



中山大學

SUN YAT-SEN UNIVERSITY

第二讲 数学基础

张瑞茂

zhangrm27@mail.sysu.edu.cn

本讲内容

01

线性代数

02

微积分

03

数学优化

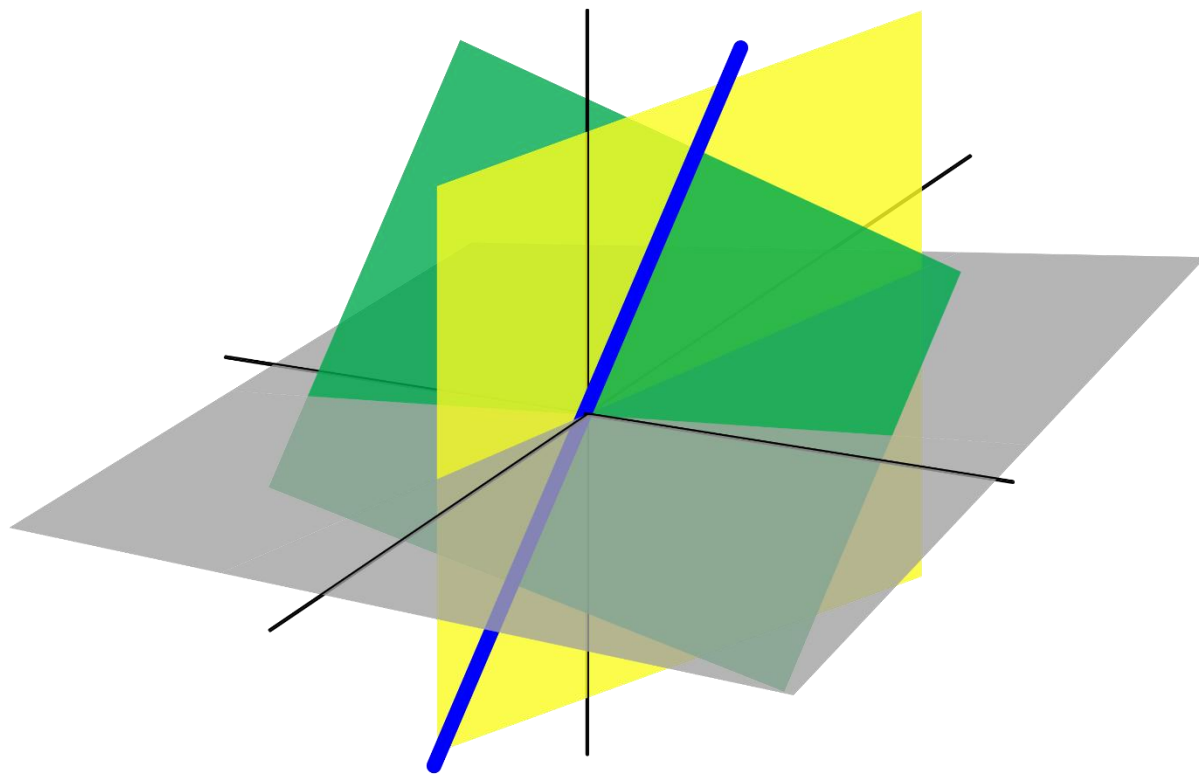
04

概率论

05

信息论

- 数学的一个分支，广泛应用于科学和工程中
- 对于理解和从事机器学习算法相关工作很有必要
- 主要包含向量、向量空间（线性空间）以及向量的线性变换和有限维的线性方程组

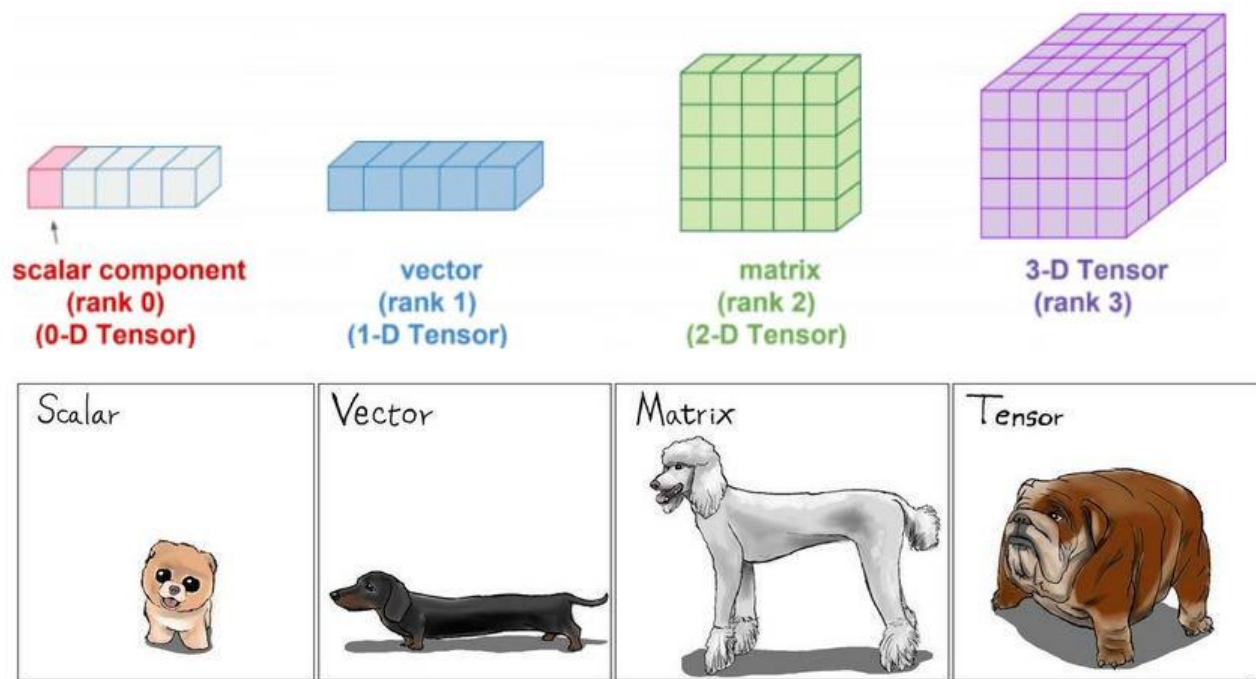


- 标量 (scalar) : 一个单独的数, 只有大小, 无方向, 用斜体小写表示, 如 x
- 向量 (vector) : 一系列有序排列的数, 索引确定具体元素, 用粗体小写表示, 如

$$\mathbf{x} = [x_1, x_2, \dots, x_n]^\top$$

- 矩阵 (matrix) : 二维数组, 两个索引确定具体元素, 粗体大写表示, 如 $\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$

- 张量 (tensor) : 数组元素分布在若干维坐标的规则网格中, 用字体 $\mathbf{A}$



重点

定义 2.1 (向量空间 (Vector Space))

也称线性空间 (LinearSpace), 是指由向量组成的集合, 并满足以下两个条件:

