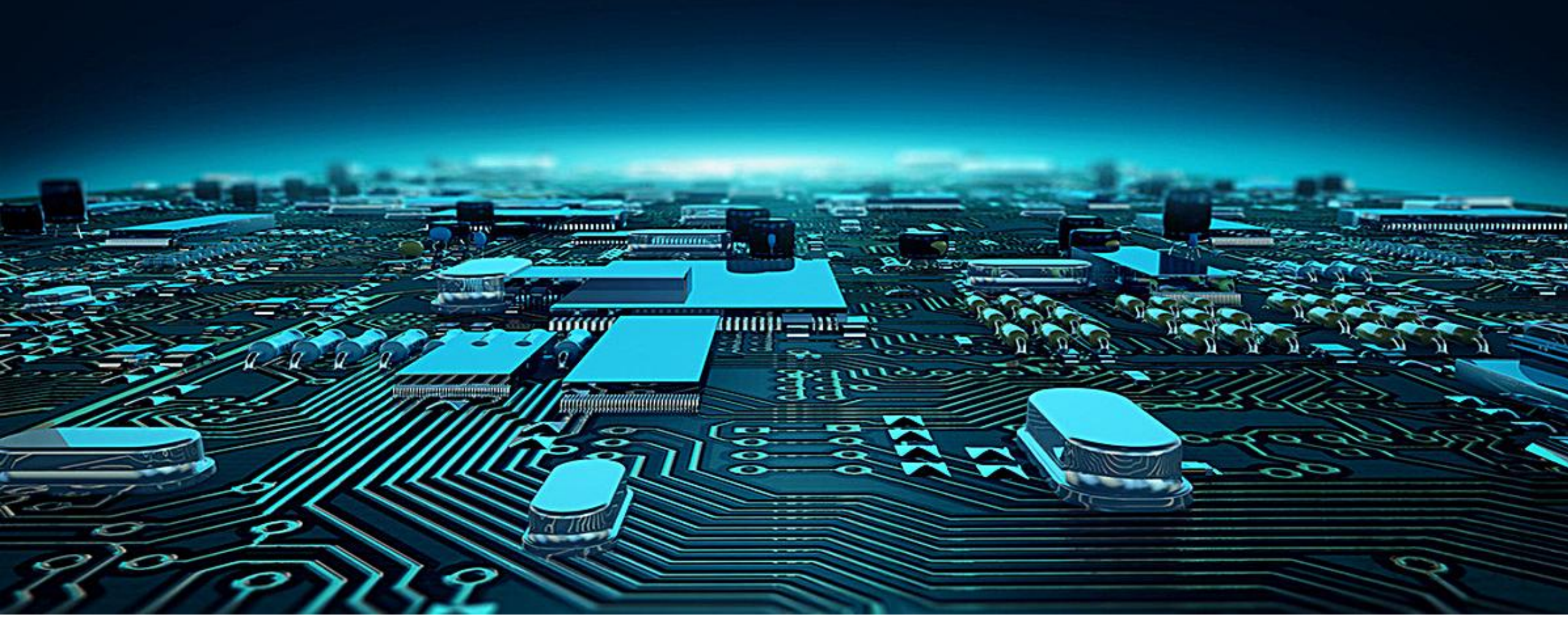
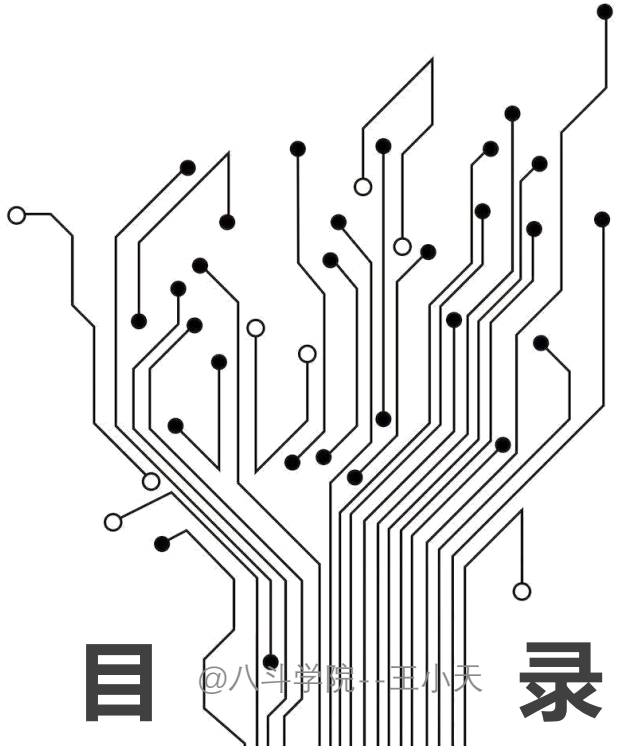


机器学习的数学基础

@八斗学院--王小天(Michael)

