第三章 度量与投影

第 6 讲 内积与范数：数据度量的观点

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

1. 6.1 向量范数
2. 6.2 内积、距离、夹角与正交性
3. 6.3 数据科学中常用的相似性度量 I
4. 6.4 矩阵的内积与范数
5. 6.5 范数在机器学习中的应用

内积和范数引例：文本聚类

在第 2.1.1 节中，我们对纽约时报在 2010 年 12 月 7 日的四则新闻提要都进行了向量化的表示，我们希望知道这四则新闻提要有哪些是相关的，可以通过对这四则新闻提要进行简单聚类来实现:

(a) Suit Over Targeted Killing in Terror Case Is Dismissed $\cdots$

(b) In Tax Deal With G.O.P, a Portent for the Next 2 Years $\cdots$

(c) Obama Urges China to Check North Koreans $\cdots$

(d) Top Test Scores From Shanghai Stun Educators $\cdots$

$$sim_{cos}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|},$$

其中 $\boldsymbol{x}, \boldsymbol{y}$ 是文本向量，$sim_{cos}(\boldsymbol{x}, \boldsymbol{y})$ 表示 $\boldsymbol{x}, \boldsymbol{y}$ 的余弦相似度。

内积和范数引例：手写数字分类

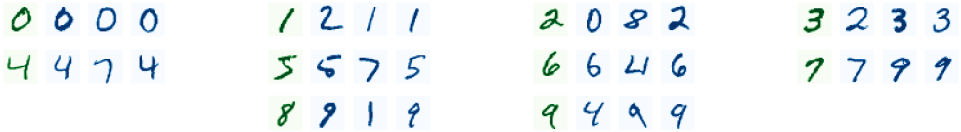


图 1: 对 MINST 数据集进行分类（绿色的为训练集，蓝色的为测试集）

$$d(A,B)=\sum_{jk}|\boldsymbol{A}_{jk}-\boldsymbol{T}_{jk}|$$

其中 d 是手写数字训练图片的表示矩阵 $\boldsymbol{A}$ 和测试图片的表示矩阵 $\boldsymbol{T}$ 之间的距离（两个矩阵同等大小），j,k 取遍矩阵所有元素。

6.1.1 向量范数：复数的模

例 1

复数 $\boldsymbol{x} = (a, b) = a + ib$ 的长度或者**模**指的是

$$||\boldsymbol{x}|| = \sqrt{a^2 + b^2}$$

显然复向量 $\boldsymbol{x}$ 的模 $||\boldsymbol{x}||$ 具有下列三条性质：

(1) $||\boldsymbol{x}|| \geq 0$，当且仅当 $\boldsymbol{x} = 0$ 时等号成立；

(2) $||\lambda\boldsymbol{x}|| = |\lambda|||\boldsymbol{x}||; (\forall\lambda \in \mathbb{R})$

(3) $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||.(\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C})$

向量的模

例 2

n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)\in\mathbb{R}^n$ 的模或长度定义为

$$||\boldsymbol{x}||=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$$

显然向量 $\boldsymbol{x}$ 的模 $||\boldsymbol{x}||$ 也具有下列三条性质：

(1) $||\boldsymbol{x}||\geq 0$，当且仅当 $\boldsymbol{x}=0$ 时等号成立；

(2) $||\lambda\boldsymbol{x}||=|\lambda|||\boldsymbol{x}||;(\forall\lambda\in\mathbb{R})$

(3) $||\boldsymbol{x}+\boldsymbol{y}||\leq||\boldsymbol{x}||+||\boldsymbol{y}||.(\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n)$

向量范数的定义

定义 1

设 $\mathbb{V}$ 是数域上 $\mathbb{K}$ 的 n 维线性空间，函数

$$\begin{aligned} ||\cdot||: \mathbb{V} &\to \mathbb{R}, \\ \boldsymbol{x} &\mapsto \|\boldsymbol{x}\|, \end{aligned}$$

它把向量 $\boldsymbol{x}$ 映射为它的长度 $\|\boldsymbol{x}\| \in \mathbb{R}$，并且使得对 $\forall \lambda \in \mathbb{R}$ 和 $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$，满足

(1) 非负性：$\|\boldsymbol{x}\| \geq 0$，$\|\boldsymbol{x}\| = 0$ 当且仅当 $\boldsymbol{x} = 0$；

(2) 齐次性：$\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|$；

(3) 三角不等式：$\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$；

称 $\|\boldsymbol{x}\|$ 是向量 $\boldsymbol{x}$ 的**向量范数**，称定义了范数的线性空间 $\mathbb{V}$ 为**赋范线性空间**。

例 3

对任给的 $\boldsymbol{x}=(x_1,x_2,x_3)^{\mathrm{T}}\in\mathbb{C}^3$，试问如下实值函数是否构成向量范数？

1. $|x_1|+|2x_2+x_3|$,
2. $|x_1|+|2x_2|-5|x_3|$.

解

1. 非负性：$|x_1|+|2x_2+x_3|\geq 0$;
   齐次性：令 $c\in\mathbb{C}$,$|cx_1|+|2cx_2+cx_3|=|c|(|x_1|+|2x_2+x_3|)$;
   三角不等式：令 $\boldsymbol{x}=(x_1,x_2,x_3)^{\mathrm{T}},\boldsymbol{y}=(y_1,y_2,y_3)^{\mathrm{T}}\in\mathbb{C}^3$，则
   $|x_1+y_1|+|2(x_2+y_2)+(x_3+y_3)|\leq|x_1|+|2x_2+x_3|+|y_1|+|2y_2+y_3|$.
2. 取 $\boldsymbol{x}=(0,0,1)$ 则 $|0|+|2\times 0|-5|1|=-5<0$ 不满足非负性。

6.1.2 几种常用的向量范数：l_p 范数

例 4

对于任意 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)\in\mathbb{R}^n$，由

$$\|\boldsymbol{x}\|_p=(\sum_{i=1}^{n}|x_i|^p)^{\frac{1}{p}},1\leq p<\infty$$

定义的 $||\cdot||_p$ 是 $\mathbb{R}^n$ 上的向量范数，称为p **范数**或 l_p 范数。

(1) 当 $p=1$ 时，得到 **1 范数**或 l_1 范数，也称为 Manhattan 范数

$$\|\boldsymbol{x}\|_1=\sum_{i=1}^{n}|x_i|$$

(2) 当 $p=2$ 时, 得到 **2 范数**或 l_2 范数，也称为欧几里得范数

$$\|\boldsymbol{x}\|_2=\sqrt{\sum_{i=1}^{n}x_i^2}$$

l_∞ 范数

在例4中，在广义实数范围内，p 能否取到正无穷大呢？具体而言，如何计算这种范数呢？

例 5

对于任意 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$，由

$$\|\boldsymbol{x}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{x}\|_p,$$

也就是，

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\cdots,n} |x_i|,$$

定义的 $||\cdot||_\infty$ 是 $\mathbb{R}^n$ 上的向量范数，称为∞ **范数**或 l_∞ 范数。

l_∞ 范数

证明.

验证 $\|\boldsymbol{x}\|_\infty \equiv \max\limits_i |\boldsymbol{x}_i|$ 是向量范数显然很容易。下证 $\max\limits_i |\boldsymbol{x}_i| = \lim\limits_{p\to+\infty} \|\boldsymbol{x}\|_p$。令 $\|\boldsymbol{x}_j\| = \max\limits_j |\boldsymbol{x}_i|$，则有

$$\begin{aligned}\|\boldsymbol{x}\|_\infty = |\boldsymbol{x}_j| &\le (\sum_{i=1}^{n} |\boldsymbol{x}_i|^p)^{(1/p)} = \|\boldsymbol{x}\|_p \\ &\le (n|\boldsymbol{x}_j|^p)^{(1/p)} = n^{(1/p)}\|\boldsymbol{x}\|_\infty\end{aligned}$$

由极限的夹逼准则，并注意到 $\lim\limits_{p\to+\infty} n^{1/p} = 1$，即得欲证结论。 □

非向量范数

例 6

当 $0 < p < 1$，由

$$\|\boldsymbol{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$$

定义的 $\|\cdot\|_p$ 不是 $\mathbb{R}^n$ 上的向量范数。

证明.

