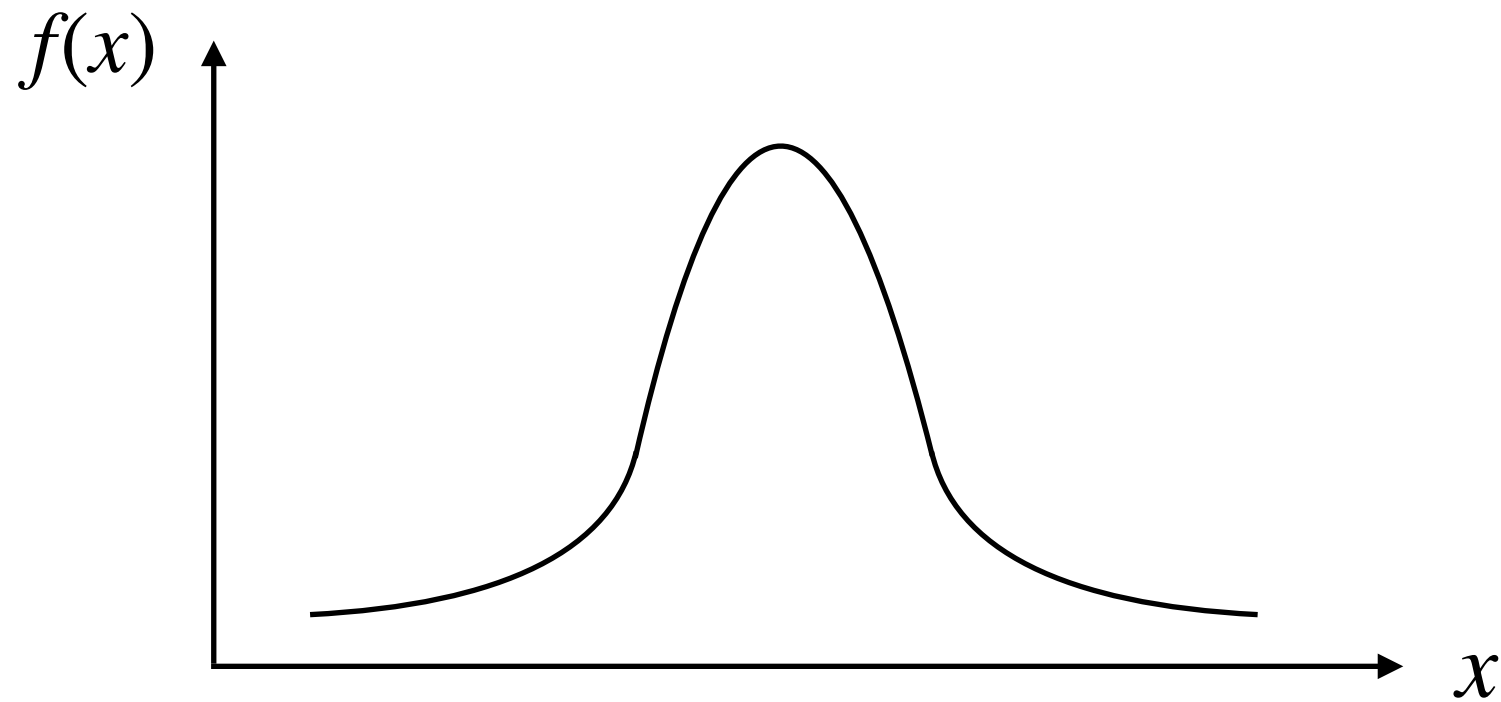


若连续型随机变量 X 具有概率密度

$$f(x) = \begin{cases} 1/(b-a) & \text{如果 } a < x < b \\ 0 & \text{否则} \end{cases}$$

则称 X 在区间 (a, b) 上服从均匀分布，记为 $X \sim U(a, b)$ 。

正态分布



若连续型随机变量 X 具有概率密度

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

其中 μ 和 σ 实常数且 $\sigma > 0$ ，则称 X 服从参数为 μ 和 σ 的**正态分布**或**高斯分布**，记作 $X \sim N(\mu, \sigma^2)$ 。