1. 向量加法 $+$: 向量空间 $\mathbb{V}$ 中的两个向量 $\mathbf{a}$ 和 $\mathbf{b}$, 它们的和 $\mathbf{a} + \mathbf{b}$ 也属于空间 $\mathbb{V}$
2. 标量乘法 $\cdot$: 向量空间 $\mathbb{V}$ 中的任一向量 $\mathbf{a}$ 和任一标量 c , 它们的乘积 $c\mathbf{a}$ 也属于空间 $\mathbb{V}$

**定义 2.3 (线性子空间)**

向量空间 $\mathbb{V}$ 的线性子空间 $\mathbb{U}$ 是 $\mathbb{V}$ 的一个子集, 并且满足向量空间的条件 (向量加法和标量乘法)

**定义 2.4 (线性无关)**

向量空间 $\mathbb{V}$ 中 n 个向量 $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, 若对任意一组标量 $\lambda_1, \lambda_2, \dots, \lambda_n$, 满足 $\sum_{i=1}^n \lambda_i \mathbf{v}_i = \mathbf{0}$, 则必然 $\lambda_1 = \lambda_2 = \dots = \lambda_n = 0$, 那么 $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ 是线性无关的, 也称为线性独立



定义 2.5 (基向量)

n 维向量空间 $\mathbb{V}$ 的基 $\mathbb{B} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ 是 $\mathbb{V}$ 的有限子集, 其元素之间线性无关。

$\mathbb{V}$ 中所有的向量都可以按唯一方式表达为 $\mathbb{B}$ 中向量的线性组合

对任意 $\mathbf{v} \in \mathbb{V}$, 存在一组标量 $(\lambda_1, \lambda_2, \dots, \lambda_n)$ 使得

$$\mathbf{v} = \sum_{i=1}^n \lambda_i \mathbf{e}_i \quad (2.3)$$

其中基 $\mathbb{B}$ 中的向量为基向量。若基向量有序, 则 $(\lambda_1, \lambda_2, \dots, \lambda_n)$ 称为 $\mathbf{v}$ 关于 $\mathbb{B}$ 的坐标。 n 维空间 $\mathbb{V}$ 的一组标准基为

$$\mathbf{e}_1 = [1, 0, 0, \dots, 0], \quad (2.4)$$

$$\mathbf{e}_2 = [0, 1, 0, \dots, 0], \quad (2.5)$$

$$\dots \quad (2.6)$$

$$\mathbf{e}_n = [0, 0, 0, \dots, 1] \quad (2.7)$$

$\mathbb{V}$ 中任一向量 $\mathbf{v} = [v_1, v_2, \dots, v_n]$ 可唯一表示为

$$[v_1, v_2, \dots, v_n] = v_1 \mathbf{e}_1 + v_2 \mathbf{e}_2 + \dots + v_n \mathbf{e}_n \quad (2.8)$$

$v_1, v_2, \dots, v_n$ 称为向量 $\mathbf{v}$ 的笛卡尔座标。



定义 2.6 (点积 (dot product))

$\langle a, b \rangle = a^\top b = \sum_i a_i b_i = c$ 也称内积或标量积

**定义 2.7 (正交 (orthogonal))**

$\langle a, b \rangle = a^\top b = 0$, 则 a 和 b 互相正交 (orthogonal)

$\mathbb{V}^n$ 中, 至多有 n 个非零向量互相正交

若 v 与子空间 $\mathbb{U}$ 中每个向量都正交, 则 v 和子空间 $\mathbb{U}$ 正交



- 范数（包括 L_p 范数）是将向量映射到非负值的函数。

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

L^2 范数

即欧几里得范数（Euclidean norm），表示从原点到向量 $\mathbf{x}$ 点的欧几里得距离，简记为 $\|\mathbf{x}\|$

$$\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}} \quad (2.10)$$

平方 L^2 范数也经常用来衡量向量的大小，可通过点积 $\mathbf{x}^\top \mathbf{x}$ 计算，在数学和计算上都比 L^2 范数方便，如平方 L^2 范数求导

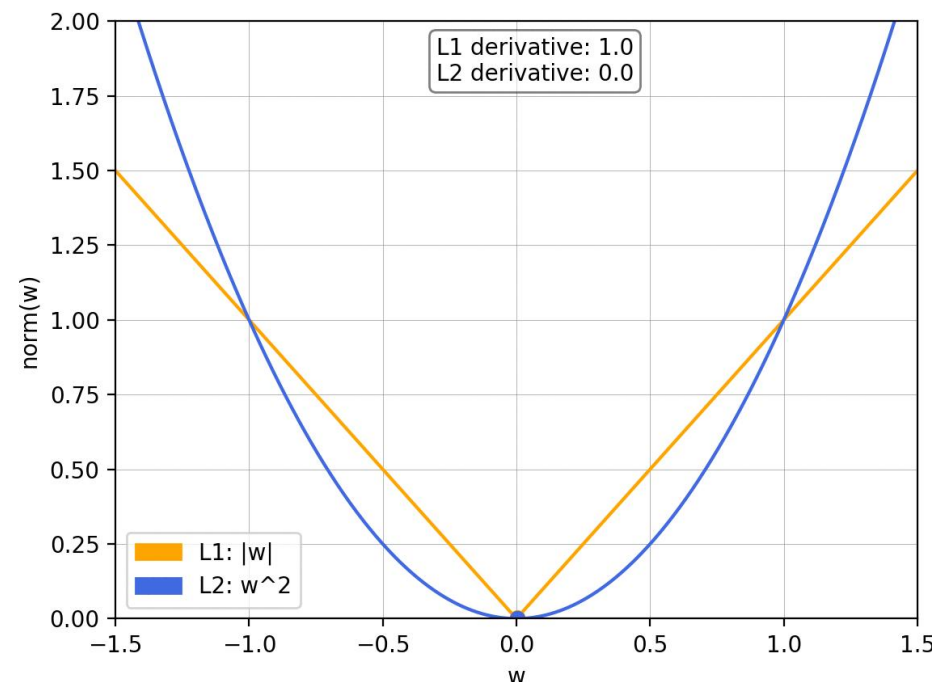
L^1 范数

平方 L^2 范数在原点附近变化缓慢，某些机器学习应用中，区分零和非零小值是很重要的，可使用 L^1 范数：

$$\|\mathbf{x}\|_1 = \sum_i |x_i|. \quad (2.11)$$

当 $\mathbf{x}$ 中某元素从 0 增加 ϵ ，对应的 L^1 范数也增加 ϵ

?