考虑 $n = 2, p = \frac{1}{2}$. 取 $\boldsymbol{\alpha} = (1, 0)^{\mathrm{T}}, \boldsymbol{\beta} = (0, 1)^{\mathrm{T}}$, 则

$$\|\boldsymbol{\alpha}\|_{\frac{1}{2}} = \|\boldsymbol{\beta}\|_{\frac{1}{2}} = 1, \|\boldsymbol{\alpha} + \boldsymbol{\beta}\|_{\frac{1}{2}} = 4$$

$$\|\boldsymbol{\alpha} + \boldsymbol{\beta}\|_{\frac{1}{2}} \geq \|\boldsymbol{\alpha}\|_{\frac{1}{2}} + |\boldsymbol{\beta}\|_{\frac{1}{2}}$$

基数函数：l_0 范数

定义 2

向量 $\boldsymbol{x}$ 的**基数函数**定义为 $\boldsymbol{x}$ 中非零元素的个数，即

$$card(\boldsymbol{x}) = \sum_{i=1}^{n} \mathcal{I}(x_i \neq 0)$$

其中，

$$\mathcal{I}(x_i \neq 0) = \begin{cases} 1 & , x_i \neq 0 \\ 0 & , x_i = 0 \end{cases}$$

基数函数也被称为 l_0 范数，但是它并不满足范数定义的条件。

例 7

求向量 $\boldsymbol{x} = (-1, 2, 4)^{\mathrm{T}}$ 的 0，1，2，和 ∞-范数。

解

$$||\boldsymbol{x}||_0 = 3$$
$$||\boldsymbol{x}||_1 = |-1| + 2 + 4 = 7$$
$$||\boldsymbol{x}||_2 = \sqrt{|-1|^2 + 2^2 + 4^2} = \sqrt{21}$$
$$||\boldsymbol{x}||_\infty = \max\{|-1|, 2, 4\} = 4$$

6.1.3 范数的几何意义：单位范数球

定义 3

对于 l_p 范数小于等于 1 的向量集合，

$$\mathcal{B}_p = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_p \leq 1\}$$

称为 l_p 的**单位范数球**。

范数的几何意义

例 8

单位范数球的形状反映了不同范数的性质，对于不同的 p，范数球有着不同的几何形状。图2分别表示了 $\mathcal{B}_2, \mathcal{B}_1, \mathcal{B}_\infty$ 在 $\mathbb{R}^2$ 的范数球形状。

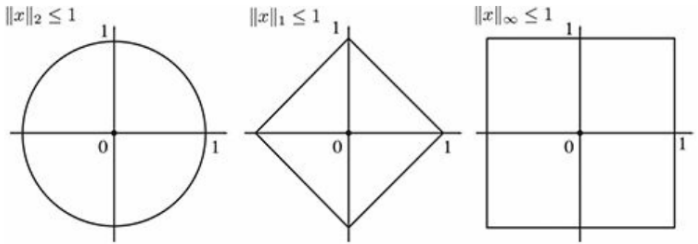


图 2: $\mathbb{R}^2$ 上的范数球

6.1.4 范数性质

定义 4

设 $\{\boldsymbol{x}^{(k)}\}$ 为 $\mathbb{R}^n$ 中一向量序列，$\boldsymbol{x}^* \in \mathbb{R}^n$，其中

$$\boldsymbol{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \cdots, x_n^{(k)})^{\mathrm{T}}, \boldsymbol{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)^{\mathrm{T}}$$

如果 $\lim_{k\to\infty} x_i^{(k)} = x_i^* (i = 1, 2, \cdots, n)$，则称 $\boldsymbol{x}^{(k)}$ **收敛**于向量 x^*，记作

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*$$

范数的连续性和收敛性质

定理 1

（范数的连续性）设非负函数 $N(\boldsymbol{x}) = \|\boldsymbol{x}\|$ 为 $\mathbb{R}^n$ 上任一向量范数，则 $N(\boldsymbol{x})$ 是 $\boldsymbol{x}$ 分量 $x_1, x_2, \cdots, x_n$ 的连续函数。

定理 2

（柯西收敛原理）实数域 $\mathbb{R}$（或者复数域 $\mathbb{C}$）上的有限维线性空间按任何范数 $\|\cdot\|$ 必定完备。

定理 3

（向量序列收敛定理）设 $\{\boldsymbol{x}^{(k)}\}$ 为 $\mathbb{R}^n$ 中一向量序列，$\boldsymbol{x}^* \in \mathbb{R}^n$ 则

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}^* \Longleftrightarrow \lim_{k\to\infty} \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\right\| = 0$$

其中 $\|\cdot\|$ 为向量的任一种范数。

范数的等价性

在 $\mathbb{R}^n$（或 $\mathbb{C}^n$）上可以定义各种向量范数，其数值大小一般不同，但是在各种向量范数之间存在下述重要的关系

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_1 \le n\|\boldsymbol{x}\|_\infty$$
$$\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_1 \le \|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1$$
$$\frac{1}{\sqrt{n}}\|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2$$

或者

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \|\boldsymbol{x}\|_1 \le \sqrt{n}\|\boldsymbol{x}\|_2 \le n\|\boldsymbol{x}\|_\infty$$

范数的等价性

定理 4

（向量范数的等价性定理）设 $\|\boldsymbol{x}\|_s, \|\boldsymbol{x}\|_t$ 为 $\mathbb{R}^n$ 上的任意两种向量范数，则存在两个与向量无关的正常数 $c_1, c_2 > 0$，使得下面的不等式成立

$$c_1 \|\boldsymbol{x}\|_s \leq \|\boldsymbol{x}\|_t \leq c_2 \|\boldsymbol{x}\|_s, \quad \text{对一切 } \boldsymbol{x} \in \mathbb{R}^n$$

并称 $\|\boldsymbol{x}\|_t$ 和 $\|\boldsymbol{x}\|_s$ 为 $\mathbb{R}^n$ 上的等价范数。

6.2.1 内积：点积的定义

定义 5

n 维实向量空间 $\mathbb{R}^n$ 的**标准内积 (点积)** 是两个向量的对应元素乘积之和，即

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

通常我们指内积都是指这种标准内积。

内积的定义

定义 6

设 $\mathbb{V}$ 是数域上 $\mathbb{K}$ 的 n 维线性空间，函数 $\langle\cdot,\cdot\rangle:\mathbb{V}\times\mathbb{V}\to\mathbb{R}$, 它把向量 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{V}$ 映射为一个实数 $\langle\boldsymbol{x},\boldsymbol{y}\rangle\in\mathbb{R}$，并且使得对 $\forall\lambda\in\mathbb{R}$ 和 $\forall\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in\mathbb{V}$，满足

(1) 非负性：$\langle\boldsymbol{x},\boldsymbol{x}\rangle\geq 0$，$\langle\boldsymbol{x},\boldsymbol{x}\rangle=0$ 当且仅当 $\boldsymbol{x}=0$;

(2) 对称性：$\langle\boldsymbol{x},\boldsymbol{y}\rangle=\langle\boldsymbol{y},\boldsymbol{x}\rangle$;

(3) 齐次性：$\langle\lambda\boldsymbol{x},\boldsymbol{y}\rangle=\lambda\langle\boldsymbol{x},\boldsymbol{y}\rangle$;

(4) 线性性：$\langle\boldsymbol{x}+\boldsymbol{y},\boldsymbol{z}\rangle=\langle\boldsymbol{x},\boldsymbol{z}\rangle+\langle\boldsymbol{y},\boldsymbol{z}\rangle$。

称 $\langle\boldsymbol{x},\boldsymbol{y}\rangle$ 是向量 $\boldsymbol{x},\boldsymbol{y}$ 的内积，称定义了内积的线性空间 $\mathbb{V}$ 为**内积空间**。若内积是点积时，称定义了标准内积的线性空间为**欧氏空间**。

内积不是点积

例 9

考虑 $\mathbb{V} = \mathbb{R}^2$. 如果我们定义

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := x_1 y_1 - (x_1 y_2 + x_2 y_1) + 2 x_2 y_2$$

则 $\langle \cdot, \cdot \rangle$ 是一个内积，但不是点积。

6.2.2 对称、正定矩阵的内积表示

- 考虑一个定义了内积的 n 维线性空间 $\mathbb{V}$ 以及其上的内积 $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ 和有序基底 $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$。对任意的 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$，可以用基向量线性表出，也即 $\boldsymbol{x} = \sum_{i=1}^n \psi_i \boldsymbol{b}_i \in \mathbb{V}$ 以及 $\boldsymbol{y} = \sum_{j=1}^n \lambda_j \boldsymbol{b}_j \in \mathbb{V}$。
- 由内积的线性性，可得 $\boldsymbol{x}, \boldsymbol{y}$ 的内积

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \sum_{i=1}^n \psi_i \boldsymbol{b}_i, \sum_{j=1}^n \lambda_j \boldsymbol{b}_j \rangle = \sum_{i=1}^n \sum_{j=1}^n \psi_i \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle \lambda_j = \hat{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{A} \hat{\boldsymbol{y}}$$

  其中 $\boldsymbol{A}_{ij} := \langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle, \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ 分别是 $\boldsymbol{x}, \boldsymbol{y}$ 的坐标。
- 所以内积是被 $\boldsymbol{A}$ 唯一决定了，而内积的对称性决定了 $\boldsymbol{A}$ 也是对称的。
- 进一步地，由内积的非负性可得

$$\forall \boldsymbol{x} \in \mathbb{V} \setminus \{\boldsymbol{0}\} : \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} > 0.$$