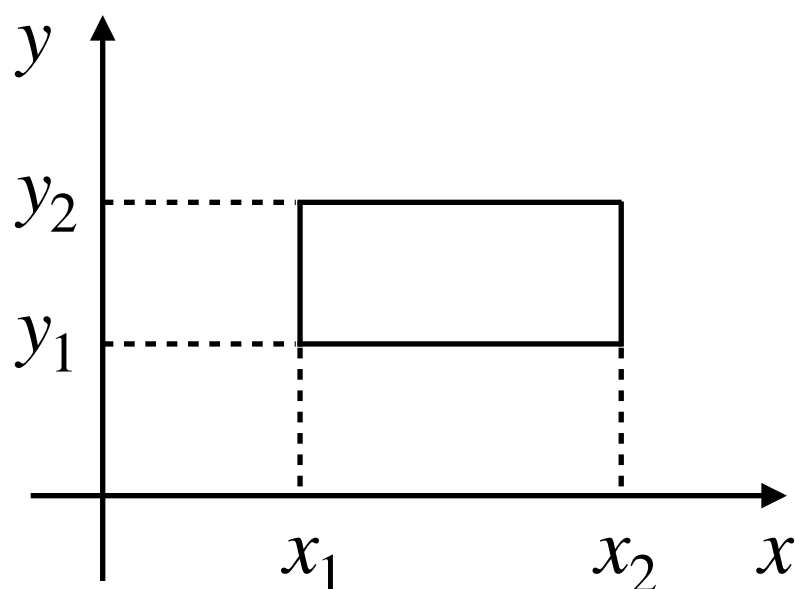
二维随机变量

之前只限于讨论单个随机变量的情况，实际问题中经常出现多个随机变量的情况。例如，为了研究某一地区某一年龄段儿童的发育情况，需要统计儿童的身高和体重。

设 (X, Y) 是二维随机变量，对于任意实数 x 和 y ，二元函数

$$F(x, y) = P(X \leq x, Y \leq y)$$

称为二维随机变量 (X, Y) 的分布函数，或随机变量 X 和 Y 的联合分布函数。



$$\begin{aligned} &P(x_1 < X \leq x_2, y_1 < Y \leq y_2) \\ &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \end{aligned}$$

二维离散型随机变量

如果二维随机变量 (X, Y) 全部可能的取值是有限对或可列无限多对，则称 (X, Y) 是离散型的随机变量。设 (X, Y) 所有的可能取值为 (x_i, y_j) ， $i, j = 1, 2, \dots$ ，则 X 和 Y 的联合分布律定义为

$$P(X = x, Y = y) = p_{ij}$$

联合分布律通常使用表格的方式来表示：

$Y \backslash X$	x_1	x_2	\dots	x_i	\dots
y_1	p_{11}	p_{21}	\dots	p_{i1}	\dots
y_2	p_{12}	p_{22}	\dots	p_{i2}	\dots
\vdots	\vdots	\vdots		\vdots	
y_j	p_{1j}	p_{2j}	\dots	p_{ij}	\dots
\vdots	\vdots	\vdots		\vdots	

二维连续型随机变量

对于二维随机变量 (X, Y) 的分布函数 $F(x, y)$ ，如果存在非负的函数 $f(x, y)$ 使得对于任意 x 和 y 都有

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(u, v) du dv$$

则称 (X, Y) 是连续型的二维随机变量，函数 $f(x, y)$ 称为二维随机变量 (X, Y) 的概率密度，或成为随机变量 X 和 Y 的联合概率密度。

例如，给定概率密度

$$f(x, y) = \begin{cases} 2e^{-(2x+y)} & \text{如果 } x > 0, y > 0 \\ 0 & \text{否则} \end{cases}$$

可计算分布函数为 $F(x, y) = (1 - e^{-2x})(1 - e^{-y})$ ，当 $x > 0$ 且 $y > 0$ 时。

边缘分布律

二维随机变量 (X, Y) 作为一个整体，具有分布函数 $F(x, y)$ ，而 X 和 Y 都是随机变量，各自也有分布函数，分别记为 $F_X(x)$ 和 $F_Y(y)$ ，分别称为二维随机变量 (X, Y) 关于 X 和关于 Y 的[边缘分布函数](#)，定义如下：

$$F_X(x) = P(X \leq x, Y < \infty) = F(x, \infty)$$

$$F_Y(y) = P(X < \infty, Y \leq y) = F(\infty, y)$$

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

上述式子也称为二维离散型随机变量 (X, Y) 关于 X 和 Y 的[边缘分布律](#)。

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	
y_2	0.0	0.2	0.0	0.0	
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$					