重点

L^∞ 范数

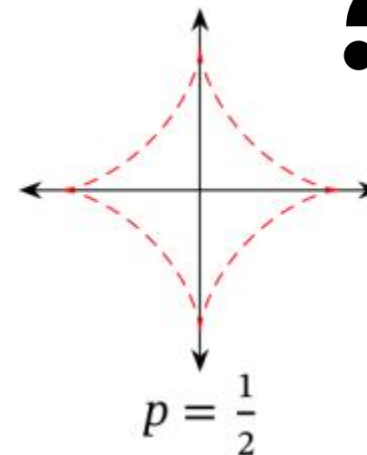
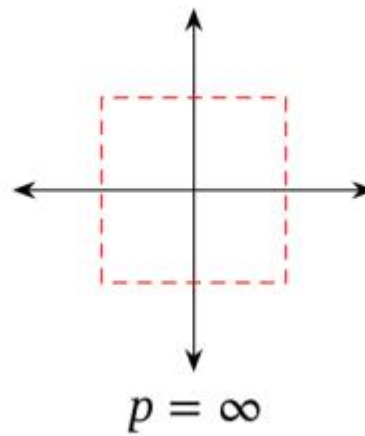
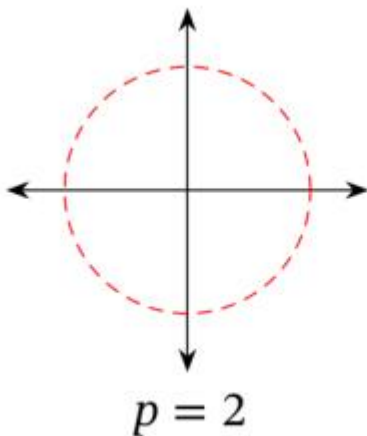
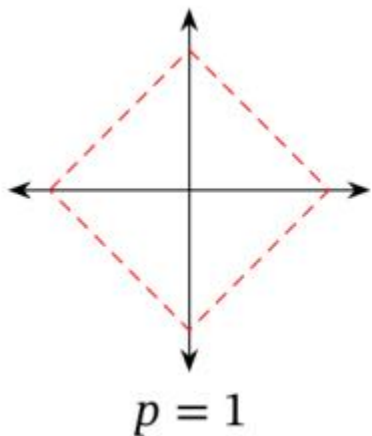
即 最大范数 (max norm), 表示向量中最大幅值元素的绝对值:

$$\|\mathbf{x}\|_\infty = \max_i |x_i|. \quad (2.12)$$

Frobenius 范数 (Frobenius norm)

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}, \quad (2.13)$$

其类似于向量的 L^2 范数。



?

- 转置 (transpose):

$$(A^{\top})_{i,j} = A_{j,i}.$$

- 加: $C = A + B$, 其中 $C_{i,j} = A_{i,j} + B_{i,j}$

- 标量和矩阵相乘: $D = a \cdot B + c$, 其中 $D_{i,j} = a \cdot B_{i,j} + c$

- 矩阵和向量相乘:

$$C = AB.$$

具体地

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}.$$

- **Hadamard 乘积** (Hadamard product): $C = A \odot B$, 其中 $C_{i,j} = A_{i,j} \cdot B_{i,j}$

- 结合律:

$$A + (B + C) = (A + B) + C$$

$$A(BC) = (AB)C$$

- 交换律:

$$A + B = B + A$$

- 分配律:

$$A(B + C) = AB + AC.$$

定义 2.9 (单位矩阵)

- 任意向量和单位矩阵相乘，都不会改变
- 保持 n 维向量不变的单位矩阵记作 $I_n \in \mathbb{R}^{n \times n}$

$$\forall \mathbf{x} \in \mathbb{R}^n, I_n \mathbf{x} = \mathbf{x}. \quad (2.22)$$

- 单位矩阵主对角线元素为 1，其他位置元素为 0

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



定义 2.10 (矩阵逆)

- 矩阵 A 的矩阵逆 (matrix inversion) 记作 A^{-1} ，满足 $A^{-1}A = I_n$



定义 2.11 (对角矩阵 (diagonal matrix))

矩阵 D 是对角矩阵, 当且仅当对于所有的 $i \neq j$, $D_{i,j} = 0$.



- $\text{diag}(\mathbf{v})$ 表示一个对角元素由向量 $\mathbf{v}$ 中元素给定的对角方阵
- 对角阵乘法: $\text{diag}(\mathbf{v})\mathbf{x} = \mathbf{v} \odot \mathbf{x}$
- 对角方阵的逆矩阵存在, 当且仅当对角元素都是非零值
- 对角方阵求逆矩: $\text{diag}(\mathbf{v})^{-1} = \text{diag}([1/v_1, \dots, 1/v_n]^\top)$
- 不是所有的对角矩阵都是方阵, 长方形的矩阵也有可能是对角矩
- 非方阵的对角矩阵没有逆矩阵

定义 2.12 (对称 (symmetric))

矩阵是转置和自己相等的矩阵:

$$\mathbf{A} = \mathbf{A}^\top. \quad (2.23)$$



对称阵举例: 距离度量矩阵 $\mathbf{A}$, $\mathbf{A}_{i,j}$ 表示点 i 到点 j 的距离, 那么 $\mathbf{A}_{i,j} = \mathbf{A}_{j,i}$

定义 2.13 (正交矩阵 (orthogonal matrix))

行向量和列向量是分别标准正交的方阵:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{A} \mathbf{A}^\top = \mathbf{I}. \quad (2.24)$$

这意味着

$$\mathbf{A}^{-1} = \mathbf{A}^\top \quad (2.25)$$



特征值与特征向量

对于方阵 A ，若存在一个标量 λ 和一个非零向量 v 满足：

$$Av = \lambda v. \quad (2.26)$$

则 λ 和 v 分别称为 A 的 **特征值** (eigenvalue) 和 **特征向量** (eigenvector)

特征分解

特征分解 (eigendecomposition) 将矩阵分解成一组特征向量和特征值

若 A 有 n 个线性无关的特征向量 $\{v^{(1)}, \dots, v^{(n)}\}$ ，对应特征值 $\{\lambda_1, \dots, \lambda_n\}$ ，记 $V = [v^{(1)}, \dots, v^{(n)}]$ ， $\lambda = [\lambda_1, \dots, \lambda_n]^T$ ，则 A 的 **特征分解** (eigendecomposition) 可记作

$$A = V \text{diag}(\lambda) V^{-1}. \quad (2.27)$$

A 为实对称矩阵则：

$$A = Q \Lambda Q^T. \quad (2.28)$$

其中 Q 是 A 特征向量组成的正交矩阵， Λ 是对角矩阵。

$\Lambda_{i,i}$ 对应特征向量是 Q 的第 i 列，记作 $Q_{:,i}$ 。可将 A 看作沿方向 $v^{(i)}$ 延展 λ_i 倍的空间。

- 任一实对称矩阵 A 都有特征分解，但特征分解可能不唯一
- 特征分解唯一当且仅当所有特征值唯一，通常按降序排列 Λ 元素
- 矩阵奇异当且仅当含有零特征值

奇异值分解

定义 2.14 (奇异值分解)

将矩阵分解为 奇异向量 (singular vector) 和 奇异值 (singular value)

实数矩阵都有奇异值分解，但不一定有特征分解。如，非方阵无特征分解，但有奇异值分解。

特征分解：

$$A = V \text{diag}(\lambda) V^{-1}. \quad (2.29)$$

奇异值分解：

$$A = U D V^T. \quad (2.30)$$

其中， A 大小 $m \times n$ ， U 大小 $m \times m$ ， D 大小 $m \times n$ ， V 大小 $n \times n$ 。 U 和 V 为正交矩阵， D 为对角矩阵， D 不一定是方阵。

D 对角线元素为 A 的 **奇异值** (singular value)。

U 列向量为 **左奇异向量** (left singular vector)， V 列向量为 **右奇异向量** (right singular vector)。