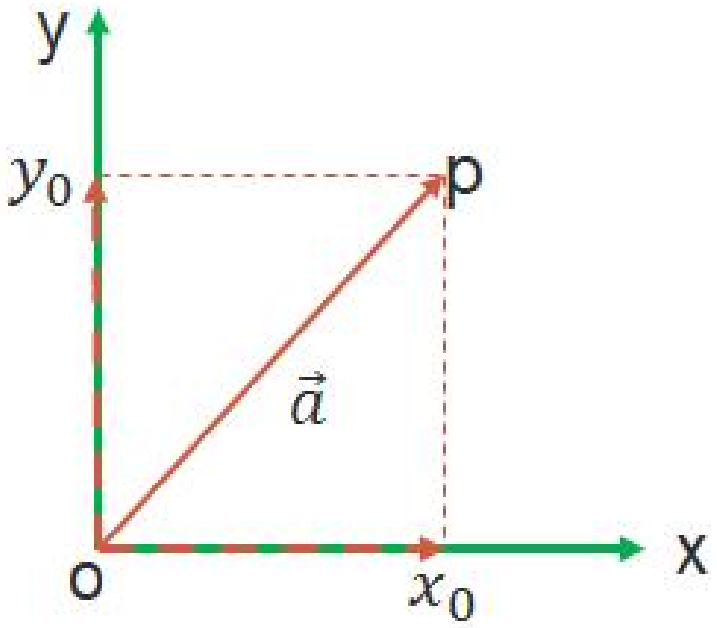
2024/03/24



目　录

向量

$$\vec{a}=(x,y)$$



线性变换

$$T(v+w)=T(v)+T(w)$$

$$T(cv)=cT(v)$$

矩阵

简单来说，矩阵是充满数字的表格。

$$\mathbf{A}=\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix},\ \mathbf{B}=\begin{bmatrix}5 & 1 & 2\\ 3 & 0 & -5\end{bmatrix}$$

矩阵加减法

两个矩阵相加或相减，需要满足两个矩阵的列数和行数一致。

加法交换律：A + B = B + A

$$\mathbf{A}=\begin{bmatrix}3 & -1\\ 2 & 0\end{bmatrix},\ \mathbf{B}=\begin{bmatrix}-7 & 2\\ 3 & 5\end{bmatrix}$$

$$\mathbf{A}+\mathbf{B}=\begin{bmatrix}3+(-7) & -1+2\\ 2+3 & 0+5\end{bmatrix}=\begin{bmatrix}-4 & 1\\ 5 & 5\end{bmatrix}$$

$$\mathbf{A}-\mathbf{B}=\begin{bmatrix}3-(-7) & -1-2\\ 2-3 & 0-5\end{bmatrix}=\begin{bmatrix}10 & -3\\ -1 & -5\end{bmatrix}$$

矩阵乘法

两个矩阵A和B相乘，需要满足A的列数等于B的行数

1. $\mathbf{A}=\begin{bmatrix}1 & -3\\ 7 & 5\end{bmatrix}$, $\mathbf{B}=\begin{bmatrix}10 & -8\\ 12 & -2\end{bmatrix}$

$$\boldsymbol{A}\times\boldsymbol{B}=\begin{bmatrix}1\times 10+(-3)\times 12 & 1\times(-8)+(-3)\times(-2)\\ 7\times 10+5\times 12 & 7\times(-8)+5\times(-2)\end{bmatrix}=\begin{bmatrix}-26 & -2\\ 130 & -66\end{bmatrix}$$

2. $\boldsymbol{A}=\begin{bmatrix}3 & 1 & 2\\ -2 & 0 & 5\end{bmatrix}$, $\mathbf{B}=\begin{bmatrix}-1 & 3\\ 0 & 5\\ 2 & 5\end{bmatrix}$

$$\boldsymbol{A}\times\boldsymbol{B}=\begin{bmatrix}3\times(-1)+1\times 0+2\times 2 & 3\times 3+1\times 5+2\times 5\\ -2\times(-1)+0\times 0+5\times 2 & -2\times 3+0\times 5+5\times 5\end{bmatrix}=\begin{bmatrix}1 & 24\\ 12 & 19\end{bmatrix}$$

$$\boldsymbol{B}\times\boldsymbol{A}=\begin{bmatrix}-1\times 3+3\times(-2) & -1\times 1+3\times 0 & -1\times 2+3\times 5\\ 0\times 3+5\times(-2) & 0\times 1+5\times 0 & 0\times 2+5\times 5\\ 2\times 3+5\times(-2) & 2\times 1+5\times 0 & 2\times 2+5\times 5\end{bmatrix}=\begin{bmatrix}-9 & -1 & 13\\ -10 & 0 & 25\\ -4 & 2 & 29\end{bmatrix}$$

单位矩阵

单位矩阵是一个n×n矩阵，从左到右的对角线上的元素是1，其余元素都为0。下面是三个单位矩阵：

$$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}$$

如果A是n×n矩阵，I是单位矩阵，则AI= A, IA = A

$$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\times\begin{bmatrix}a & b\\ c & d\end{bmatrix}=\begin{bmatrix}1\times a+0\times c & 1\times b+0\times d\\ 0\times a+1\times c & 0\times b+1\times d\end{bmatrix}=\begin{bmatrix}a & b\\ c & d\end{bmatrix}$$