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2				

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2			

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5		

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5	0.1	

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5	0.1	1.0

随机变量 X 和 Y 的分布律分别定义为

$$P(X = x_i) = \sum_{j=1}^{\infty} p_{ij}$$

$$P(Y = y_j) = \sum_{i=1}^{\infty} p_{ij}$$

边缘概率密度

对于连续型随机变量 (X, Y) ，设其概率密度为 $f(x, y)$ ，由于

$$F_X(x) = F(x, \infty) = \int_{-\infty}^x \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx$$

由此可知 X 是一个连续型随机变量，而且其概率密度函数为

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

同样， Y 也是一个连续型随机变量，其概率密度为

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

$f_X(x)$ 和 $f_Y(y)$ 分别是关于 X 和关于 Y 的**边缘概率密度**。

条件分布律

下面来考虑事件 $\{Y = y_j\}$ 在已发生的条件下事件 $\{X = x_i\}$ 发生的概率，也就是求事件 $\{X = x_i | Y = y_j\}$ 的概率。

设 (X, Y) 是二维离散型随机变量，对于固定的 j ，若 $P(Y = y_j) > 0$ ，则称

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

为在 $Y = y_j$ 条件下随机变量 X 的**条件分布律**。

类似地，对于固定的 i ，若 $P(X = x_i) > 0$ ，则称

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

为在 $X = x_i$ 条件下随机变量 Y 的**条件分布律**。

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	
y_2	0.0	0.2	0.0	0.0	
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	
$P(X = x_i)$					

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$					

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2				

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2			

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5		

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5	0.1	

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件分布律

$Y \backslash X$	x_1	x_2	x_3	x_4	$P(Y = y_j)$
y_1	0.1	0.0	0.3	0.0	0.4
y_2	0.0	0.2	0.0	0.0	0.2
y_3	0.1	0.0	0.2	0.1	0.4
$P(X = x_i)$	0.2	0.2	0.5	0.1	1.0

$$P(Y = y_1 | X = x_1) = \frac{P(X = x_1, Y = y_1)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

$$P(Y = y_2 | X = x_1) = \frac{P(X = x_1, Y = y_2)}{P(X = x_1)} = \frac{0.0}{0.2} = 0.0$$

$$P(Y = y_3 | X = x_1) = \frac{P(X = x_1, Y = y_3)}{P(X = x_1)} = \frac{0.1}{0.2} = 0.5$$

条件概率密度

设二维随机变量 (X, Y) 的概率密度为 $f(x, y)$ ， (X, Y) 关于 Y 的边缘概率密度为 $f_Y(y)$ 。若对于固定的 y ， $f_Y(y) > 0$ ，则在 $Y = y$ 条件下 X 的**条件概率密度**定义为：

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

与之对应地，在 $Y = y$ 条件下 X 的**条件分布函数**定义为：

$$F_{X|Y}(x|y) = \int_{-\infty}^x \frac{f(x, y)}{f_Y(y)} dx$$

类似地，我们也可以定义在 $X = x$ 条件下 Y 的条件概率密度和条件分布函数。

相互独立的随机变量

设 $F(x, y)$ 、 $F_X(x)$ 和 $F_Y(y)$ 分别是二维随机变量 (X, Y) 的分布函数及边缘概率分布，如果对于所有的 x 和 y 有

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

即

$$F(x, y) = F_X(x)F_Y(y)$$

则称随机变量 X 和 Y **相互独立**。

当 X 和 Y 是离散型随机变量时， X 和 Y 相互独立的条件是

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

当 X 和 Y 是连续型随机变量时， X 和 Y 相互独立的条件是

$$f(x, y) = f_X(x)f_Y(y)$$

数学期望

设离散型随机变量 X 的分布律为 $P(X = x_k) = p_k$ ($k \geq 1$)，其数学期望定义为：

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} x_k p_k$$

类似地，设连续型变量 X 的概率密度为 $f(x)$ ，其数学期望定义为：

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

例如，假定 $P(X = 0) = 0.3$ ， $P(X = 1) = 0.5$ ， $P(X = 2) = 0.2$ ，则 X 的数学期望计算如下：

$$\mathbb{E}(X) = 0 \times 0.3 + 1 \times 0.5 + 2 \times 0.2 = 0.9$$

随机变量函数的数学期望

设 Y 是随机变量 X 的连续函数，即 $Y = g(X)$ 。如果 X 是离散型随机变量，其分布律为 $P(X = x_k) = p_k$ ($k \geq 1$)，则 Y 的数学期望定义为：