对称、正定矩阵的内积表示

定义 7

若对称矩阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 满足

$$\forall \boldsymbol{x} \in \mathbb{V} \setminus \{\boldsymbol{0}\} : \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} > 0$$

则我们称 $\boldsymbol{A}$ 对称正定矩阵或正定矩阵；若满足

$$\forall \boldsymbol{x} \in \mathbb{V} \setminus \{\boldsymbol{0}\} : \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} \geq 0,$$

则我们称 $\boldsymbol{A}$ 对称正半定矩阵或正半定矩阵。

对称、正定矩阵的内积表示

例 10

考虑下列矩阵

$$\boldsymbol{A}_1 = \begin{pmatrix} 9 & 6 \\ 6 & 5 \end{pmatrix}, \boldsymbol{A}_2 = \begin{pmatrix} 9 & 6 \\ 6 & 3 \end{pmatrix}$$

则 $\boldsymbol{A}_1$ 是对称正定矩阵。因为他是对称的且对于任意的 $\boldsymbol{x} \in \mathbb{V} \setminus \{\boldsymbol{0}\}$ 有

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}_1 \boldsymbol{x} = \begin{pmatrix} x_1, x_2 \end{pmatrix} \begin{pmatrix} 9 & 6 \\ 6 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ \boldsymbol{x}_2 \end{pmatrix} = 9\boldsymbol{x}_1^2 + 12x_1x_2 + 5x_2^2 = (3x_1 + 2x_2)^2 + x_2^2 > 0$$

$\boldsymbol{A}_2$ 对称的但不正定。因为 $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}_2 \boldsymbol{x} = 9x_1^2 + 12x_1x_2 + 3x_2^2 = (3x_1 + 2x_2)^2 - x_2^2$ 可以小于 0。（比如 $\boldsymbol{x} = (2, -3)^{\mathrm{T}}$ 时）

对称、正定矩阵的内积表示

如果 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是对称、正定的，则

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{A} \hat{\boldsymbol{y}}$$

定义了一个关于有序基底 B 的内积，其中 $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ 是 $\mathbb{V}$ 中向量 $\boldsymbol{x}, \boldsymbol{y}$ 关于 B 下的坐标。

定理 5

对于一个实值有限维空间 $\mathbb{V}$ 和 $\mathbb{V}$ 下一个有序基底 B，如果 $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ 是一个内积当且仅当存在一个对称、正定矩阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 满足

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \hat{\boldsymbol{x}}^{\mathrm{T}} \boldsymbol{A} \hat{\boldsymbol{y}}$$

对称、正定矩阵的内积表示

如果矩阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是对称正定矩阵，则

- $\boldsymbol{A}$ 的核（零空间）只包含 $\boldsymbol{0}$ 因为 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} > 0$ 对于任意 $\boldsymbol{x} \neq 0$ 成立，即如果 $\boldsymbol{x} \neq 0$ 则 $\boldsymbol{A}\boldsymbol{x} \neq 0$
- $\boldsymbol{A}$ 的对角元是正的，因为 $a_{ii} = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{e}_i > 0$，其中 e_i 是 $\mathbb{R}^n$ 中的标准基

6.2.3 内积导出的范数

定义 8

设 $\mathbb{V}$ 是内积空间，则由

$$||\boldsymbol{x}|| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}, \quad \forall \boldsymbol{x} \in \mathbb{V}$$

定义的函数 $||\cdot||$ 是 $\mathbb{V}$ 上的向量范数，称为由内积 $\langle \cdot, \cdot \rangle$ 导出的范数。

- 并不是每个范数都可以由内积导出，如 l_1 和 l_∞ 范数不能由内积导出
- 标准内积与 2-范数之间存在联系：

$$\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}$$

柯西施瓦兹不等式

定理 6

柯西施瓦兹不等式: 若 $\|\cdot\|$ 是由 $(\mathbb{V}, \langle\cdot,\cdot\rangle)$ 导出的范数，那么

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2 \leq \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 .$$

当 $\boldsymbol{y} = \boldsymbol{0}$ 时，不等式成立是显然的。下面证明 $\boldsymbol{y} \neq \boldsymbol{0}$ 的情况。

柯西施瓦兹不等式

证明.

当 $\boldsymbol{y} \neq \boldsymbol{0}$ 时，对任意 $\lambda \in \mathbb{R}$，

$$\begin{aligned} 0 &\leq \langle \boldsymbol{x} - \lambda \boldsymbol{y}, \boldsymbol{x} - \lambda \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x} - \lambda \boldsymbol{y}, \boldsymbol{x} \rangle - \lambda \langle \boldsymbol{x} - \lambda \boldsymbol{y}, \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \lambda^2 \langle \boldsymbol{y}, \boldsymbol{y} \rangle \\ &= \|\boldsymbol{x}\|^2 - 2\lambda \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \lambda^2 \|\boldsymbol{y}\|^2 \end{aligned}$$

取 $\lambda = \langle \boldsymbol{x}, \boldsymbol{y} \rangle \|\boldsymbol{y}\|^{-2}$，得 $0 \leq \|\boldsymbol{x}\|^2 - \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \|\boldsymbol{y}\|^{-2}$

从而得到

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 \leq \|\boldsymbol{x}\|^2 \|\boldsymbol{y}\|^2 .$$

用内积计算向量的长度

例 11

令 $\boldsymbol{x}=(1,1)^{\mathrm{T}}\in\mathbb{R}^2$，如果我们把点积作为内积，则向量 $\boldsymbol{x}$ 的长度为

$$\|\boldsymbol{x}\|=\sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}=\sqrt{1^2+1^2}=\sqrt{2}.$$

我们现在采用一个不同的内积

$$\langle\boldsymbol{x},\boldsymbol{y}\rangle:=\boldsymbol{x}^{\mathrm{T}}\begin{pmatrix}1&-\frac{1}{2}\\-\frac{1}{2}&1\end{pmatrix}\boldsymbol{y}=x_1y_1-\frac{1}{2}(x_1y_2+x_2y_1)+x_2y_2,$$

则向量长度为

$$\langle\boldsymbol{x},\boldsymbol{x}\rangle=x_1^2-x_1x_2+x_2^2=1-1+1=1\implies\|\boldsymbol{x}\|=\sqrt{1}=1.$$

所以相对于点积这个内积使得 $\boldsymbol{x}$ 变短了。事实上，在 x_1,x_2 同号的情况下，上述内积会给出一个比点积更小的向量长度值；如果异号则给出更大的值。

6.2.4 距离和度量

定义 9

考虑一个内积空间 $\langle \mathbb{V}, \langle \cdot, \cdot \rangle \rangle$。我们称

$$d(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}$$

为 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$ 的距离。如果我们用点积作为内积，则上述距离称为**欧几里得距离**，简称**欧氏距离**。

和向量长度类似，向量间的距离不必需要内积，使用范数就足够了。如果我们使用内积导出的范数，则距离会依赖于内积的选择。

度量

定义 10

考虑一个内积空间 $\langle \mathbb{V}, \langle \cdot, \cdot \rangle \rangle$，我们称映射

$$d : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$$
$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto d(\boldsymbol{x}, \boldsymbol{y})$$

为度量。

度量空间

定义 11

一个度量空间由一个有序对 $(\mathbb{V}, d)$ 表示，其中 $\mathbb{V}$ 是一种集合，d 是定义在 V 上的一种度量：

$$d : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$$

且对任意 $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{V}$，需满足

- 非负性：即 $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$，且 $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Longleftrightarrow \boldsymbol{x} = \boldsymbol{y}$；
- 对称性：即 $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$；
- 三角不等式：$d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$。

所以赋范线性空间按由范数导出的距离构成一个特殊的**度量空间**。度量空间也称为**距离空间**。

6.2.5 夹角

有了内积和范数，便可以定义两个向量之间的夹角。考虑直角坐标系中两个非零向量 $\boldsymbol{x}, \boldsymbol{y}$，它们与原点 $\boldsymbol{o}$ 构成一个三角形，如图3所示。

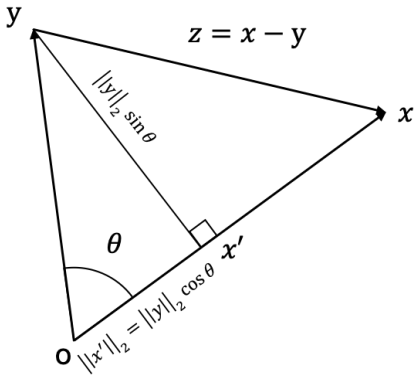
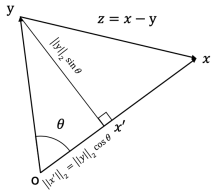