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}\times\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=\begin{bmatrix}a\times 1+c\times 0 & b\times 1+d\times 0\\ a\times 0+c\times 1 & b\times 0+d\times 1\end{bmatrix}=\begin{bmatrix}a & b\\ c & d\end{bmatrix}$$

单位矩阵在矩阵乘法中的作用相当于数字1。

逆矩阵

矩阵A的逆矩阵记作A-1， A A-1=A-1A= I，I是单位矩阵。

1. $\boldsymbol{A}=\begin{bmatrix} a & b \\ c & d \end{bmatrix}$的逆矩阵

   $|\boldsymbol{A}| = ad - bc$, $|\boldsymbol{A}|$是$\boldsymbol{A}$的二阶行列式

   $$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

2. $\boldsymbol{B}=\begin{bmatrix} 3 & -4 \\ 2 & -5 \end{bmatrix}$的逆矩阵

   $$\boldsymbol{B}^{-1} = \frac{1}{|\boldsymbol{B}|}\begin{bmatrix} -5 & 4 \\ -2 & 3 \end{bmatrix} = \frac{1}{3\times(-5)-(-4)\times 2}\begin{bmatrix} -5 & 4 \\ -2 & 3 \end{bmatrix} = -\frac{1}{7}\begin{bmatrix} -5 & 4 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 5/7 & -4/7 \\ 2/7 & -3/7 \end{bmatrix}$$

奇异矩阵

当一个矩阵没有逆矩阵的时候，称该矩阵为奇异矩阵。
->当且仅当一个矩阵的行列式为零时，该矩阵是奇异矩阵。

$$A=\begin{bmatrix}a & b\\ c & d\end{bmatrix},\quad A^{-1}=\frac{1}{|A|}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}=\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$$

当ad-bc=0时，|A|没有定义，A-1不存在，A是奇异矩阵。

矩阵的转置

简单地说，矩阵的转置就是行列互换，用A^T表示A的转置矩阵。

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix},\qquad A^T=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$$

转置运算特性：

$$(A^T)^T=A$$

$$(AB)^T=B^TA^T$$

对称矩阵

如果一个矩阵转置后等于原矩阵，那么这个矩阵称为对称矩阵。

一个矩阵转置和这个矩阵的乘积就是一个对称矩阵。

$$A=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix},\qquad A^T=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}$$

$$A^TA=\begin{bmatrix}1 & 4\\ 2 & 5\\ 3 & 6\end{bmatrix}\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix}=\begin{bmatrix}17 & 22 & 27\\ 22 & 29 & 36\\ 27 & 36 & 45\end{bmatrix}$$

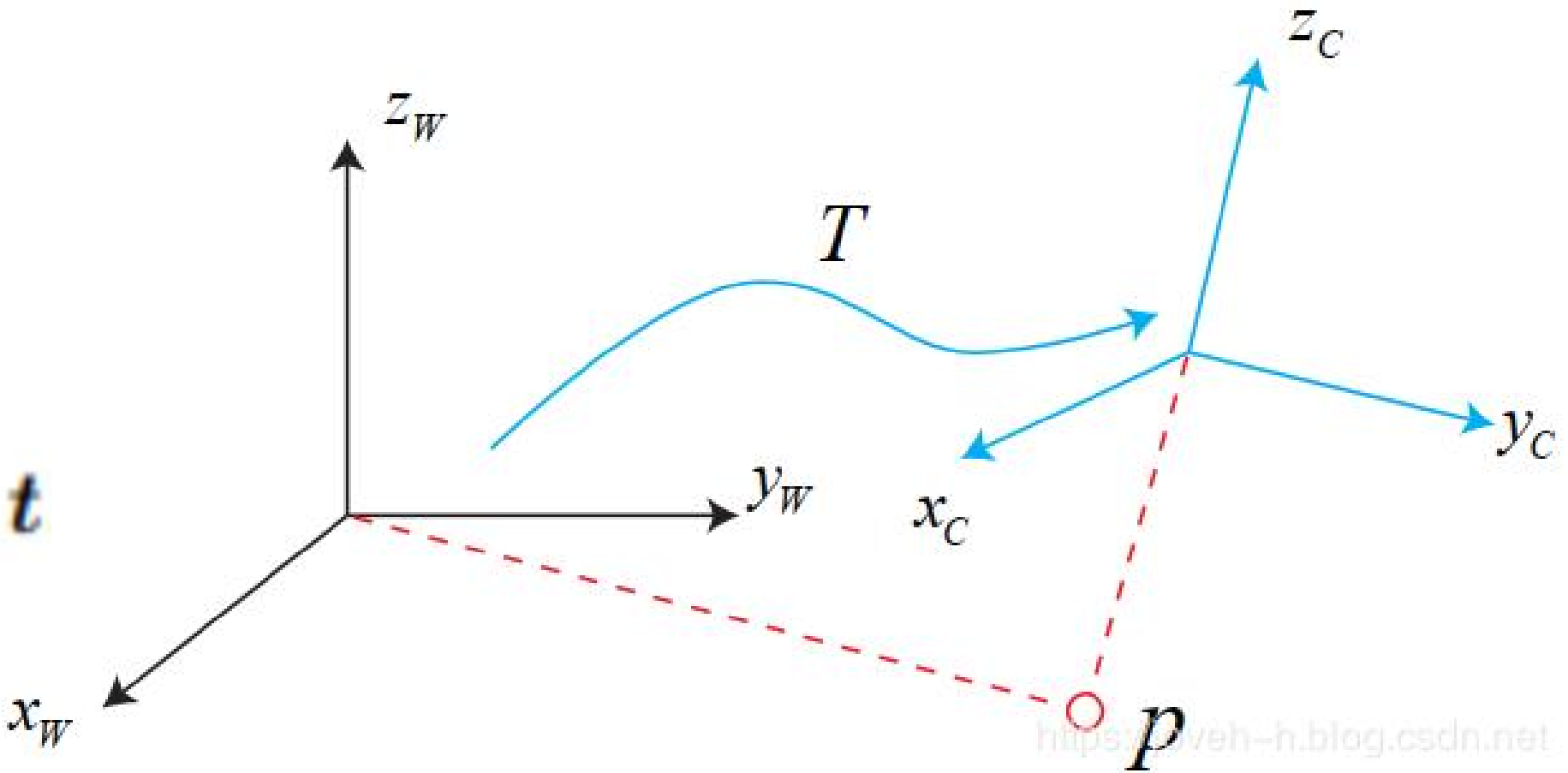
$$\begin{bmatrix}17 & 22 & 27\\ 22 & 29 & 36\\ 27 & 36 & 45\end{bmatrix}^T=\begin{bmatrix}17 & 22 & 27\\ 22 & 29 & 36\\ 27 & 36 & 45\end{bmatrix}$$