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{k=1}^{\infty} g(x_k)p_k$$

如果 X 是连续型随机变量，其概率密度为 $f(x)$ ，则 Y 的数学期望定义为：

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

随机变量函数的数学期望在深度学习中有着十分广泛的应用，尤其是在估计概率模型参数方面，是必须熟练掌握的重要概念。

数学期望的性质

数学期望有以下重要性质：

- ① 设 C 为实常数，则有 $\mathbb{E}(C) = C$ 。
- ② 设 X 是一个随机变量， C 是常数，则有 $\mathbb{E}(CX) = C\mathbb{E}(X)$ 。
- ③ 设 X 和 Y 是两个随机变量，则有 $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ 。这一性质可以推广到任意有限个随机变量之和的情况。
- ④ 设 X 和 Y 是两个相互独立的随机变量，则有 $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ 。这一性质可以推广到任意有限个相互独立的随机变量之积的情况。

方差

方差用于度量随机变量与其均值的偏离程度。设 X 是一个随机变量， X 的方差定义为：

$$D(X) = \text{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

我们通常将 $\sqrt{D(X)}$ 记为 $\sigma(X)$ ，称为标准差或者均方差。

对于离散型随机变量，方差计算公式为

$$D(X) = \sum_{k=1}^{\infty} (x_k - \mathbb{E}(X))^2 p_k$$

对于连续型随机变量，方差计算公式为

$$D(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx$$

方差

X	-1	2	5
p_k	0.10	0.70	0.20

$$\mathbb{E}(X) = 0.1 \times (-1) + 0.7 \times 2 + 0.2 \times 5 = 2.3$$

$$D(X) = 0.1 \times (-1 - 2.3)^2 + 0.7 \times (2 - 2.3)^2 + 0.3 \times (5 - 2.3)^2 = 2.61$$

X	-1	2	5
p_k	0.30	0.40	0.30

$$\mathbb{E}(X) = 0.3 \times (-1) + 0.4 \times 2 + 0.3 \times 5 = 2.0$$

$$D(X) = 0.3 \times (-1 - 2.0)^2 + 0.4 \times (2 - 2.0)^2 + 0.3 \times (5 - 2.0)^2 = 5.40$$

内容提要

微积分

概率论

线性代数

信息论

向量

n 个有次序的数 a_1, a_2, \dots, a_n 所组成的数组称为 n 维向量。这 n 个数称为该向量的 n 个分量，第 i 个数 a_i 称为第 i 个分量。向量通常表示为

$$\mathbf{a} = (a_1, a_2, \dots, a_n)$$

向量的模也称为向量的大小，定义如下：

$$||\mathbf{a}|| = \sqrt{a_1^2 + \dots + a_n^2}$$

给定两个 n 维向量 $\mathbf{a} = (a_1, \dots, a_n)$ 和 $\mathbf{b} = (b_1, \dots, b_n)$ ，主要运算公式如下：

- ① 加法： $\mathbf{a} + \mathbf{b} = (a_1 + b_1, \dots, a_n + b_n)$ 。
- ② 与数的乘法：设 λ 是一个实数，则 $\lambda\mathbf{a} = (\lambda a_1, \dots, \lambda a_n)$ 。
- ③ 内积： $\mathbf{a} \cdot \mathbf{b} = (a_1 b_1, \dots, a_n b_n)$ 。

矩阵

由 $m \times n$ 个数 a_{ij} ($i = 1, \dots, m; j = 1, \dots, n$) 排成的 m 行 n 列的数表称为 $m \times n$ 矩阵，记作

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

这 $m \times n$ 个数称为矩阵 \mathbf{A} 的**元素**。行数和列数都等于 n 的矩阵称为 **n 阶方阵**。只有一行的矩阵称为**行向量**：

$$\mathbf{A} = (a_1, a_2, \dots, a_n)$$

只有一列的矩阵称为**列向量**：

$$\mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$$

矩阵的加法

设有两个 $m \times n$ 矩阵 $\mathbf{A} = (a_{ij})$ 和 $\mathbf{B} = (b_{ij})$ ，那么矩阵 \mathbf{A} 和 \mathbf{B} 的和记为