图 3: 向量 $\boldsymbol{x}, \boldsymbol{y}$ 之间的夹角 θ

夹角



令 θ 是 $\boldsymbol{ox}$ 与 $\boldsymbol{oy}$ 之间的夹角，$\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{y}$。对三角形 yxx' 运用勾股定理，有

$$\begin{aligned}\|\boldsymbol{z}\|_2^2 &= (\|\boldsymbol{y}\|_2 \sin\theta)^2 + (\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2 \cos\theta)^2 \\ &= \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos\theta.\end{aligned}$$

由于

$$\|\boldsymbol{z}\|_2^2 = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y},$$

则有

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos\theta.$$

夹角

因此向量 $\boldsymbol{x}, \boldsymbol{y}$ 之间的夹角 θ 可以由下式来确定：

$$\cos\theta = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}.$$

当 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0$ 时，向量 $\boldsymbol{x}, \boldsymbol{y}$ 之间的角度为 $90°$，称为正交。当 θ 为 $0°$ 或者 $180°$ 时，$\boldsymbol{x}, \boldsymbol{y}$ 成一直线，即 $\boldsymbol{y} = k\boldsymbol{x}, k \in \mathbb{K}$，称为平行。

6.2.6 正交

定义 12

设向量 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$，如果 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$，则称 $\boldsymbol{x}, \boldsymbol{y}$ **正交**，记作 $\boldsymbol{x} \perp \boldsymbol{y}$。特别地，如果 $||\boldsymbol{x}|| = 1 = ||\boldsymbol{y}||$，也即是单位向量时，称 $\boldsymbol{x}, \boldsymbol{y}$ **标准正交**。

零向量与任何向量正交。

对于非零向量组 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d\}$，如果对于 $\forall i \neq j$，有 $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = 0$，则称向量组两两正交，并且具有如下性质：

命题 1

两两正交的向量组线性无关。

正交向量

例 12

考虑两个向量 $\boldsymbol{x} = (1,1)^{\mathrm{T}}, \boldsymbol{y} = (-1,1)^{\mathrm{T}} \in \mathbb{R}^2$。我们用两种不同的内积来确定他们之间的夹角 ω。使用点积作为内积则可以得到 ω 为 $90°$，所以 $\boldsymbol{x} \perp \boldsymbol{y}$。

而我们选择内积

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{y}$$

计算 $\boldsymbol{x}, \boldsymbol{y}$ 之间的角度 ω 时，

$$\cos\omega = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = -\frac{1}{3} \implies \omega \approx 109.5°$$

所以 $\boldsymbol{x}, \boldsymbol{y}$ 不是正交的。

因此向量在一种内积下正交并不代表他们在其他内积下也正交。

正交矩阵

定义 13

方阵 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ 是一个正交矩阵当且仅当它的列向量是标准正交的，即

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{I} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A},$$

因此 $\boldsymbol{A}^{-1} = \boldsymbol{A}^{\mathrm{T}}$。

正交矩阵的特性

正交矩阵变换是特殊的，因为用正交矩阵 $\boldsymbol{A}$ 作用一个向量 $\boldsymbol{x}$ 时，向量 $\boldsymbol{x}$ 的长度不变。对于点积，我们得到

$$\|\boldsymbol{A}\boldsymbol{x}\|^2 = (\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = \|\boldsymbol{x}\|^2.$$

并且两个向量 $\boldsymbol{x}, \boldsymbol{y}$ 的夹角也不会在正交矩阵的作用下改变。同样用点积作为内积，则 $\boldsymbol{A}\boldsymbol{x}$ 和 $\boldsymbol{A}\boldsymbol{y}$ 的夹角为

$$\cos\omega = \frac{(\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x})}{\|\boldsymbol{A}\boldsymbol{x}\|\|\boldsymbol{A}\boldsymbol{y}\|} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}}{\sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y}}} = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|},$$

这就是向量 $\boldsymbol{x}, \boldsymbol{y}$ 之间的夹角。这就意味着正交矩阵 $\boldsymbol{A}$ 能够保持角度和长度不变。

本小节主要讨论非概率相关的相似性度量。

假设有 m 个样本，每个样本由 n 个属性的特征向量组成。样本集合可以用矩阵 X 表示

$$\boldsymbol{X} = [x_{ij}]_{m\times n} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

矩阵的第 j 列表示第 j 个样本，第 i 行表示第 i 个属性，矩阵元素 x_{ij} 表示第 j 个样本的第 i 个属性值；$i = 1, 2, \cdots, m$，$j = 1, 2, \cdots, n$。

6.3.1 闵可夫斯基距离

定义 14

给定特征空间或样本集合 X，X 是由范数或内积导出的 m 维度量空间 $\mathbb{R}^m$ 中点的集合，其中 $\boldsymbol{x}_i, \boldsymbol{x}_j \in X$, $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{mi})^T$，$\boldsymbol{x}_j = (x_{1j}, x_{2j}, ..., x_{mj})^T$，样本 $\boldsymbol{x}_i$ 与样本 $\boldsymbol{x}_j$ 的**闵可夫斯基距离**，简称**闵氏距离**，定义为

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[p]{\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_p,$$

其中 $1 \leq p < \infty$。

- 当 $p = 2$ 时，对应欧氏距离
- 当 $p = 1$ 时，对应曼哈顿距离
- 当 $p \to \infty$ 时，对应切比雪夫距离

欧氏距离

定义 15

欧氏距离是指的是 n 维空间中各个点之间的直线距离，计算公式如下：

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt[2]{\sum_{k=1}^{m} |x_{ki} - x_{kj}|^2} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2,$$

注意到在定义欧氏距离的时候利用到了 2 范数，我们在考虑两个点之间的距离就是在考虑这两个点对应的向量的差的长度（即范数），那么同样可以用其他范数来定义不同的距离。

曼哈顿距离

定义 16

曼哈顿距离，也称出租车距离，用以标明两个点在标准坐标系上的绝对轴距总和，计算公式如下：

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{m} |x_{ki} - x_{kj}| = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_1$$

在 1-范数意义下的距离，我们称为曼哈顿距离。这是因为曼哈顿城的道路总是横着或者竖着，我们要计算从一点走到另外一点的距离不能够使用两点之间的直线距离。

曼哈顿距离

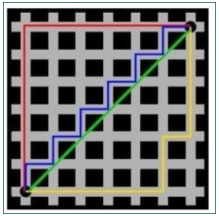


图 4: 曼哈顿距离和欧式距离

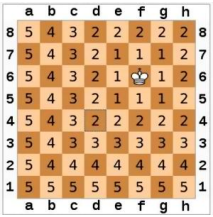
如图4所示，绿线代表欧氏距离，红线代表曼哈顿距离，蓝、黄线代表等价的曼哈顿距离。

切比雪夫距离

定义 17

切比雪夫距离是将二个点其各坐标数值差绝对值的最大值作为距离：

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max_k |x_{ki} - x_{kj}| = ||\boldsymbol{x}_i - \boldsymbol{x}_j||_\infty$$



切比雪夫距离起源于国际象棋中国王的走法，相当于国王从一个格子走到另一个格子的最小步数。

图 5: 国际象棋中的切比雪夫距离

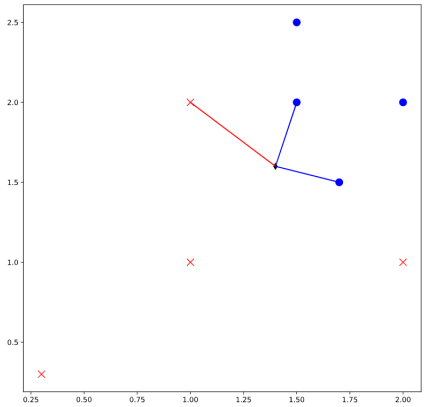
例子：k-NN

下面我们将以闵式距离为例，用 k-NN 算法来展示不同相似性度量对于模型的影响。

例 13

k-近邻算法 (k-NN) 是机器学习中一种非常简单的算法。给定带类别的数据。当预测新的数据属于哪一类别时，我们只需比较距离这一数据最近的 k 个已知数据点中哪种类别是多数，则认为这个数据点就是该类别。

比如取 $k = 3$, 图中的黑色点 (菱形) 即是所要预测的数据点，而蓝色 (圆点) 为正例，红色 (叉) 为负例。因为距离最近的三个点中，有两个是正例，一个为负例。故我们认为这个数据点为正例。

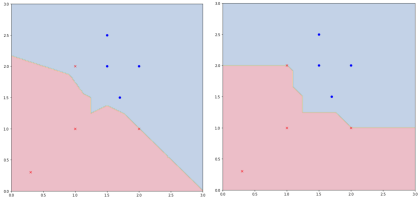


例子:k-NN

为了说明不同度量对模型的影响，给定训练集：
正例为：$(1.5, 2), (1.7, 1.5), (2, 2), (1.5, 2.5)$
负例为：$(1, 2), (0.3, 0.3), (2, 1), (1, 1)$
固定 $k = 3$，然后将平面分成两部分，一部分涂上红色表示某模型将此区域的点预测为负例，另外一部分涂成蓝色表示正例。