欧氏变换

欧氏变换由两部分组成:

- 旋转
- 平移

$$a' = Ra + t$$



齐次坐标

简而言之，齐次坐标就是用N+1维来代表N维坐标

我们可以在一个2D坐标末尾加上一个额外的变量w来形成2D齐次坐标
因此，一个点(X,Y)在齐次坐标里面变成了（x,y,w），并且有
X = x/w
Y = y/w

$$\underset{\text{Homogeneous}}{(x,y,w)} \quad \Leftrightarrow \quad \underset{\text{Cartesian}}{\left(\frac{x}{w},\frac{y}{w}\right)}$$

例如：
（1，2）的齐次坐标可以表示为（1，2，1）。

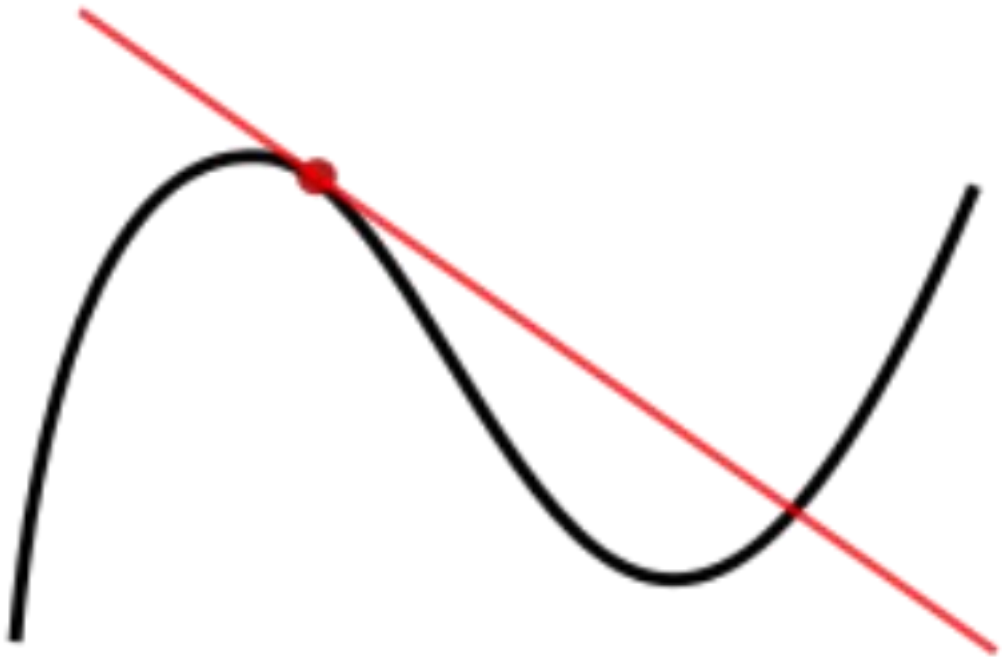
如果点（1，2）移动到无限远处，在笛卡尔坐标下它变为(∞,∞)，然后它的齐次坐标表示为（1，2，0），因为(1/0, 2/0) = (∞,∞)，我们可以不用"∞"来表示一个无穷远处的点了