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}$$

需要注意，只有两个矩阵的行数和列数相同时，才可以进行加法运算。
设 \mathbf{A} 、 \mathbf{B} 和 \mathbf{C} 都是 $m \times n$ 矩阵，则矩阵加法满足以下运算律：

- ① 交换律： $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ 。
- ② 结合律： $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

数与矩阵的乘法

实数 λ 与矩阵 \mathbf{A} 的乘积记作 $\lambda\mathbf{A}$ 或 $\mathbf{A}\lambda$ ，计算如下

$$\lambda\mathbf{A} = \mathbf{A}\lambda = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \dots & \lambda a_{mn} \end{pmatrix}$$

设 \mathbf{A} 和 \mathbf{B} 为 $m \times n$ 矩阵， λ 和 μ 为实数，则数与矩阵的乘法满足以下规律：

- ① $(\lambda\mu)\mathbf{A} = \lambda(\mu\mathbf{A})$
- ② $(\lambda + \mu)\mathbf{A} = \lambda\mathbf{A} + \mu\mathbf{A}$
- ③ $\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$

矩阵与矩阵相乘

设 \mathbf{A} 是一个 $m \times s$ 矩阵， \mathbf{B} 是一个 $s \times n$ 矩阵，那么矩阵 \mathbf{A} 与矩阵 \mathbf{B} 的乘积是一个 $m \times n$ 矩阵 $\mathbf{C} = \mathbf{AB}$ ，其中

$$c_{ij} = \sum_{k=1}^s a_{ik} b_{kj}$$

其中， a_{ik} 是矩阵 \mathbf{A} 的元素， b_{kj} 是矩阵 \mathbf{B} 的元素， c_{ij} 是矩阵 \mathbf{C} 的元素。注意，当且仅当左矩阵的列数等于右矩阵的行数时，两个矩阵才能相乘。

$$\begin{pmatrix} 1 & 0 & 3 & -1 \\ 2 & 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 4 & 1 & 0 \\ -1 & 1 & 3 \\ 2 & 1 & 0 \\ 1 & 3 & 4 \end{pmatrix} = \begin{pmatrix} 9 & -2 & -1 \\ 9 & 9 & 11 \end{pmatrix}$$

矩阵的转置

把矩阵 \mathbf{A} 的行换成同序数的列得到一个新矩阵，称为 \mathbf{A} 的转置矩阵，记作 \mathbf{A}^\top 。例如：

$$\mathbf{A} = \begin{pmatrix} 9 & -2 & -1 \\ 9 & 9 & 11 \end{pmatrix} \quad \mathbf{A}^\top = \begin{pmatrix} 9 & 9 \\ -2 & 9 \\ -1 & 11 \end{pmatrix}$$

矩阵的转置满足下述运算规律：

- ① $(\mathbf{A}^\top)^\top = \mathbf{A}$
- ② $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
- ③ $(\lambda \mathbf{A})^\top = \lambda \mathbf{A}^\top$
- ④ $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

方阵的行列式

由 n 阶方阵 \mathbf{A} 的元素所构成的行列式，称为方阵 \mathbf{A} 的行列式，记作 $|\mathbf{A}|$ 或 $\det\mathbf{A}$ 。给定一个两行两列的方阵，其行列式计算公式为

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \det\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

例如，给定一个方阵

$$\mathbf{A} = \begin{pmatrix} 3 & -2 \\ 2 & 1 \end{pmatrix}$$

其行列式计算如下

$$\det\mathbf{A} = \begin{vmatrix} 3 & -2 \\ 2 & 1 \end{vmatrix} = 3 \times 1 - 2 \times (-2) = 7$$

三阶行列式

三行三列的方阵的行列式的计算更复杂一些，基本规律是先按照正向（即从上方往右下方）对角线求和，再按照反向（即从上方往左下方）对角线求和，最后计算两者之差。

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - \\ a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$$

对角矩阵

不在对角线上的元素都是0的矩阵称为**对角矩阵**:

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

该对角矩阵通常也记作: $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$

一个特殊的对角阵是**单元阵**, 所有的对角线元素都为1:

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

逆矩阵

对于 n 阶矩阵 \mathbf{A} ，如果有一个 n 阶矩阵 \mathbf{B}

$$\mathbf{AB} = \mathbf{BA} = \mathbf{E}$$

则说矩阵 \mathbf{A} 是可逆的，并把矩阵 \mathbf{B} 称为 \mathbf{A} 的逆矩阵。 \mathbf{A} 的逆矩阵通常记为 \mathbf{A}^{-1} 。对于可逆矩阵，有以下性质：