哪种度量方式更好呢？这取决于具体的问题以及给出的数据。如果我们确定给出的数据都是准确值没有任何误差。我们就有理由相信右边的模型比左边的模型更好。如果我们不能保证给出的数据都是准确值，那左边的模型也有可能比右边的更好。

下面左图采用的距离度量方式是欧氏距离，右图采用的是曼哈顿距离。



闵可夫斯基距离度量的缺陷

需要注意，闵氏距离 (包括曼哈顿距离、欧氏距离和切比雪夫距离) 存在明显的缺点，我们通过下例进行说明。

例：给定二维样本 (身高, 体重)，其中身高范围是 $150 \sim 190$，体重范围是 $50 \sim 60$，有三个样本：$a(180, 50), b(190, 50), c(180, 60)$。

- 通过计算可以得出 ab 之间的闵氏距离等于 ac 之间的闵氏距离，但是身高的 $10cm$ 不等价于体重的 $10kg$。

在学习了概率部分之后，我们将会给出解决这个问题的方案。

6.3.2 余弦相似度

前面我们使用样本特征向量之间的闵氏距离作为相似性度量，我们也可以考虑从特征向量之间的夹角来界定相似程度。

定义 18

余弦相似度是通过计算两个样本特征向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间夹角的余弦值，以此作为两个样本间相似度大小的衡量，计算公式如下

$$sim_{cos}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|} = \frac{\sum_{k=1}^{m} x_{ki}x_{kj}}{[\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2]^{\frac{1}{2}}}$$

- 因为夹角余弦取值范围为 [-1,1]，所以余弦相似度的取值范围也是 [-1,1]。
- 夹角余弦越大表示两个向量的夹角越小，夹角余弦越小表示两向量的夹角越大。
- 当两个向量的方向重合时夹角余弦取最大值 1，当两个向量的方向完全相反夹角余弦取最小值 -1。

余弦相似度应用举例

例 14

回顾 *2.1.1* 节纽约时报的四则新闻提要的例子，我们知道它们的分别可以用向量表示为：

$$\begin{aligned}
\boldsymbol{a}^{'} &= (\frac{1}{3}, \frac{2}{3}, 0, 0, 0, 0)^{\mathrm{T}}, \\
\boldsymbol{b}^{'} &= (\frac{1}{10}, 0, \frac{3}{10}, \frac{1}{5}, \frac{2}{5}, 0)^{\mathrm{T}}, \\
\boldsymbol{c}^{'} &= (0, 0, 0, \frac{1}{2}, 0, \frac{1}{2})^{\mathrm{T}}, \\
\boldsymbol{d}^{'} &= (0, 0, 0, 0, 0, 1)^{\mathrm{T}}.
\end{aligned}$$

利用夹角的概念，经过计算，可得两两新闻提要之间的余弦相似度如表1所示：

Table 1: 四则新闻标题两两之间的余弦夹角

$\cos\theta$	$\boldsymbol{a}^{'}$	$\boldsymbol{b}^{'}$	$\boldsymbol{c}^{'}$	$\boldsymbol{d}^{'}$
$\boldsymbol{a}^{'}$	1	0.0816	0	0
$\boldsymbol{b}^{'}$	0.0816	1	0.2582	0
$\boldsymbol{c}^{'}$	0	0.2582	1	0.7071
$\boldsymbol{d}^{'}$	0	0	0.7071	1

当两则新闻提要之间没有重复的单词出现，夹角余弦值为 0；当两则新闻提要是相同的，夹角余弦值为 1。

余弦相似度

余弦相似度从夹角上区分差异，而对绝对的数值不敏感，因此没法衡量每个维度上数值的差异，我们通过下例进行说明：

例：用户对内容评分，按 5 分制，X 和 Y 两个用户对两个内容的评分分别为 $(1,2)$ 和 $(4,5)$。

- X 和 Y 之间的余弦相似度 0.98，两者极为相似。但从评分上看 X 似乎不喜欢这两个内容，而 Y 则比较喜欢。
- 余弦相似度对数值的不敏感导致了结果的误差，需要修正这种不合理性就出现了调整余弦相似度，即所有维度上的数值都减去一个均值。
- 假设两个内容评分均值都是 3，那么调整后为 $(-2,-1)$ 和 $(1,2)$，再用余弦相似度计算，得到 -0.8，相似度为负值并且差异不小，但显然更加符合现实。

6.3.3 汉明距离

除了闵氏距离和余弦相似度，还可以使用汉明距离作为相似性度量。

定义 19

汉明距离表示两个（相同长度）字符串对应位置上的值不等的个数。

例如：

- 1011101 与 1001001 之间的汉明距离是 2。
- 2143896 与 2233796 之间的汉明距离是 3。
- "toned" 与"roses" 之间的汉明距离是 3。

这个距离常常用在字符串的处理上，我们可以将其拓展应用到向量上。

6.4.1 矩阵内积：矩阵的向量化

将向量的内积与范数加以推广，即可引出矩阵的内积与范数。

定义 20

令 $m \times n$ 实矩阵 $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n]$，将这个矩阵“拉长”为 $mn \times 1$ 向量

$$\boldsymbol{a} = vec(\boldsymbol{A}) = \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix}$$

$vec(\boldsymbol{A})$ 称为**矩阵 $\boldsymbol{A}$ 的（列）向量化。**

利用向量的内积和范数表达，即可以得到下面有关矩阵内积和范数的定义。

矩阵内积

定义 21

设矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是线性空间 $\mathbb{R}^{m\times n}$ 中任意两个实矩阵，其**矩阵内积**为：

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \langle vec(\boldsymbol{A}), vec(\boldsymbol{B})\rangle = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{b}_i = \sum_{i=1}^{n} \langle \boldsymbol{a}_i, \boldsymbol{b}_i\rangle$$

或等价写作

$$\langle \boldsymbol{A}, \boldsymbol{B}\rangle = vec(\boldsymbol{A})^{\mathrm{T}} vec(\boldsymbol{B}) = Tr(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{B})$$

6.4.2 矩阵范数

定义 22

对于任意的 $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$，与 $c \in \mathbb{R}$。如果函数 $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ 满足条件

(1) $\|\boldsymbol{A}\| \geq 0 (\|\boldsymbol{A}\| = 0 \Longleftrightarrow \boldsymbol{A} = \boldsymbol{O})$ （正定条件）；

(2) $\|c\boldsymbol{A}\| = |c|\|\boldsymbol{A}\|$，$c$ 为实数 （齐次条件）；

(3) $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ （三角不等式）；

则称 $\|\cdot\|$ 是 $\mathbb{R}^{m\times n}$ 上的一个（广义）矩阵范数。

l_1 范数

例 15

对任意 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，由

$$\|\boldsymbol{A}\|_{m_1} := \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$$

定义的 $\|\cdot\|_{m_1}$ 是 $\mathbb{R}^{m\times n}$ 上的矩阵范数，称为 l_1 范数。

容易验证：

(1)$\|\boldsymbol{A}\|_{m_1} \geq 0$, 并且当 $\boldsymbol{A} = \boldsymbol{O}$ 即 $a_{ij} \equiv 0$ 时 $\|\boldsymbol{A}\|_{m_1} = 0$。

(2)$\|c\boldsymbol{A}\|_{m_1} = \sum_{i=1}^{m}\sum_{j=1}^{n}|ca_{ij}| = |c|\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}| = |c|\,\|\boldsymbol{A}\|_{m_1}$。

(3)$\|\boldsymbol{A}+\boldsymbol{B}\|_{m_1} = \sum_{i=1}^{m}\sum_{j=1}^{n}(|a_{ij}+b_{ij}|) \leq \sum_{i=1}^{m}\sum_{j=1}^{n}(|a_{ij}|+|b_{ij}|) = \|\boldsymbol{A}\|_{m_1} + \|\boldsymbol{B}\|_{m_1}$。

因此, 实函数 $\|\boldsymbol{A}\|_{m_1} = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|$ 是一种矩阵范数。

实际上，这个范数就是 $vec(\boldsymbol{A})$ 的 l_1 范数。

l_∞ 范数

例 16

对任意 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，由

$$\|\boldsymbol{A}\|_{m_\infty} := \max_{1\le i\le m, 1\le j\le n} |a_{ij}|$$

定义的 $\|\cdot\|_{m_\infty}$ 是 $\mathbb{R}^{m\times n}$ 上的（广义）矩阵范数，称为 l_∞ 范数。

实际上，这个范数就是 $vec(\boldsymbol{A})$ 的 l_∞ 范数。

Frobenius 范数

例 17

对任意 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，由

$$\|\boldsymbol{A}\|_F := (\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2)^{\frac{1}{2}} = (Tr(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}))^{1/2}$$