导数&偏导数

导数（微分）：是代表函数（曲线）的斜率，是描述函数（曲线）变化快慢的量，同时曲线的极大值点也可以使用导数来判断，即极大值点的导数为0，此时斜率为零。

偏导数：是指在多元函数的情况下，对其每个变量进行求导，求导时，把其他变量看做常量进行处理，物理意义就是查看这一个变量在其他情况不变的情况下对函数的影响程度。

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



梯度

梯度: 梯度的本意是一个**向量**（矢量），表示某一函数在该点处的方向导数沿着该方向取得最大值，即函数在该点处沿着该方向（此梯度的方向）变化最快，变化率最大（为该梯度的模）

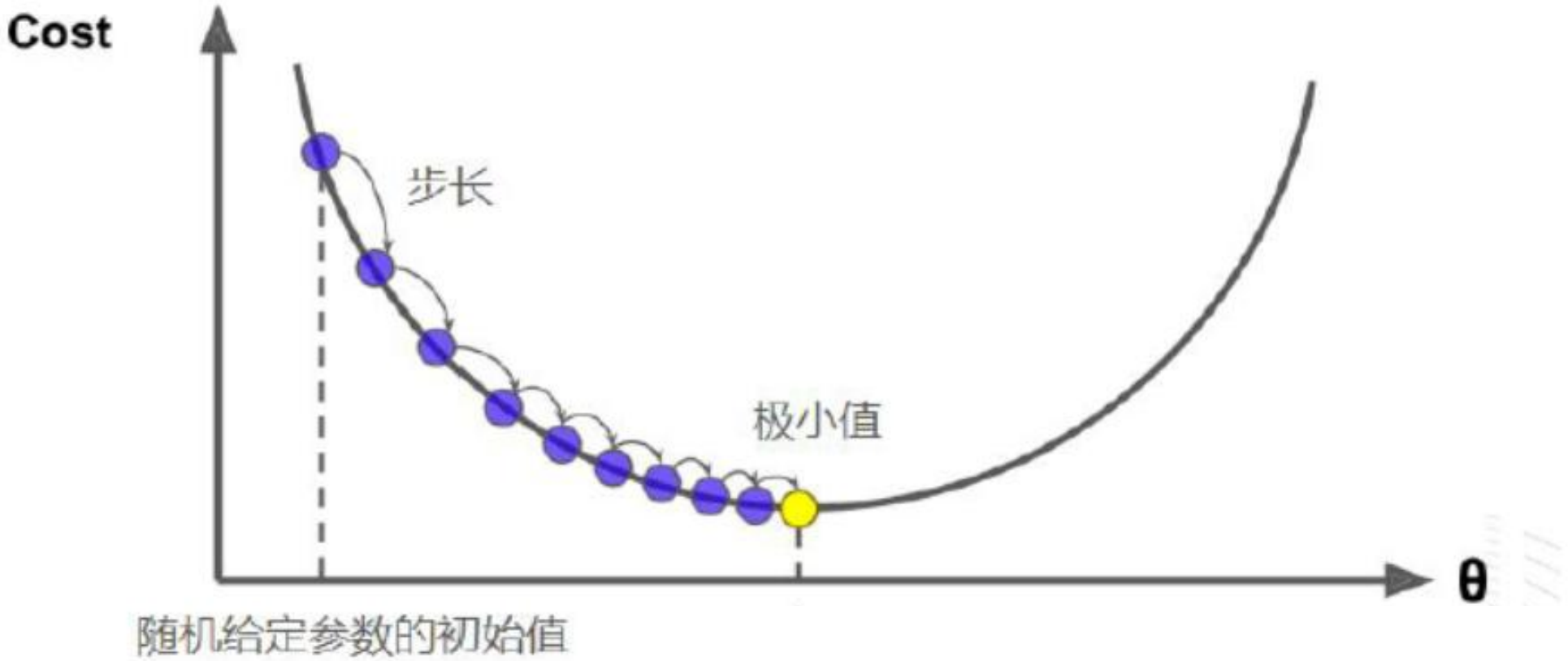
简而言之，对多元函数的各个自变量求偏导数，并把求得的这些偏导数写成向量形式，就是梯度。