- ① 如果矩阵 \mathbf{A} 是可逆的，那么 \mathbf{A} 的逆矩阵是唯一的。
- ② 如果矩阵 \mathbf{A} 可逆，则 $|\mathbf{A}| \neq 0$ 。
- ③ 如果 $\mathbf{AB} = \mathbf{E}$ 或 $\mathbf{BA} = \mathbf{E}$ ，则 $\mathbf{B} = \mathbf{A}^{-1}$ 。
- ④ 如果 \mathbf{A} 可逆，则 \mathbf{A}^{-1} 亦可逆，且 $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$ 。
- ⑤ 如果 \mathbf{A} 和 \mathbf{B} 为同阶矩阵且均可逆，则 \mathbf{AB} 亦可逆，且 $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ 。

矩阵的初等变换

给定一个矩阵，以下三种变换称为初等行变换：

- ① 对调第 i 行和第 j 行，记作 $r_i \leftrightarrow r_j$ 。
- ② 第 i 行的所有元素乘以实数 k ，记作 kr_i 。
- ③ 把第 j 行所有元素的 k 倍加到第 i 行对应的元素上，记作 $r_i + kr_j$ 。

同理，可以定义矩阵的初等列变换：

- ① 对调第 i 列和第 j 列，记作 $c_i \leftrightarrow c_j$ 。
- ② 第 i 列的所有元素乘以实数 k ，记作 kc_i 。
- ③ 把第 j 列所有元素的 k 倍加到第 i 列对应的元素上，记作 $r_i + kc_j$ 。

矩阵的初等变换

执行多步初等变换操作将矩阵 \mathbf{A} 转换为矩阵 \mathbf{F} 。

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & -1 & 1 & 2 \\ 1 & 1 & -2 & 1 & 4 \\ 4 & -6 & 2 & -2 & 4 \\ 3 & 6 & -9 & 7 & 9 \end{pmatrix}$$

$$\mathbf{F} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

标准形与矩阵的秩

对于 $m \times n$ 矩阵 \mathbf{A} ，总可以经过初等行变换和列变换将其化简为以下形式

$$\mathbf{F} = \begin{pmatrix} \mathbf{E}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}_{m \times n}$$

其中， \mathbf{E}_r 表示维度为 r 的单元方阵， \mathbf{O} 表示元素全为0的矩阵。 \mathbf{F} 称为**标准形**， r 称为矩阵的**秩**。

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & -1 & 1 & 2 \\ 1 & 1 & -2 & 1 & 4 \\ 4 & -6 & 2 & -2 & 4 \\ 3 & 6 & -9 & 7 & 9 \end{pmatrix} \Rightarrow \mathbf{F} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

方阵的特征值和特征向量

设 \mathbf{A} 是 n 阶矩阵，如果存在实数 λ 和 n 维非零列向量 \mathbf{x} 使得以下等式成立：

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

则称 λ 是矩阵 \mathbf{A} 的**特征值**，非零向量 \mathbf{x} 为 \mathbf{A} 的对应于特征 λ 的**特征向量**。

例如，以下等式成立

$$\begin{pmatrix} -1 & 1 & 0 \\ -4 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 2 \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

矩阵的特征值和特征向量并不唯一。

$$\begin{pmatrix} -1 & 1 & 0 \\ -4 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix} \times \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix} = 1 \times \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}$$

内容提要

微积分

概率论

线性代数

信息论

信息量

什么是信息量？假设我们听到了两件事，分别如下：

- 事件A：巴西队获得了2022年FIFA世界杯冠军。
- 事件B：中国队获得了2022年FIFA世界杯冠军。

仅凭直觉来说，显而易见事件B的信息量比事件A的信息量要大（也就是“大新闻”）。究其原因，是因为事件A发生的概率很大，事件B发生的概率很小。所以当越不可能的事件发生了，我们获得的信息量就越大，而越可能发生的事件发生了，我们获得的信息量就越小。

因此，信息量应该和事件发生的概率有关。

熵

如果 X 是一个离散型随机变量，其概率分布为 $P(X = x) = p(x)$ ， $x \in \mathcal{X}$ 。其中， \mathcal{X} 表示随机变量所有取值的集合，则该随机变量的熵为：