事实上，可用与 A 相关的特征分解解释 A 的奇异值分解。 A 的 **左奇异向量** (left singular vector) 是 AA^T 的特征向量； A 的 **右奇异向量** (right singular vector) 是 $A^T A$ 的特征向量。 A 非零奇异值是 $A^T A$ 特征值的平方根，也是 AA^T 特征值的平方根。

重点

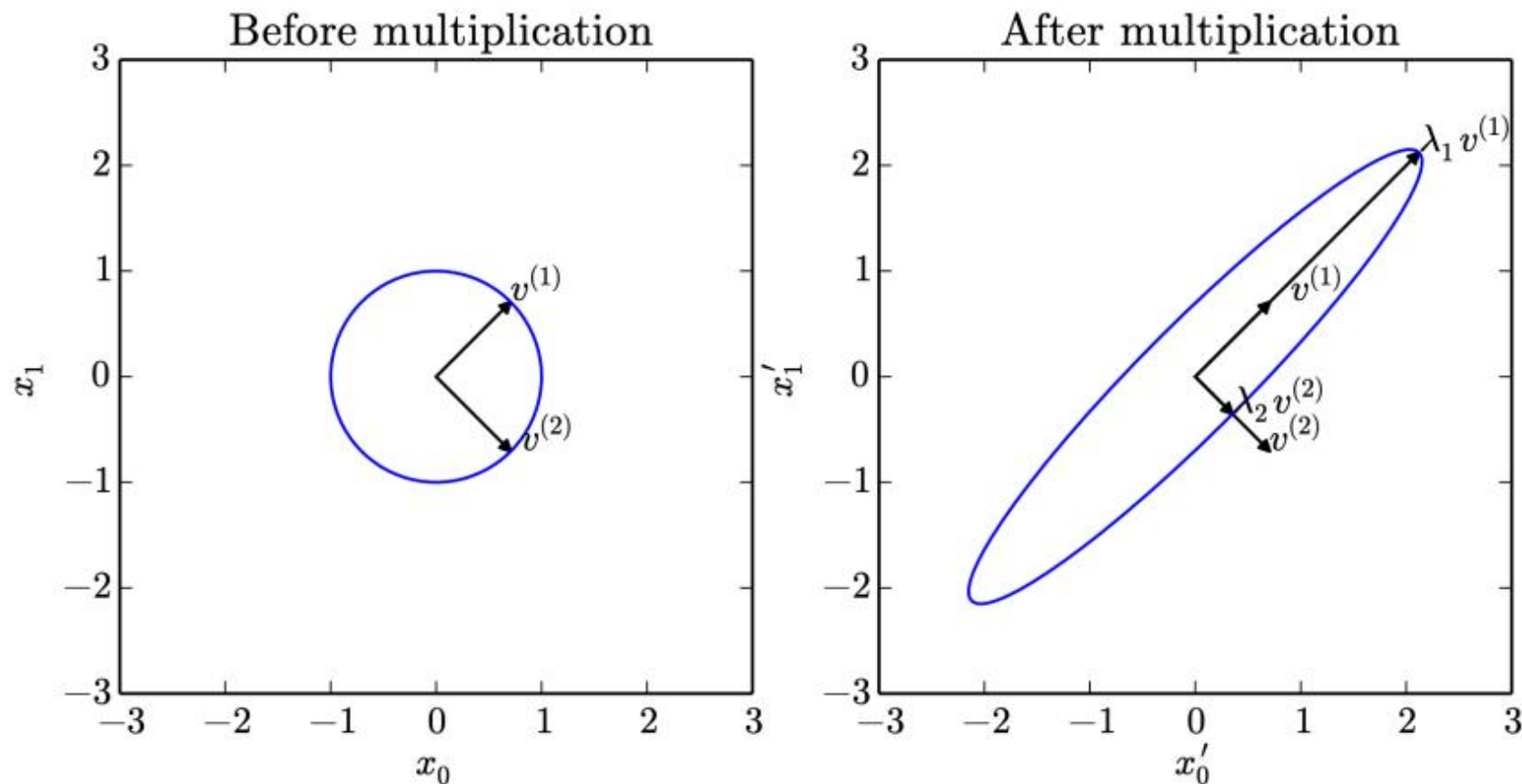


图 2.2: A 有两个标准正交特征向量, 特征值 λ_1 的 $v^{(1)}$, 特征值 λ_2 的 $v^{(2)}$ 。(左) 单位向量 $u \in \mathbb{R}^2$ 的集合, 构成一个单位圆。(右) 所有 Au 点集合。 A 拉伸单位圆, 将 $v^{(i)}$ 方向的空间拉伸了 λ_i 倍。

迹运算返回的是矩阵对角元素的和：

$$\text{Tr}(\mathbf{A}) = \sum_i \mathbf{A}_{i,i}.$$

迹运算描述矩阵Frobenius 范数：

$$\|\mathbf{A}\|_F = \sqrt{\text{Tr}(\mathbf{A}\mathbf{A}^\top)}.$$

$$\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{A}^\top).$$

$$\text{Tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \text{Tr}(\mathbf{C}\mathbf{A}\mathbf{B}) = \text{Tr}(\mathbf{B}\mathbf{C}\mathbf{A}).$$

更一般地，

$$\text{Tr}\left(\prod_{i=1}^n \mathbf{F}^{(i)}\right) = \text{Tr}\left(\mathbf{F}^{(n)} \prod_{i=1}^{n-1} \mathbf{F}^{(i)}\right).$$

定义 2.15 (行列式)

$\det(\mathbf{A})$ ，将方阵 $\mathbf{A}$ 映射到实数的函数，行列式等于矩阵特征值的乘积

- 行列式绝对值，用来衡量矩阵参与矩阵乘法后空间扩大或缩小了多少
- 行列式是 0，空间至少沿着某一维完全收缩，使其失去了所有的体积
- 行列式是 1，这个转换保持空间体积不变

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots \quad \cdots \quad \cdots \quad \cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_n \end{cases}$$

可写作如下线性变换的形式

$$\mathbf{x} \in \mathbb{R}^n$$

$$\mathbf{Ax} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\mathbf{Ax} = \mathbf{b}$$

线性方程组求解:

$$\mathbf{Ax} = \mathbf{b}$$

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{I}_n\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

取决于逆矩阵 $\mathbf{A}^{-1}$ 是否存在。

当 $\mathbf{A}^{-1}$ 存在时, 求解算法有很多种。

重点

定义 2.16 (线性组合 (linear combination))

将 A 列向量看作从原点 (origin) 发的不同方向, 确定有多少种方法可以到达向量 b , 则向量 x 元素表示沿这些方向走多远, 即 x_i 表示沿着第 i 个向量方向走多远:

$$Ax = \sum_i x_i A_{:,i}. \quad (2.45)$$

一组向量的线性组合 (linear combination), 是指每个向量乘以对应标量系数之后的和, 即:

$$\sum_i c_i v^{(i)}. \quad (2.46)$$

**定义 2.17 (生成子空间 (span))**

原始向量线性组合后所能抵达的点的集合



确定 $Ax = b$ 是否有解, 相当于确定向量 b 是否在 A 列向量的生成子空间 (span), 称为 A 的列空间 (column space) 或值域 (range)。