$$\nabla f(x_0, y_0) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y})$$

梯度下降法：是一种寻找函数极小值的方法。

该方法最普通的做法是：在已知**参数当前值**的情况下，按当前点对应的**梯度向量的反方向**，并按事先给定好的**步长**大小，对参数进行调整。

按如上方法对参数做出多次调整之后，函数就会逼近一个极小值。

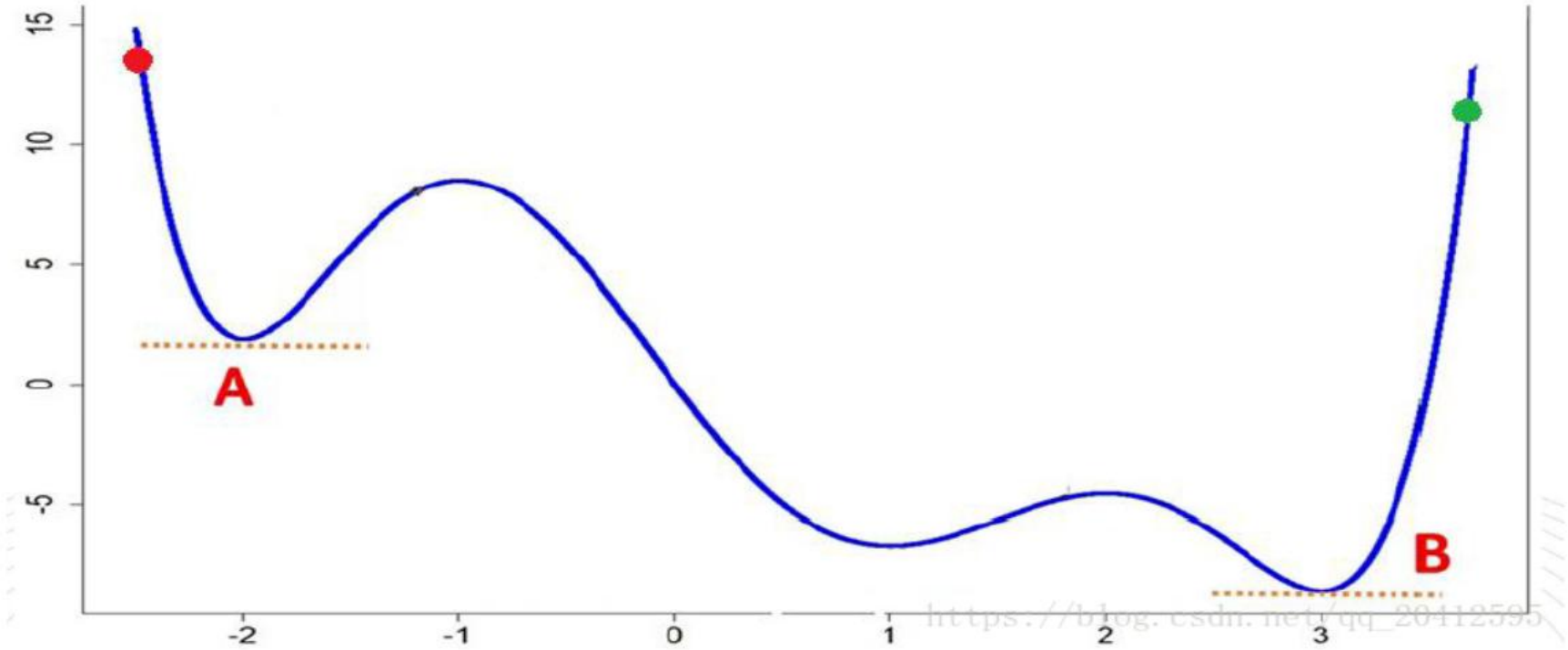


梯度

梯度下降法存在的问题：

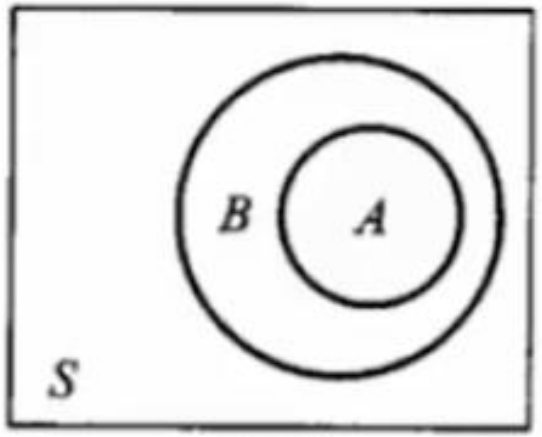
1.参数调整缓慢

2.收敛于局部最小值

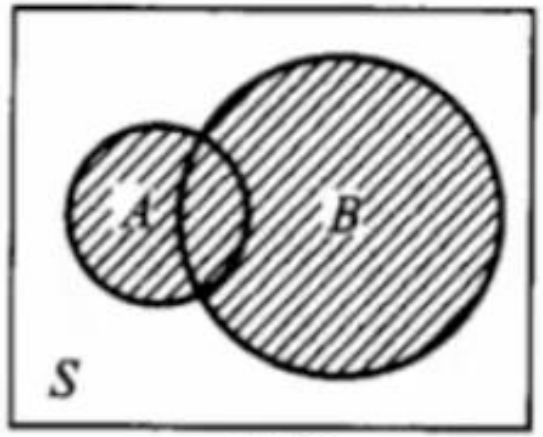


Machine Learning与Traditional statistical analyses的一些区别，主要在关注主体和验证性作区分。前者不关心模型的复杂度有多么的高，仅仅要求模型有良好的泛化性以及准确性。而后者在模型本身有一定的要求——不可过于复杂。