定义的 $\|\cdot\|_F$ 是 $\mathbb{R}^{m\times n}$ 上的矩阵范数，称为 l_2 范数或 *Frobenius* 范数（F 范数）。

实际上，这个范数就是 $vec(\boldsymbol{A})$ 的 l_2 范数。

相容性条件

考虑到矩阵乘法的重要地位，因此讨论矩阵范数时一般附加“相容性”条件。

定义 23

若矩阵范数 $\|\cdot\|$ 满足：

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|, \text{对任意} \boldsymbol{A} \in \mathbb{R}^{m\times p}, \boldsymbol{B} \in \mathbb{R}^{p\times n}$$

则称矩阵范数满足**相容性条件**。

不满足相容性条件的矩阵范数我们可以称其为广义矩阵范数。

相容性条件

例 18

$\|\cdot\|_{m_1}$ 满足相容性条件。

$$\|\boldsymbol{A}\boldsymbol{B}\|_{m_1} \leq \|\boldsymbol{A}\|_{m_1} \|\boldsymbol{B}\|_{m_1}, \text{对任意} \boldsymbol{A} \in \mathbb{R}^{m\times p}, \boldsymbol{B} \in \mathbb{R}^{p\times n}$$

例 19

$\|\cdot\|_F$ 满足相容性条件。

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F, \text{对任意} \boldsymbol{A} \in \mathbb{R}^{m\times p}, \boldsymbol{B} \in \mathbb{R}^{p\times n}$$

相容性条件

例 20

$\|\cdot\|_{m_\infty}$ 不满足相容性条件。

取

$$\boldsymbol{A}=\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$$

那么

$$\left\|\boldsymbol{A}^2\right\|_{m_\infty}=\|2\boldsymbol{A}\|_{m_\infty}=2\not\leqslant\|\boldsymbol{A}\|_{m_\infty}^2=1$$

我们只需要对 $\|\cdot\|_{m_\infty}$ 做一点修改，就可以使其满足相容性条件：

$$\|\boldsymbol{A}\|_{m_\infty}:=n\max_{1\leq i\leq m,1\leq j\leq n}|a_{ij}|$$

6.4.3 算子范数：相容性

定义 24

若矩阵范数 $\|\cdot\|_M$ 和向量范数 $\|\cdot\|_v$ 满足

$$\|\boldsymbol{A}\boldsymbol{x}\|_v \leq \|\boldsymbol{A}\|_M \|\boldsymbol{x}\|_v, \boldsymbol{A} \in \mathbb{R}^{m\times n}, \boldsymbol{x} \in \mathbb{R}^n,$$

则称矩阵范数 $\|\cdot\|_M$ 与向量范数 $\|\cdot\|_v$ 是相容的。

对于给定的任意向量范数，我们都可以如下构造一个与该向量范数相容的矩阵范数。

算子范数

定义 25

称 $m \times n$ 矩阵空间上如下定义的范数 $\|\cdot\|$ 为**从属于向量范数 $\|\cdot\|_v$ 的矩阵范数**，也称其为由向量范数 $\|\cdot\|_v$ 诱导出的**算子范数**

$$\begin{aligned}\|\boldsymbol{A}\| &= \max\{\|\boldsymbol{A}\boldsymbol{x}\|_v : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\|_v = 1\} \\ &= \max\left\{\frac{\|\boldsymbol{A}\boldsymbol{x}\|_v}{\|\boldsymbol{x}\|_v} : \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq 0\right\}\end{aligned}$$

显然，该矩阵范数和向量范数 $\|\cdot\|_v$ 是相容的。

因为，对任意 $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq 0$，

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|_v}{\|\boldsymbol{x}\|_v} \leq \max\left\{\frac{\|\boldsymbol{A}\boldsymbol{x}\|_v}{\|\boldsymbol{x}\|_v} : \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq 0\right\} = \|\boldsymbol{A}\|$$

所以 $\|\boldsymbol{A}\boldsymbol{x}\|_v \leq \|\boldsymbol{A}\|_M \|\boldsymbol{x}\|_v$。

算子范数

定理 7

算子范数都满足相容性条件。

证明.

设矩阵范数 $\|\cdot\|$ 是由向量范数 $\|\cdot\|_v$ 诱导出的算子范数，$\boldsymbol{A} \in \mathbb{R}^{m\times p}$，$\boldsymbol{B} \in \mathbb{R}^{p\times n}$，$\boldsymbol{x} \in \mathbb{R}^n$，

$$\|\boldsymbol{AB}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{ABx}\|_v \leq \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\|\|\boldsymbol{Bx}\|_v = \|\boldsymbol{A}\| \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{Bx}\|_v = \|\boldsymbol{A}\|\|\boldsymbol{B}\|$$

□

常用的算子范数

经常利用向量的 l_p-范数诱导出算子范数：

$$\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}$$

定理 8

设 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$，$p = 1, \infty, 2$ 时，向量的 l_p-范数诱导出的算子范数分别为

$$\|\boldsymbol{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

$$\|\boldsymbol{A}\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$$

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$$

当 $\boldsymbol{A} = \boldsymbol{O}$ 时，以上三式显然成立。假定 $\boldsymbol{A} \neq \boldsymbol{O}$，对以上的三个范数进行证明。

1 范数证明

证明.

对于 1 范数，将给定的 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 按列分块 $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n]$，并记 $\delta = \left\|\boldsymbol{a}_{j_0}\right\|_1 = \max_{1\le j\le n} \left\|\boldsymbol{a}_j\right\|_1$，则对任意满足 $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i| = 1$ 的 $\boldsymbol{x} \in \mathbb{R}^n$，有

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{x}\|_1 = \left\|\sum_{j=1}^n x_j \boldsymbol{a}_j\right\| &\le \sum_{j=1}^n |x_j| \left\|\boldsymbol{a}_j\right\|_1 \\ &\le \sum_{j=1}^n |x_j| \max_{1\le j\le n} \left\|\boldsymbol{a}_j\right\|_1 = \left\|\boldsymbol{a}_{j_0}\right\|_1 = \delta\end{aligned}$$

□

此处我们证明了 $\|\boldsymbol{A}\|_1 := \max_{\|\boldsymbol{x}\|_1=1} \|\boldsymbol{A}\boldsymbol{x}\|_1 \le \delta$。

1 范数证明

此外，令 $\boldsymbol{x}$ 为第 j_0 个元素为 1，其余分量为 0 的向量 $\boldsymbol{e}_{j_0}$，则有 $\left\|\boldsymbol{e}_{j_0}\right\|_1 = 1$，而且

$$\left\|\boldsymbol{A}\boldsymbol{e}_{j_0}\right\|_1 = \left\|\boldsymbol{a}_{j_0}\right\|_1 = \delta$$

这样我们证明了存在满足 $\|\boldsymbol{x}\|_1 = 1$ 的 $\boldsymbol{x}$，使得 $\|\boldsymbol{A}\boldsymbol{x}\|_1 = \delta$。

因此有

$$\|\boldsymbol{A}\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \|\boldsymbol{A}\boldsymbol{x}\|_1 = \delta = \max_{1\le j\le n} \|\boldsymbol{a}_j\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|.$$

∞ 范数证明

证明.

对于 ∞ 范数，记

$$\eta = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|,$$

则对任意满足 $\|\boldsymbol{x}\|_\infty = 1$ 的 $\boldsymbol{x} \in \mathbb{R}^n$，有

$$\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_{1\le i\le m} |\sum_{j=1}^{n} a_{ij}x_j| \le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}||x_j| \le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| = \eta$$

□

此处我们证明了 $\|\boldsymbol{A}\|_\infty := \max_{\|\boldsymbol{x}\|_\infty=1} \|\boldsymbol{A}\boldsymbol{x}\|_\infty \le \eta$。

∞ 范数证明

设 $\boldsymbol{A}$ 的第 k 行的元素的绝对值之和最大，即 $\eta = \sum_{j=1}^{n} |a_{kj}|$。令

$$\tilde{\boldsymbol{x}} = (sgn(a_{k1}), \cdots, sgn(a_{kn}))^{\mathrm{T}}$$

则 $\boldsymbol{A} \neq \boldsymbol{O}$ 蕴含 $\|\tilde{\boldsymbol{x}}\|_\infty = 1$，有 $\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|_\infty = \sum_{j=1}^{n} |a_{kj}| = \eta$。

这里证明了存在满足 $\|\boldsymbol{x}\|_\infty = 1$ 的 $\boldsymbol{x}$，使得 $\|\boldsymbol{A}\boldsymbol{x}\|_\infty = \eta$。

则

$$\|\boldsymbol{A}\|_\infty = \eta = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|.$$

2 范数证明

证明.