方程 $Ax = b$ 对于任意向量 $b \in \mathbb{R}^m$ 都存在解, 则 A 的列空间, 必须构成整个 $\mathbb{R}^m$, A 至少有 m 列, 即 $n \geq m$; 否则, A 列空间的维数会小于 m 。

线性相关 (linear dependence): 列向量可能存在冗余, $n \geq m$ 是方程对每一点都有解的必要非充分条件; 若某向量是一组向量中某些向量的线性组合, 加入该向量这组向量, 不会增加这组向量的生成子空间。

线性无关 (linearly independent): 一组向量中的任一向量都不能表示成其他向量的线性组合

- 若矩阵列空间涵盖 $\mathbb{R}^m$, 则必须包含至少一组 m 个线性无关的向量, 是式 (2.36) 对每个向量 $\mathbf{b}$ 都有解的充分必要条件
- 矩阵可逆, 式 (2.36) 对每一个 $\mathbf{b}$ 值至多有一个解, 需确保该矩阵至多有 m 个列向量

综上, $\mathbf{A}$ 必须是 **方阵** (square), 即 $m = n$, 且所有列向量线性无关。

一个列向量线性相关的方阵称为 **奇异的** (singular)。若 $\mathbf{A}$ 非方阵或是奇异阵, 该方程仍可能有解, 但无法使用矩阵逆求解。

本讲内容

01

线性代数

02

微积分

03

数学优化

04

概率论

05

信息论

定义 2.19 (导数)

对定义域和值域都是实数域的函数 $f: \mathbb{R} \rightarrow \mathbb{R}$, 若 $f(x)$ 在点 x_0 的邻域 Δx 内, 极限

$$f'(x_0) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x_0)}{\Delta x} \quad (2.51)$$

存在, 则称函数 $f(x)$ 在点 x_0 处可导, $f'(x_0)$ 为其导数

在几何上, 导数可以看作函数曲线上的切线斜率

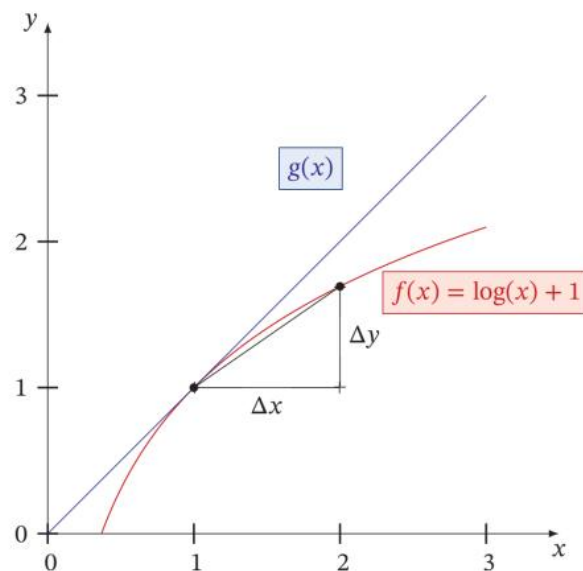


图 2.3: 函数 $f(x) = \log(x) + 1$ 导数

重点

定义 2.20 (高阶导数)

对一个函数的导数继续求导

$$f^n(x) = \frac{d^n f(x)}{dx^n} \quad (2.52)$$

**定义 2.21 (偏导数)**

对于一个多元变量函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$, 其偏导数定义为关于期中一个变量 x_i 的导数, 而其余变量固定, 记为 $\frac{\partial}{\partial x_i} f(\mathbf{x})$

**定义 2.22 (微分)**

给定一个连续函数, 计算其导数的过程称为微分

一个函数 $f(x)$ 在定义域中的所有点都存在导数, 则 $f(x)$ 为可微函数

可微函数一定连续, 但连续函数不一定可微



重点



定义 2.23 (泰勒公式)

是函数 $f(x)$ 在已知某一点的各阶导数值的情况之下, 可用这些导数值做系数构建一个多项式来近似函数在这一点邻域中的值

如果函数 $f(x)$ 在 x_0 点处 $n, n > 1$ 次可导, 在一个包含点 x_0 的区间上的任意 x , 都有

$$f(x) = f(x_0) + \frac{1}{1!}f'(x_0)(x-x_0) + \frac{1}{2!}f^{(2)}(x_0)(x-x_0)^2 + \cdots + \frac{1}{n!}f^{(n)}(x_0)(x-x_0)^n + R_n(x) \quad (2.53)$$

多项式部分称为函数 $f(x)$ 在 x_0 处的 n 阶泰勒展开式, 剩余的 $R_n(x)$ 是泰勒公式的余项, 是 $(x-x_0)^n$ 的高阶无穷小

**重点**

指数函数 e^x 在 $x = 0$ 处的泰勒展开表达式：

作答

矩阵微积分 (Matrix Calculus) 是多元微积分的一种表达方式, 即使用矩阵和向量来表示因变量每个成分关于自变量每个成分的偏导数

标量关于向量的偏导数

对于 $\mathbf{x} \in \mathbb{R}^n$ 和函数 $y = f(\mathbf{x}) \in \mathbb{R}$, y 关于 $\mathbf{x}$ 的偏导数为:

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial x_n} \right]^\top \quad (2.56)$$

向量关于标量的偏导数

对于 $x \in \mathbb{R}$ 和函数 $\mathbf{y} = f(x) \in \mathbb{R}^n$, $\mathbf{y}$ 关于 x 的偏导数为:

$$\frac{\partial \mathbf{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \dots, \frac{\partial y_n}{\partial x} \right]^\top \quad (2.57)$$

向量关于向量的偏导数

对于 $\mathbf{x} \in \mathbb{R}^n$ 和函数 $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^m$, $f(\mathbf{x})$ 关于 $\mathbf{x}$ 的偏导数为:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

为 $f(\mathbf{x})$ 的雅克比矩阵 (Jacobian Matrix)

Hessian 矩阵

对于 $\mathbf{x} \in \mathbb{R}^n$ 和函数 $y = f(\mathbf{x}) \in \mathbb{R}$, $f(\mathbf{x})$ 关于 $\mathbf{x}$ 的二阶偏导数为:

$$\mathbf{H} = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}^2} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1^2} & \dots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 y}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

加减法则

若 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^n, \mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^n$, 则

$$\frac{\partial(\mathbf{y} + \mathbf{z})}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{m \times n}$$

乘法法则

1. 若 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^n, \mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^n$, 则

$$\frac{\partial \mathbf{y}^\top \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \mathbf{z} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \mathbf{y} \in \mathbb{R}^m$$

2. 若 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^s, \mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^t, \mathbf{A} \in \mathbb{R}^{s \times t}$ 与 $\mathbf{x}$ 无关, 则

$$\frac{\partial \mathbf{y}^\top \mathbf{A} \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \mathbf{A} \mathbf{z} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \mathbf{A}^\top \mathbf{y} \in \mathbb{R}^m$$

3. 若 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} = f(\mathbf{x}) \in \mathbb{R}, \mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^n$, 则

$$\frac{\partial \mathbf{y} \mathbf{z}}{\partial \mathbf{x}} = \mathbf{y} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \mathbf{z}^\top \in \mathbb{R}^{m \times n}$$