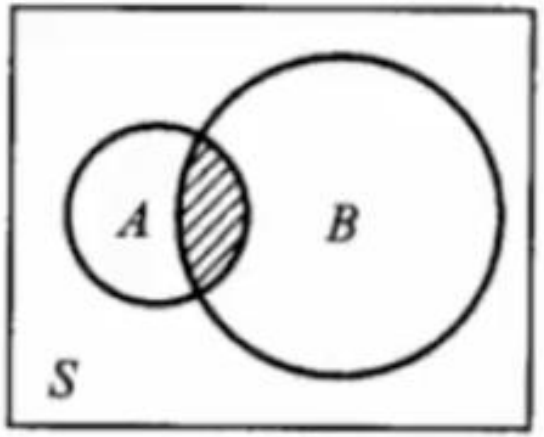
事件与关系运算



$A \subset B$

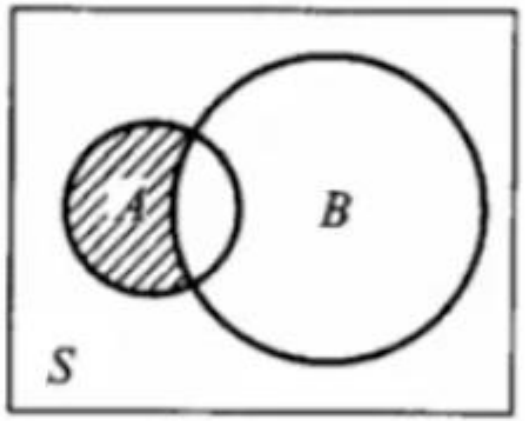


$A \cup B$

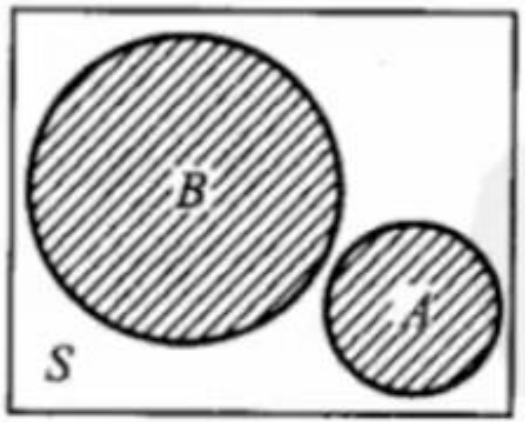


$A \cap B$

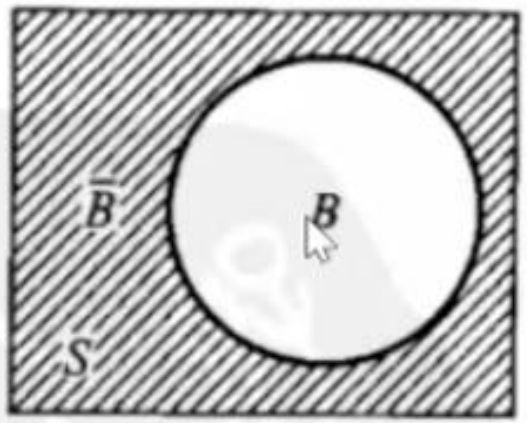
事件与关系运算



$A-B$



$A \cap B=\varnothing$



$B \cup \bar{B}=S, B \cap \bar{B}=\varnothing$

事件运算定律

(1) 交换律：$A \cup B = B \cup A, A \cap B = B \cap A$

(2) 结合律：$(A \cup B) \cup C = A \cup (B \cup C)$；　$(A \cap B) \cap C = A \cap (B \cap C)$

(3) 分配律：$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

概率的基本概念

(1) 概率：事件发生的可能性大小的度量，其严格定义如下：

概率P(g)为定义在事件集合上的满足下面2个条件的函数：

- 1)对任何事件A，P(A) >= 0
- 2)对必然事件B，P(B) = 1

(2) 概率的基本性质:

1) $P(\overline{A}) = 1 - P(A)$;

2) $P(A - B) = P(A) - P(AB)$;

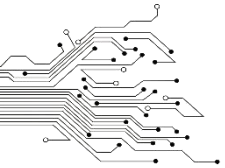
(3) 古典型概率: 实验的所有结果只有有限个，且每个结果发生的可能性相同，其概率计算公式:

$$P(A) = \frac{\text{事件 } A \text{ 发生的基本事件数}}{\text{基本事件总数}}$$

独立性

设A，B为随机事件，若同时发生的概率等于各自发生的概率的乘积，则A，B**相互独立**。

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



离散

离散就是不连续。

数学期望、方差、标准差

1. 数学期望（均值）：表示一件事平均发生的概率，记为E(x), E(x) = x1p1+x2p2+...+ xnpn。或者：

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx$$

2. 方差：用来刻画随机变量x和数学期望E(x)之间的偏离程度，记做D(x)。

$$D(x) = \sum_{k=1}^{\infty}[x_k - E(X)]^2 p_k \quad 或 \quad D(x) = \int_{-\infty}^{\infty}[x_k - E(X)]^2 f(x)dx$$