对于 2 范数，应有

$$\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \left[(\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right]^{\frac{1}{2}}$$

$$= \max_{\|\boldsymbol{x}\|_2=1} \left[\boldsymbol{x}^{\mathrm{T}}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)\boldsymbol{x}\right]^{\frac{1}{2}}$$

注意，$\boldsymbol{A}^T\boldsymbol{A}$ 是半正定矩阵，设其特征值为

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0,$$

以及其对应的正交规范特征向量为 $\boldsymbol{q}_1, \cdots, \boldsymbol{q}_n \in \mathbb{R}^n$。 □

2 范数证明

则对任一满足 $\|\boldsymbol{x}\|_2 = 1$ 的向量 $\boldsymbol{x} \in \mathbb{R}^n$ 有

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{q}_i$$

$$\sum_{i=1}^{n} \alpha_i^2 = 1$$

于是，有

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} = \sum_{i=1}^{n} \lambda_i \alpha_i^2 \leqslant \lambda_1$$

这里我们证明了 $\|\boldsymbol{A}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \left[\boldsymbol{x}^{\mathrm{T}} \left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{A}\right) \boldsymbol{x}\right]^{\frac{1}{2}} \leq \sqrt{\lambda_1}$。

2 范数证明

另一方面，若取 $\boldsymbol{x}=\boldsymbol{q}_1$，则有

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{q}_1^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{q}_1=\boldsymbol{q}_1^{\mathrm{T}}\lambda_1\boldsymbol{q}_1=\lambda_1$$

这里我们证明了存在满足 $\|\boldsymbol{x}\|_2=1$ 的 $\boldsymbol{x}$，使得 $\|\boldsymbol{A}\boldsymbol{x}\|_2=\sqrt{\lambda_1}$。

所以

$$\|\boldsymbol{A}\|_2=\max_{\|\boldsymbol{x}\|_2=1}\|\boldsymbol{A}\boldsymbol{x}\|_2=\sqrt{\lambda_1}=\sqrt{\lambda_{\max}\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)}$$

2 范数的性质

我们通常分别称矩阵的 1 范数、∞ 范数和 2 范数为列和范数、行和范数和谱范数。显然矩阵列和范数与行和范数容易计算，而矩阵的谱范数不易计算，它需要计算 $\boldsymbol{A}^T\boldsymbol{A}$ 的最大特征值，但是谱范数具有几个好的性质，使它在理论研究中很有用处。下面给出谱范数几个常用的性质。

定理 9

设 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$，则

(1) $||\boldsymbol{A}||_2 = \max\{|\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x}| : x, y \in \mathbb{C}^n, ||\boldsymbol{x}||_2 = ||\boldsymbol{y}||_2 = 1\}$;

(2) $||\boldsymbol{A}^T||_2 = ||\boldsymbol{A}||_2 = \sqrt{||\boldsymbol{A}^T\boldsymbol{A}||_2}$;

(3) 对于任意的正交矩阵 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 有，$||\boldsymbol{U}\boldsymbol{A}||_2 = ||\boldsymbol{A}\,\boldsymbol{V}||_2 = ||\boldsymbol{A}||_2$。

例 21

设矩阵 $\boldsymbol{A}=\begin{pmatrix}2 & -1\\ -2 & 4\end{pmatrix}$，求 $\|\boldsymbol{A}\|_p,(p=1,2,\infty)$ 以及 $\|\boldsymbol{A}\|_F$

$$\|\boldsymbol{A}\|_1=\max\{2+|-2|,|-1|+4\}=5$$

$$\|\boldsymbol{A}\|_\infty=\max\{2+|-1|,|-2|+4\}=6$$

因为 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=\begin{pmatrix}2 & -2\\ -1 & 4\end{pmatrix}\begin{pmatrix}2 & -1\\ -2 & 4\end{pmatrix}=\begin{pmatrix}8 & -10\\ -10 & 17\end{pmatrix}$ 由

$|\boldsymbol{I}\lambda-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}|=\begin{vmatrix}\lambda-8 & 10\\ 10 & \lambda-17\end{vmatrix}=0$ 解得 $\lambda_1=23.466,\lambda_2=1.534$ 故

$\|\boldsymbol{A}\|_2=\sqrt{23.466}=4.844$

$$\|\boldsymbol{A}\|_F=(2^2+(-1)^2+(-2)^2+4^2)^{\frac{1}{2}}=5$$

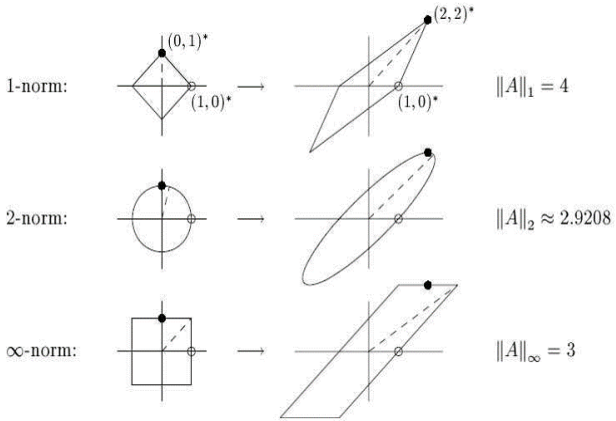
算子范数的几何意义

例 22

对应于 $p = 1, 2, \infty$ 三种向量范数的单位球面 $\mathbb{S} = \{\boldsymbol{x} \in \mathbb{R}^2 | \|\boldsymbol{x}\|_p = 1\}$ 在矩阵

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$$

作用下的效果分别为



其他矩阵范数

$$||\boldsymbol{A}||_{1,2} = (\sum_{j=1}^{n} ||\boldsymbol{a}_j||_1^2)^{\frac{1}{2}} = (\sum_{j=1}^{n}(\sum_{i=1}^{m}|A_{ij}|)^2)^{\frac{1}{2}}$$

$$||\boldsymbol{A}||_{2,1} = \sum_{j=1}^{n} ||\boldsymbol{a}_j||_2 = \sum_{j=1}^{n}(\sum_{i=1}^{m}|A_{ij}|^2)^{\frac{1}{2}}$$

$$||\boldsymbol{A}||_{p,q} = (\sum_{j=1}^{n}(\sum_{i=1}^{m}|A_{ij}|^p)^{\frac{q}{p}})^{\frac{1}{q}}$$

$$||\boldsymbol{A}||_* = \mathsf{Tr}(\sqrt{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}})$$

6.5.1 回顾

- 在第一章中，对于监督学习问题，常常将其等价为求下列函数的最小值问题：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

其中 y_i 是特征 x_i 的标签，而 $f(x_i)$ 则是模型 f 对于特征 x_i 给出的一个预测值；

$L(y_i, f(x_i))$ 是损失函数，用于衡量单个样本预测值和真实值的误差；

$\frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$ 是误差项（也称为代价函数），误差项主要用来衡量输出的预测值和真实值之间的整体误差；

$\lambda J(f)$ 是正则化项，正则化项主要用于防止模型过拟合。

- 那么在监督学习中损失函数 L 和正则化函数 J 具有什么形式呢？

损失函数 L 的形式

- 基于距离度量的损失
- 非距离度量形式的损失

常见的距离相关损失函数

- 0-1 损失函数

$$L(Y, f(X)) = \begin{cases} 1 & Y \neq f(X) \\ 0 & Y = f(X) \end{cases},$$

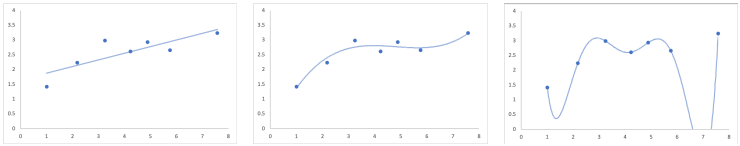
- 绝对损失函数

$$L(Y, f(X)) = |Y - f(X)|,$$

- 平方损失函数

$$L(Y, f(X)) = (Y - f(X))^2,$$

过拟合



(a) $y = ax + b$ (b) $y = ax^4 + bx^3 + cx^2 + dx + e$ (c) $y = ax^6 + bx^5 + cx^4 + dx^3 + e$

图 6: 欠拟合、正常拟合、过拟合

- 欠拟合的模型因为模型假设过于简单，而无法反应数据的真实情况。
- 若增加模型的复杂性则可得一个合适的拟合，从而能够很好地反应数据的分布和趋势。
- 若继续增加模型的复杂性就会产生过拟合的现象。这种模型不仅仅拟合了数据，并且还拟合了噪音。这将使得模型在新数据上表现很差。