重点

链式法则

在微积分中求复合函数导数的一种常用方法

1. 若 $x \in \mathbb{R}, \mathbf{y} = g(x) \in \mathbb{R}^m, \mathbf{z} = f(\mathbf{y}) \in \mathbb{R}^n$, 则

$$\frac{\partial \mathbf{z}}{\partial x} = \frac{\partial \mathbf{y}}{\partial x} \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \in \mathbb{R}^{1 \times n}$$

2. 若 $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} = g(\mathbf{x}) \in \mathbb{R}^k, \mathbf{z} = f(\mathbf{y}) \in \mathbb{R}^n$, 则

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \in \mathbb{R}^{m \times n}$$

3. 若 $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{y} = g(\mathbf{X}) \in \mathbb{R}^k, \mathbf{z} = f(\mathbf{y}) \in \mathbb{R}$, 则

$$\frac{\partial \mathbf{z}}{\partial x_{ij}} = \frac{\partial \mathbf{y}}{\partial x_{ij}} \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \in \mathbb{R}$$

函数 $\sigma(x) = \frac{1}{1 + e^{-x}}$ 一阶导数 (请用 (x) 表示)

作答



计算机编程如何求导？

本讲内容

01

线性代数

02

微积分

03

数学优化

04

概率论

05

信息论

- 数学优化 (Mathematical Optimization) 问题, 也叫最优化问题, 是指在一定约束条件下, 求解一个目标函数的最大值 (或最小值) 问题

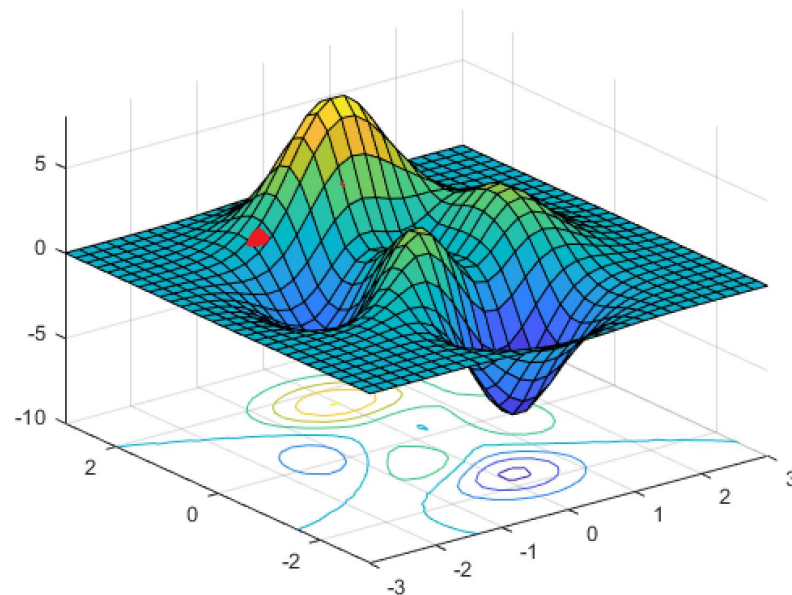
定义 2.24 (数学优化)

给定一个目标函数 (代价函数) $f: \mathbb{A} \rightarrow \mathbb{R}$, 寻找一个变量 (参数) $\mathbf{x}^* \in \mathbb{D} \subset \mathbb{A}$, 使得对于所有 $\mathbb{D}$ 中的 $\mathbf{x}$, 都满足 $f(\mathbf{x}^*) \leq f(\mathbf{x})$ (最小化), 或 $f(\mathbf{x}^*) \geq f(\mathbf{x})$ (最大化), 其中 $\mathbb{D}$ 为变量 $\mathbf{x}$ 的约束集, $\mathbb{D}$ 中的变量称为可行解。



- 很多非线性优化问题, 会存在若干个局部最小值, 其存在一个对应解称为局部最小解

为什么最优化一般是求最小值, 而不是最大值



定义 2.25 (局部最小解)

存在一个 $\delta > 0$, 对于所有满足 $\|x - x^*\| \leq \delta$ 的 x , 都有 $f(x^*) \leq f(x)$

**定义 2.26 (全局最小解)**

对于所有 $x \in \mathbb{D}$, 都有 $f(x^*) \leq f(x)$

**定理 2.1 (局部最小解的一阶必要条件)**

若 x^* 为局部最小解, 且函数 f 在 x^* 的邻域内一阶可微, 则 $\nabla f(x^*) = 0$

**定理 2.2 (局部最小解的二阶必要条件)**

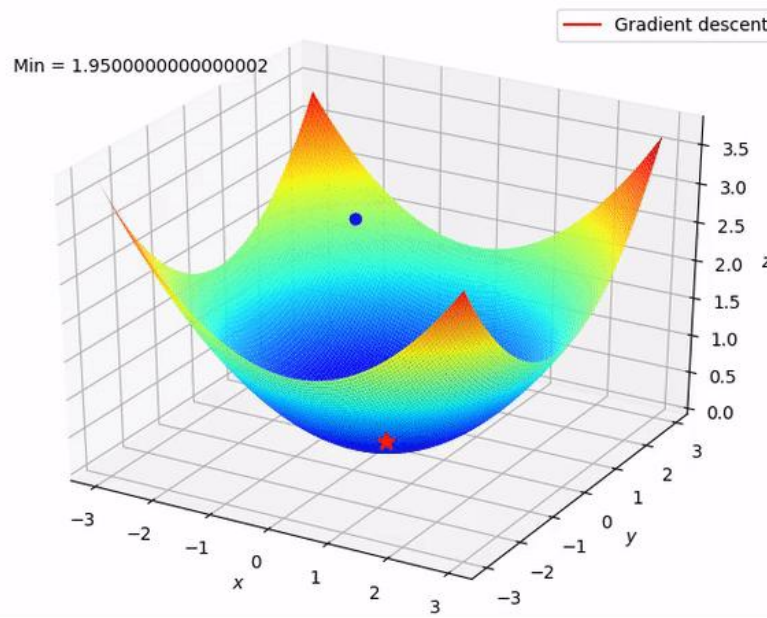
若 x^* 为局部最小解, 且函数 f 在 x^* 的邻域内二阶可微, 则 $\nabla f(x^*) = 0$, 且 $\nabla^2 f(x^*)$ 为半正定矩阵



- **梯度下降法** (Gradient Descent Method), 即**最速下降法** (Steepest Descend Method), 常用来求解无约束优化的最小值问题

梯度下降

1. 对于函数 $f(\mathbf{x})$, 在点 $\mathbf{x}_0$ 附近连续可微, 那么 $f(\mathbf{x})$ 下降最快的方向是 $f(\mathbf{x})$ 在 $\mathbf{x}_0$ 点的梯度方向的反方向
2. 梯度下降法为一阶收敛算法, 当靠近局部最小解时梯度变小, 收敛速度会变慢, 并且可能以“之字形”的方式下降
3. 如果目标函数为二阶连续可微, 可采用牛顿法
4. 牛顿法 (Newton's method) 为二阶收敛算法, 收敛速度更快, 但是每次迭代需要计算 Hessian 矩阵的逆矩阵, 复杂度较高