变换后可由下式来计算：

$$D(X) = E(X^2) - [E(X)]^2$$

$$s^2 = \frac{1}{n}[(x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2]$$

3. 标准差（均方差）：标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。

正态分布（高斯分布）

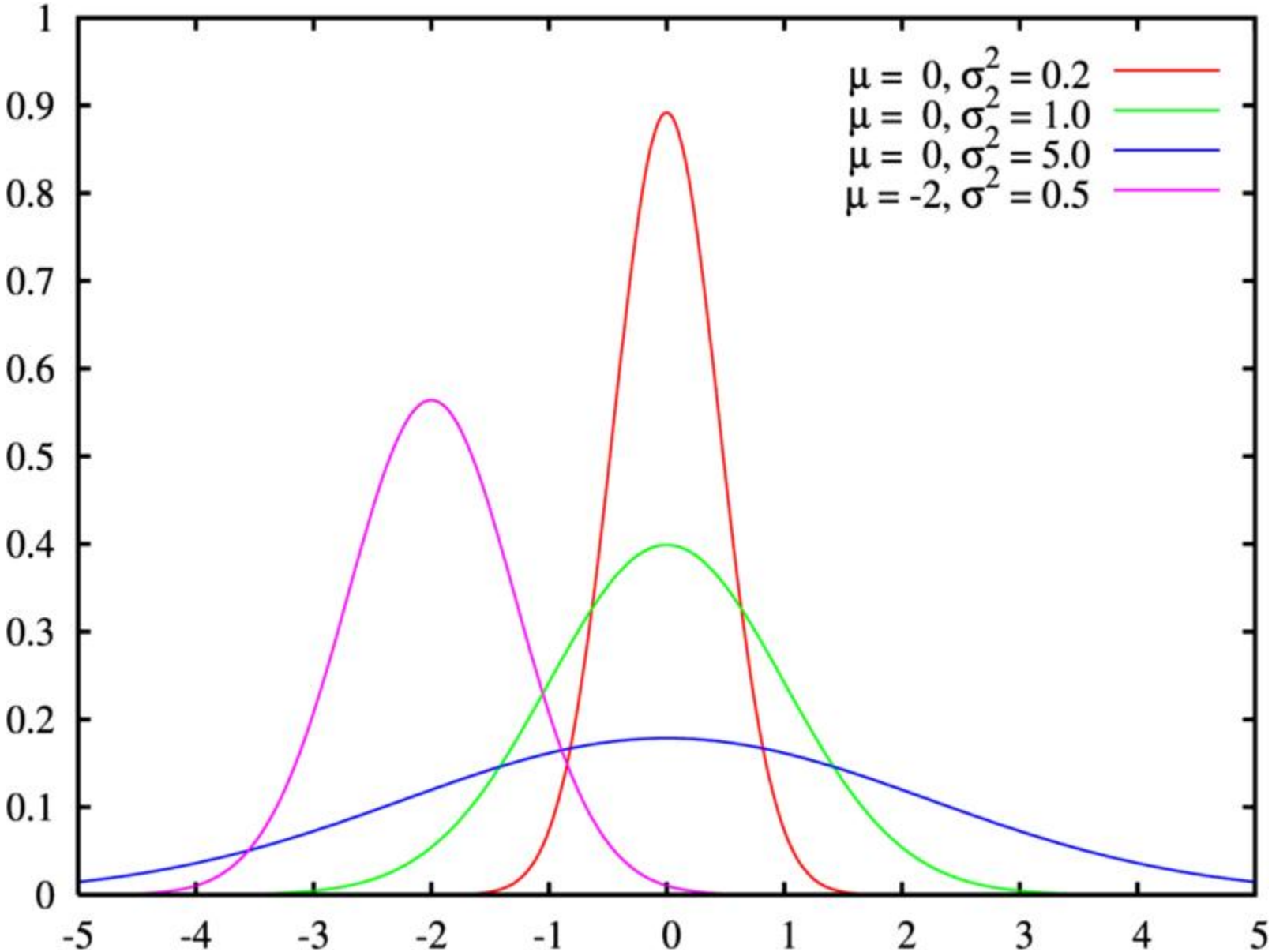
正态分布：若随机变量X服从一个数学期望为μ、方差为σ^2的正态分布，记为N(μ，σ^2)。

μ决定了其位置（中心线），其标准差σ决定了分布的幅度（胖瘦）。

标准正态分布：当μ = 0，σ = 1时的正态分布是标准正态分布

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

正态分布（高斯分布）



熵 entropy

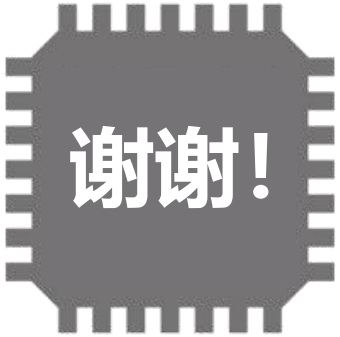
物理学上，是“混乱”程度的量度。
系统越有序，熵值越低；系统越混乱或者分散，熵值越高。

信息理论：
1、当系统的有序状态一致时，数据越集中的地方熵值越小，数据越分散的地方熵值越大。这是从信息的完整性上进行的描述。
2、当数据量一致时，系统越有序，熵值越低；系统越混乱或者分散，熵值越高。这是从信息的有序性上进行的描述。

若不确定性越大，则信息量越大，熵越大
若不确定性越小，则信息量越小，熵越小

假如事件A的分类划分是（A1,A2,⋯,An），每部分发生的概率是(p1,p2,⋯,pn)，那信息熵定义为公式如下：

$$Ent(A) = -\sum_{k=1}^{n} p_k log_2 p_k$$


谢谢！