正则化

欠拟合问题易解决，但是过拟合，则需要通过其他一些手段——正则化，来解决。在实际中，正则化则是在要求模型的训练误差很小的情况下，使得模型本身的复杂度也变小，即

$$\begin{cases} \min Loss(\theta, \boldsymbol{X}, \boldsymbol{Y}) \\ \min J(\theta) \end{cases}$$

其中 $Loss(\theta, \boldsymbol{X}, \boldsymbol{Y})$ 是经验风险，$J(\theta)$ 是正则化项，θ 是模型的参数，可代表模型。通常我们将上述双目标函数优化任务转化为下列优化任务

$$\min Loss(\theta, \boldsymbol{X}, \boldsymbol{Y}) + \lambda J(\theta)$$

这个式子就是我们之前提到的结构风险，λ 用于调节经验风险和正则化项的关系。那么这个正则化项该如何选择呢？

正则化：范数的选择

- 在前面多项式拟合的例子中，想要避免过拟合，则需让模型不出现用更高次函数去拟合四次函数产生的带有噪声的数据的情况，高次函数拟合效果虽好但也拟合了数据噪声
- 我们应让模型尽可能表现为用待定系数的四次函数去拟合四次函数产生的带有噪声的数据。如果要让模型不出现过拟合情况，需要让模型参数向量 θ 中元素个数尽可能少，对于拟合四次函数的问题来说，模型求解得到的参数向量中元素个数应该是 4，而不是其它。
- 一个向量的元素个数正好是向量的 l_0 范数，因此让模型参数向量 θ 中元素的个数最小化其实等于优化 $\min \|\theta\|_0$，这样就建立了过拟合解决方案和范数的联系：

$$\min Loss(\theta, \boldsymbol{X}, \boldsymbol{Y}) + \lambda \|\theta\|_0$$

6.5.2 损失函数和正则化项中的范数：常用的向量范数

- l_0 范数（并不满足范数的定义）

$$\|\boldsymbol{x}\|_0 = \sum_{i=1}^{n} \mathcal{I}(x_i \neq 0)$$

- l_1 范数

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

- l_2 范数

$$\|\boldsymbol{x}\|_2 = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$$

l_0 范数的含义

l_0 范数指向量中非 0 的元素的个数，优化 l_0 范数可以得到一些稀疏解。

好处

- 稀疏解存储成本低
- 自动实现特征选择 (Feature Selection)
- 可解释性强 (Interpretability)

坏处

- l_0 范数很难优化求解，是个 NP-hard 问题。

l_1 范数的含义

l_1 范数是 l_0 范数的最优凸近似，而且它比 l_0 范数要容易优化求解，所以 l_1 范数被称为“稀疏规则算子”（Lasso ）。它常常可以应用在：

- 稀疏编码
- 特征选择
- 压缩感知

$$\min \|\boldsymbol{x}\|_0 \quad s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \quad \xleftarrow{\text{在一定条件下，以概率 1 意义下等价}} \quad \min \|\boldsymbol{x}\|_0 \quad s.t. \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$

l_1 范数和 l_0 范数都可以实现稀疏，l_1 因具有比 l_0 更好的优化求解特性而被广泛应用。

l_2 范数的含义

l_2 范数又称“岭回归”（Ridge Regression）或“权值衰减（weight decay）”，最小化 l_2 范数，可以使得参数向量的元素值都很小，大都接近于 0。它的好处是可以改善过拟合、易于优化。

关于“过拟合”：在数学上称为“病态”（ill-condition)：即函数的输入改变一点点，输出却改变非常大。l_2 范数限制了参数都很小，实际上就限制了多项式各分量的影响很小，一定程度上避免了模型出现“病态”的情况。

与 l_2 范数相比，l_1 范数更有可能得到值为 0 的解，所以导致稀疏。l_2 范数得到的解在各个分量上更为均衡。

l_1 范数存在不可导点，这导致了在这一点上无法进行有效地优化。而 l_2 范数是光滑的。但正因为如此，l_2 范数鲁棒性更差一些。因为对于异常值 l_2 范数倾向于把他的影响变得更大。

6.5.3 损失函数和正则化项中的范数：常用的矩阵范数

$$\|A\|_F = (\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2)^{\frac{1}{2}} = \sqrt{\mathsf{Tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$$

$$||\boldsymbol{A}||_p = ||vec(A)||_p = (\sum_{j=1}^{n}\sum_{i=1}^{m} |A_{ij}|^p)^{\frac{1}{p}}$$

$$||\boldsymbol{A}||_{2,1} = \sum_{j=1}^{n} ||\boldsymbol{a}_j||_2 = \sum_{j=1}^{n}(\sum_{i=1}^{m} |A_{ij}|^2)^{\frac{1}{2}}$$

$$||\boldsymbol{A}||_{1,2} = (\sum_{j=1}^{n} ||\boldsymbol{a}_j||_1^2)^{\frac{1}{2}} = (\sum_{j=1}^{n}(\sum_{i=1}^{m} |A_{ij}|)^2)^{\frac{1}{2}}$$

$$||\boldsymbol{A}||_* = \mathsf{Tr}(\sqrt{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}})$$

F 范数的含义

$$\|A\|_F = (\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2)^{\frac{1}{2}} = \sqrt{\mathsf{Tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$$

- 最常用的矩阵范数为 F 范数。最小化矩阵的 F 范数，会使得矩阵的每个元素都很小，接近于 0
- $\|A-B\|_F$ 可度量矩阵 A, B 之间的差异，最小化可使得两者尽可能的相等

p 范数的含义

$$||\boldsymbol{A}||_p = ||vec(A)||_p = (\sum_{j=1}^{n}\sum_{i=1}^{m}|A_{ij}|^p)^{\frac{1}{p}}$$

- 当 $p=1$ 时，为矩阵 l_1 范数，最小化 $||\boldsymbol{A}||_1$ 能让矩阵 A 元素稀疏。稀疏矩阵的优点：计算速度更快、存储成本低、可解释性强（例如：文本分类中，可知哪些词对类别起重要作用）
- 当 $p=2$ 时，为矩阵 l_2 范数，也即 F 范数

$\|\boldsymbol{A}\|_{2,1}$ 和 $\|A\|_{1,2}$ 范数的含义

$$\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{n} ||\boldsymbol{a}_j||_2 = \sum_{j=1}^{n} (\sum_{i=1}^{m} |A_{ij}|^2)^{\frac{1}{2}}$$

最小化 $||A||_{2,1}$ 范数能让矩阵 $\boldsymbol{A}$ 不同行之间（列向量）稀疏，在机器学习领域这属于 Group Lasso。应用于文本分类领域：Lasso 对应于找出关键词，Group Lasso 找出关键句子，Hierarchical Lasso 找出关键段。

$$||\boldsymbol{A}||_{1,2} = (\sum_{j=1}^{n} ||\boldsymbol{a}_j||_1^2)^{\frac{1}{2}} = (\sum_{j=1}^{n} (\sum_{i=1}^{m} |A_{ij}|)^2)^{\frac{1}{2}}$$

最小化 $\|A\|_{1,2}$ 范数能让矩阵行内元素互斥，也即行内存在 0 元素但不能全为 0。这可以应用在特征选择的时候不同的类别可以选择互斥的特征。

核范数

$$||\boldsymbol{A}||_* = \mathrm{Tr}(\sqrt{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}})$$

- 核范数指矩阵奇异值的和，英文为 Nuclear norm
- 最小化核范数 $\|\cdot\|_*$ 可以导致低秩矩阵 (Low-Rank)。低秩矩阵的应用：矩阵填充 (Matrix Completion)，例如 –推荐系统、鲁棒 PCA、背景建模、变换不变低秩纹理（TILT）

内积和范数小结

内积和范数

- 向量范数和内积
- 矩阵范数和内积
- ...

内积和范数的应用

- 相似性度量
- 过拟合
- ...

从优化或者数值计算的角度来说，范数有助于处理“病态”的问题或 condition number 不好的问题，与方程组解的敏感性相关。

内积和范数小结

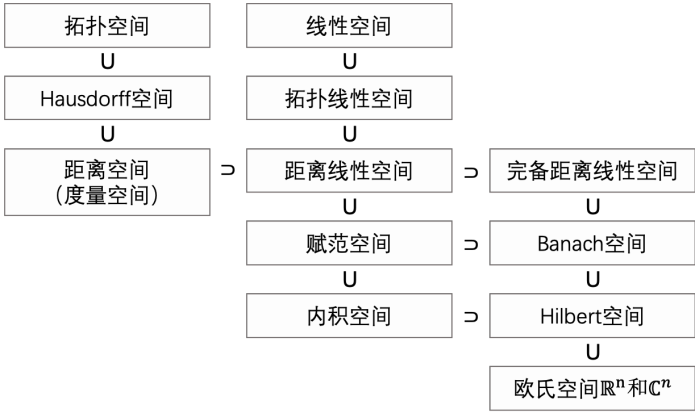


图 7: 各类空间的层次关系