重点

本讲内容

01

线性代数

02

微积分

03

数学优化

04

概率论

05

信息论

- 概率论用于表示不确定性的数学框架
 - 提供了量化不确定性的方法
 - 提供了导出新不确定性 声明 (statement) 的公理
- 在人工智能领域，概率论主要用于：
 - 概率法则解释AI系统如何推理，据此设计算法计算或估算，由概率论导出的表达式
 - 用概率和统计，从理论上分析AI系统行为

概率可以被看作是用于处理不确定性的逻辑扩展。

逻辑提供了一套形式化的规则，在某些命题为真或假的条件下，判断另外一些命题的真假

概率论提供了一套形式化的规则，在已知某些命题的似然，计算其他命题为真的似然

- 机器学习通常必须处理不确定量，不确定性和随机性可能来自多个方面。
 - 被建模系统**内在随机性**
 - **不完全观测**。即使是确定系统，当不能观测到所有驱动系统行为的变量时，系统也会呈现随机性
 - **不完全建模**。使用一些必须舍弃某些观测信息的模型时，舍弃信息会导致模型预测出现不确定性

- **随机变量** (random variable) 是可以随机地取不同值的变量, 用小写字母表示。
 - 随机变量可以是离散或连续的, 离散随机变量有有限或可数无限多的状态, 连续随机变量为实数值。
一个随机变量只是对可能状态的描述, 需要概率分布来确定每个状态的可能性。
- **概率分布** (probability distribution) 描述随机变量或一簇随机变量每个状态的可能性大小

离散型变量和概率分布律

离散型变量的概率分布用概率分布律来描述, 用大写字母 P 来表示

- P 定义域必须是 x 所有可能状态的集合
- $\forall x \in x, 0 \leq P(x) \leq 1.$
- $\sum_{x \in x} P(x) = 1.$

多个变量分布律称为 **联合概率分布** (joint probability distribution), $P(x = x, y = y)$ 表示 $x = x$ 和 $y = y$ 同时发生的概率, 简写为 $P(x, y)$ 。

连续型变量和概率密度函数

连续型随机变量用 **概率密度函数** (probability density function, PDF) 描述概率分布

- p 的定义域必须是 x 所有可能状态的集合
- $\forall x \in x, p(x) \geq 0.$ 注意, 并不要求 $p(x) \leq 1.$
- $\int p(x)dx = 1.$

概率密度函数 $p(x)$ 并未直接对特定的状态给出概率, 给出了落在面积为 δx 的无限小的区域内的概率为 $p(x)\delta x$ 。

可对概率密度函数求积分来获得点集的真实概率分布。如, x 落在区间 $[a, b]$ 的概率是 $\int_{[a,b]} p(x)dx$ 。

- 已知一组变量联合概率分布，其中一个子集的概率分布，称为**边缘概率分布**

离散型随机变量 x 和 y ，且已知 $P(x, y)$ ：

$$\forall x \in x, P(x = x) = \sum_y P(x = x, y = y).$$

连续型变量：

$$p(x) = \int p(x, y) dy.$$

条件概率

某个事件，在给定其他事件发生时出现的概率。

给定 $x = x$ ， $y = y$ 发生的条件概率记为 $P(y = y | x = x)$ 。

$$P(y = y | x = x) = \frac{P(y = y, x = x)}{P(x = x)}. \quad (2.69)$$

概率 **链式法则** (chain rule) 或 **乘法法则** (product rule)：多维随机变量的联合概率分布，可分解为单变量条件概率相乘的形式：

$$P(x^{(1)}, \dots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^n P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)}). \quad (2.70)$$

独立性和条件独立性

随机变量 x 和 y **相互独立的** (independent)：

$$\forall x \in x, y \in y, p(x = x, y = y) = p(x = x)p(y = y). \quad (2.71)$$

x 和 y 在给定随机变量 z 时 **条件独立的** (conditionally independent)：

$$\forall x \in x, y \in y, z \in z, p(x = x, y = y | z = z) = p(x = x | z = z)p(y = y | z = z). \quad (2.72)$$

$x \perp y$ 表示 x 和 y 相互独立， $x \perp y | z$ 表示 x 和 y 在给定 z 时条件独立。

定义 2.27 (期望 (expectation))

函数 $f(x)$ 关于分布 $P(x)$ 的期望 (expectation), 当 x 由 P 产生, f 作用于 x 时, $f(x)$ 的平均值。离散型随机变量:

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x)f(x), \quad (2.73)$$

连续型随机变量:

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x)f(x)dx. \quad (2.74)$$


 重点

期望是线性的:

$$\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)], \quad (2.75)$$

其中 α 和 β 不依赖于 x 。

定义 2.28 (方差 (variance))

依据 x 的概率分布采样时, x 函数值的差异:

$$\text{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2].$$

定义 2.29 (协方差 (covariance))

给出了两个变量线性相关性程度

$$\text{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])].$$

- 两个变量互独立, 协方差为零
 - 两个变量协方差不为零, 一定相关
 - 两个变量协方差为零, 一定没有线性关系, 但不一定相互独立
- 随机向量 $\mathbf{x} \in \mathbb{R}^n$ 的 **协方差矩阵** (covariance matrix) 是一个 $n \times n$ 的矩阵, 且满足

$$\text{Cov}(\mathbf{x})_{i,j} = \text{Cov}(x_i, x_j).$$

协方差矩阵对角元是方差:

$$\text{Cov}(x_i, x_i) = \text{Var}(x_i).$$

Bernoulli 分布

二值随机变量分布，由单参数 $\phi \in [0, 1]$ 控制， ϕ 给出了随机变量等于 1 的概率：

$$P(x = 1) = \phi \quad (2.80)$$

$$P(x = 0) = 1 - \phi \quad (2.81)$$

$$P(x = x) = \phi^x (1 - \phi)^{1-x} \quad (2.82)$$

$$\mathbb{E}_x[x] = \phi \quad (2.83)$$

$$\text{Var}_x(x) = \phi(1 - \phi) \quad (2.84)$$

Multinoulli 分布

Multinoulli 分布 (multinoulli distribution) 指有 k 个不同状态的单个离散型随机变量上的分布，其中 k 是一个有限值

Multinoulli 分布由向量 $\mathbf{p} \in [0, 1]^{k-1}$ 参数化，其中每一个分量 p_i 表示第 i 个状态的概率
第 k 个状态的概率可以通过 $1 - \mathbf{1}^\top \mathbf{p}$ 给出