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

我们约定 $0 \log_2 0 = 0$ 。

熵表示信源每发出一个符号所提供的平均信息量。一个随机变量的熵越大，其不确定性越大，相应地正确估计其值的可能性就越小。越不确定的随机变量需要越大的信息量来确定其值。

熵是一个非常重要的概念，在计算机科学中的很多领域都有着重要的应用，必须数量掌握。

熵的计算

x	-1	2	3
$p(x)$	0.25	0.50	0.25

$$H(X) = -0.25 \times \log_2 0.25 - 0.5 \times \log_2 0.5 - 0.25 \times \log_2 0.25 \\ = 1.50$$

x	-1	2	3
$p(x)$	0.33	0.34	0.33

$$H(X) = -0.33 \times \log_2 0.33 - 0.34 \times \log_2 0.34 - 0.33 \times \log_2 0.33 \\ = 1.58$$

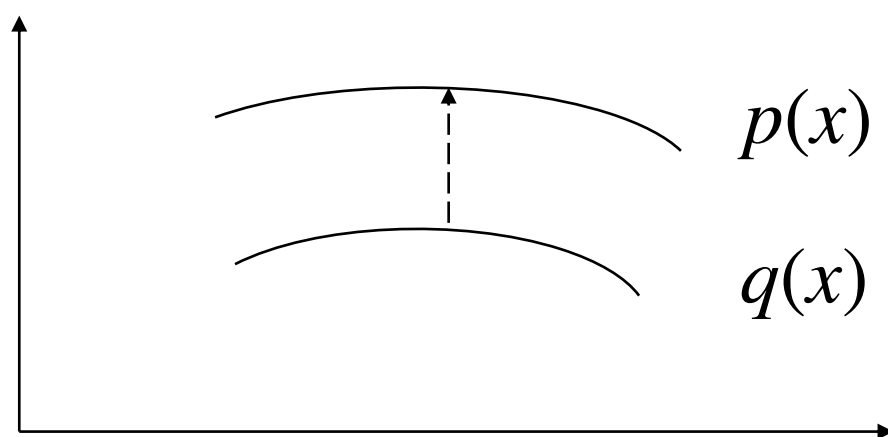
相对熵

两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵也称为KL散度（英文全称：Kullback-Leibler divergence），定义如下：

$$\text{KL}(p || q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

约定 $0 \log(0/q) = 0$ ， $p \log(p/0) = \infty$ 。

相对熵通常用于衡量两个概率分布的差距。当两个随机分布相同时，其相对熵为0。当两个随机分布的差别增加时，其相对熵也增加。



相对熵的计算

X	0	1
$p(X)$	$1/2$	$1/2$
$q_1(X)$	$1/4$	$3/4$
$q_2(X)$	$1/8$	$7/8$

给定上述概率分布，分别计算两个交叉熵

$$\text{KL}(p || q_1) = \frac{1}{2} \log_2 \left(\frac{1}{2} \times \frac{4}{1} \right) + \frac{1}{2} \log_2 \left(\frac{1}{2} \times \frac{4}{3} \right) = 0.21$$

$$\text{KL}(p || q_2) = \frac{1}{2} \log_2 \left(\frac{1}{2} \times \frac{8}{1} \right) + \frac{1}{2} \log_2 \left(\frac{1}{2} \times \frac{8}{7} \right) = 0.60$$

由于 $q_1(X)$ 相对于 $q_2(X)$ 更接近于 $p(X)$ ，相对熵的值也更小。

交叉熵

相对熵的公式可以表述为

$$\begin{aligned}\text{KL}(p||q) &= \sum_x p(x)\log p(x) - \sum_x p(x)\log q(x) \\ &= -H(p(x)) - \sum_x p(x)\log q(x)\end{aligned}$$

等式的前一部分是 p 的熵，而后一部分则是交叉熵：

$$H(p, q) = \sum_x p(x)\log q(x)$$

在人工智能中，往往需要评估模型分布和真实分布之间的差距，使用KL散度非常合适。但由于KL散度的前一部分跟真实分布相关，在优化过程中不变化，因此一般使用交叉熵作为损失函数并评估模型。

总结

- 数学是人工智能的基石，必须牢固掌握必备的数学基础知识，才能理解后续核心模型和算法。
- 主要知识点如下：
 - 微积分：多元函数偏导和极值的计算
 - 概率论：多维随机变量、数学期望
 - 线性代数：向量与矩阵的运算
 - 信息论：熵、相对熵、交叉熵

谢谢