高斯分布

正态分布 (normal distribution)，也称为 **高斯分布** (Gaussian distribution)：

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (2.85)$$

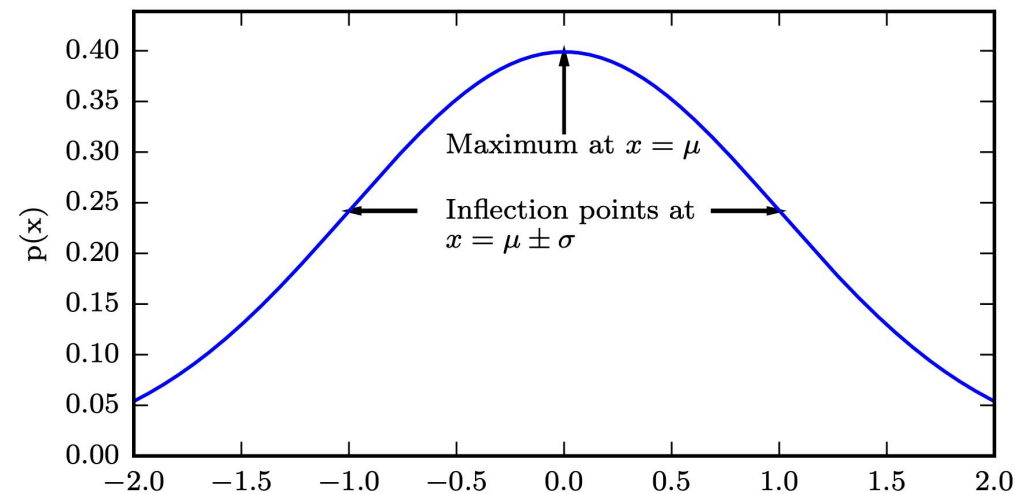
由 $\mu \in \mathbb{R}$ 和 $\sigma \in (0, \infty)$ 控制

μ 给出了峰值坐标，即分布均值： $\mathbb{E}[x] = \mu$

σ 表示分布标准差， σ^2 表示方差

当缺乏关于某分布先验知识时，正态分布是默认的较好的选择：

- 待建模的很多分布比较接近正态分布。**中心极限定理** (central limit theorem) 表明独立随机变量和近似服从正态分布。
- 方差相同的所有概率分布中，正态分布不确定性最大。可以认为正态分布是对模型加入先验知识量最少的分布。



正态分布 $N(x; \mu, \sigma^2)$ 呈钟形曲线，峰值位置 $x = \mu$ ，峰宽受 σ 控制。

重点

logistic sigmoid

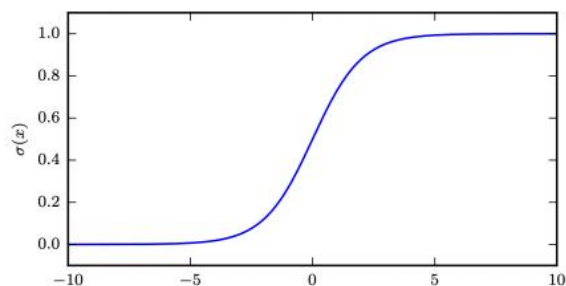
$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \quad (2.87)$$

$x \rightarrow +\infty$ 或 $x \rightarrow -\infty$, 出现 **饱和** (saturate), 对输入微小改变不敏感

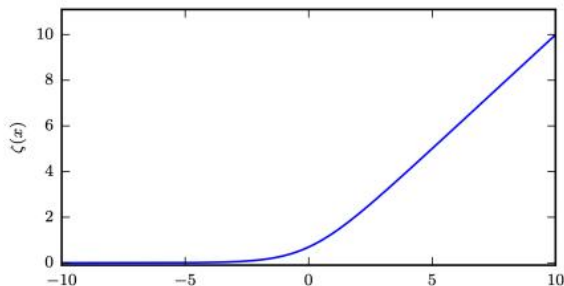
softplus 函数 (softplus function)

$$x^+ = \max(0, x). \quad (2.88)$$

函数的平滑 (或“软化”) 形式



(a) logistic sigmoid函数



(b) softplus 函数。

重点

常用函数性质

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\zeta(-x)$$

$$\frac{d}{dx}\zeta(x) = \sigma(x)$$

$$\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

$$\forall x > 0, \zeta^{-1}(x) = \log(\exp(x) - 1)$$

$$\zeta(x) = \int_{-\infty}^x \sigma(y) dy$$

$$\zeta(x) - \zeta(-x) = x$$

函数 $\sigma^{-1}(x)$ 在统计学中被称为 **分对数** (logit), 但这个函数在机器学习中很少用到。

本讲内容

01

线性代数

02

微积分

03

数学优化

04

概率论

05

信息论

- 是应用数学的一个分支，主要研究信号信息量
- 是电子工程和计算机科学中许多领域的基础
- 基本原理是一个不太可能的事件居然发生了，比一个非常可能的事件发生，能提供更多的信息

信息论

- 非常可能发生的事件信息量要比较少，确定发生的事件无信息量
- 较不可能发生的事件具有更高的信息量
- 独立事件具有增量信息，如投掷硬币两次正面朝上的信息量，是投掷一次硬币正面朝上的两倍

定义事件 $x = x$ 的 **自信息** (self-information)

$$I(x) = -\log P(x). \quad (2.99)$$

定义 2.32 (信息熵)

对整个概率分布中的不确定性总量进行量化

$$H(x) = \mathbb{E}_{x \sim P}[I(x)] = -\mathbb{E}_{x \sim P}[\log P(x)], \quad (2.100)$$

一个分布的信息熵是指遵循该分布的事件所产生的期望信息总量



信息熵

1. 熵值越大，确定性越低，混乱程度越大，随机变量的信息越多
2. 熵值越小，确定性越高，混乱程度越小，随机变量的信息越少
3. 信息是确定性的增加
4. 对于一个确定的信息，其熵为 0，信息量也为 0
5. 一个变量概率分布为均匀分布，则熵最大

重点

如何理解熵？

对于两个离散随机变量 $x \in \mathbb{X}$ 和 $y \in \mathbb{Y}$ ，其联合概率分布满足为 $p(x, y)$ ，则

定义 2.33 (联合熵)

$$H(x, y) = -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} p(x, y) \log p(x, y) \quad (2.101)$$



定义 2.34 (条件熵)

$$\begin{aligned} H(x|y) &= -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} p(x, y) \log p(x|y) \\ &= -\sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} \\ &= H(x, y) - H(y) \end{aligned} \quad (2.102)$$



互信息 (Mutual Information) 是衡量已知一个变量时, 另一个变量不确定性的减少程度

定义 2.35 (互信息)

$$I(x, y) = - \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2.103)$$

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x) \quad (2.104)$$

如果变量 x 和 y 互相独立, 它们的互信息为零

定义 2.36 (交叉熵 (cross-entropy))

对于分布为 $p(x)$ 的随机变量, 熵 $H(p)$ 表示其最优编码长度。交叉熵 (Cross Entropy) 是按照概率分布 q 的最优编码对真实分布为 p 的信息进行编码的长度, 定义为

$$H(p, q) = -\mathbb{E}_{x \sim p} \log q(x) \quad (2.105)$$

给定 p , p 和 q 越接近, 交叉熵越小; 越远, 交叉熵越大

定义 2.37 (KL 散度 (Kullback-Leibler (KL) divergence))

衡量同一随机变量 x 两个单独的概率分布 $p(x)$ 和 $q(x)$ 分布差异:

$$D_{\text{KL}}(p||q) = \mathbb{E}_{x \sim p} \left[\log \frac{p(x)}{q(x)} \right] = \mathbb{E}_{x \sim p} [\log p(x) - \log q(x)]. \quad (2.106)$$

KL 散度非负

KL 散度为 0 当且仅当 p 和 q 在离散型变量情况下同分布, 或在连续型变量情况下“几乎处处”相同

$$D_{\text{KL}}(p||q) \neq D_{\text{KL}}(q||p)$$

重点



中山大學

SUN YAT-SEN UNIVERSITY

谢 谢