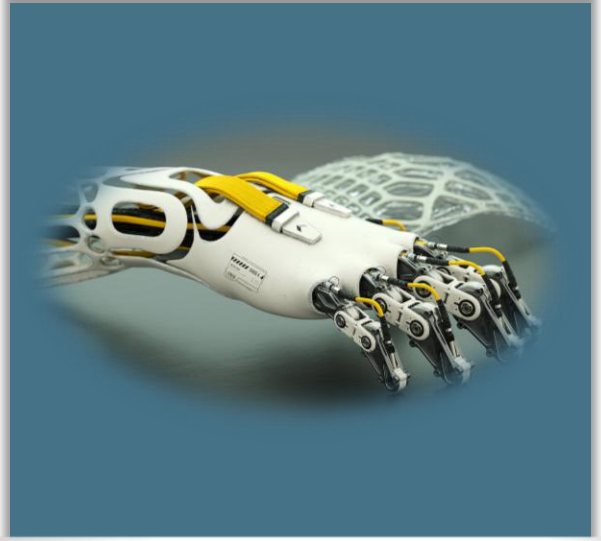
机器学习

授课人：王闻博

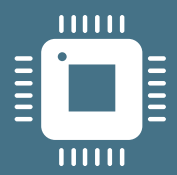
Email: wenbo_wang@kust.edu.cn

昆明理工大学 机电工程学院

2026年03月13日



机器学习的数学基础




1. 机器学习中的常用符号
2. 矩阵运算回顾
3. 概率论回顾
4. 优化理论初步

机器学习中的常用符号

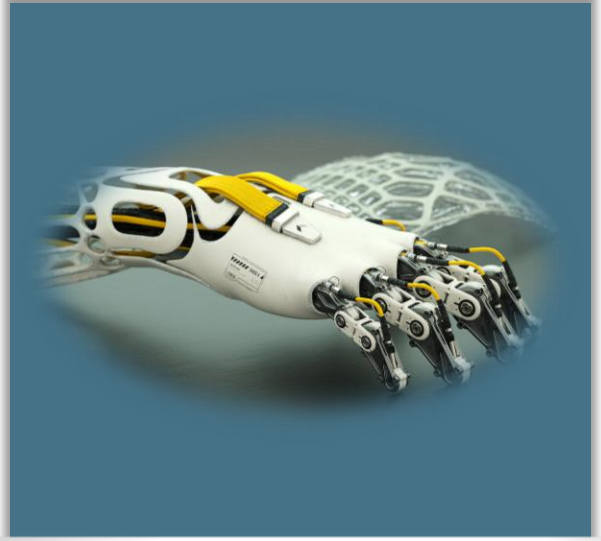
- a, b, c：标量（整数或实数）；
- **x, y, z**：向量（粗体，小写）；
- **A, B, C**：矩阵（粗体，大写）；
- **A, B, C**：张量（粗体，大写）；
- X, Y, Z：随机变量（普通字体，大写）；
- $a \in \mathcal{A}$：集合成员关系：a 是集合 $\mathcal{A}$ 的成员，集合使用花体表示(Latex中符号为\mathcal{A})；
- $|\mathcal{A}|$：基数（cardinality）——集合 A 中元素的数量；
- ||v||：向量 v 的范数；
- **u · v 或 〈u, v〉**：向量 u 和 v 的内积（点积）；

机器学习中的常用符号（续）

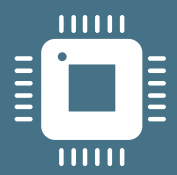
- $\mathbb{R}$：实数集（Latex中符号为\mathbb{R}）;
- $\mathbb{R}^n$：维度为 n 的实数空间;
- $\mathbf{y = f(x)}$ **或** $\mathbf{x \rightarrow f(x)}$：函数（映射）：为每个输入值 x 分配一个唯一的值 f(x);
- $\mathbf{f: \mathbb{R}^n \rightarrow \mathbb{R}}$：函数（映射）：将一个 n 维向量映射为一个标量;
- $\mathbf{A \odot B}$：矩阵 A 和 B 的元素间对位乘积（Hadamard积）；
- $\mathbf{A}^{\dagger}$：矩阵 A 的伪逆（Latex中的符号为A^{\dagger}）；
- $d^n f / dx^n$：函数 f 关于 x 的第 n 阶导数;
- $\nabla_x f(x)$：函数 f 关于 x 的梯度;
- $\mathbf{H}_f$：函数 f 的Hessian矩阵;

机器学习中的常用符号（续）

- $X \sim P$：随机变量 X 服从分布 P；
- $\mathrm{P}(X|Y)$：给定 Y 时 X 的概率；
- $\mathcal{N}(\mu, \sigma^2)$：均值为 μ，方差为 σ^2 的高斯分布；
- $\mathrm{E}_{X\sim P}[f(X)]$：关于分布 $\mathrm{P}(X)$ 的 函数$f(X)$ 的期望；
- $\mathrm{Var}(f(X))$：$f(X)$ 的方差；
- $\mathrm{Cov}(f(X), g(Y))$：$f(X)$ 和 $g(Y)$ 的协方差；
- $\mathrm{corr}(\mathrm{X}, \mathrm{Y})$：X 和 Y 的相关系数$\frac{Cov\ (X,Y)}{\sigma_X \sigma_Y}$；
- $D_{KL}(P||Q)$：分布 P 和 Q 的Kullback-Leibler散度；
- $\mathrm{CE}(\mathrm{P}, \mathrm{Q})$：分布 P 和 Q 的交叉熵。



机器学习
导论




1. 机器学习中的常用符号
2. 矩阵运算回顾
3. 概率论回顾
4. 优化理论初步

向量

- 向量定义
  - 计算机科学中：向量是一个一维的有序实数值标量组成的数组；
  - 数学中：向量是一个具有大小和方向的量，用箭头表示方向，箭头的长度与大小成比例；
- 向量的表示形式
  - 向量可以写成列向量或行向量，用粗体小写字母表示：

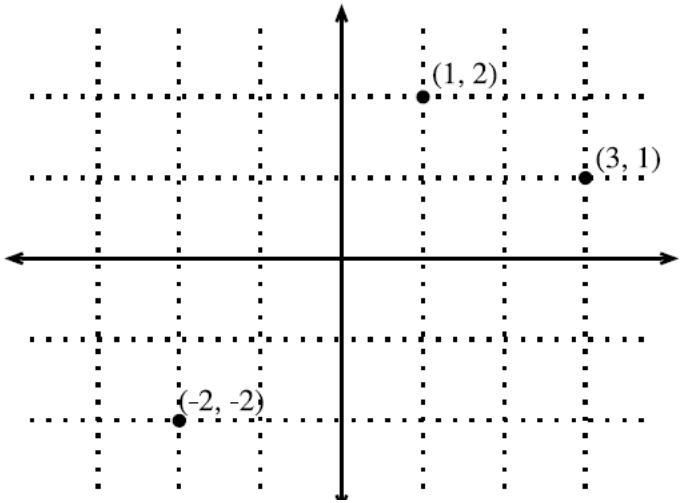
$$\mathbf{x} = \begin{bmatrix} 1 \\ 7 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{x} = [1 \quad 7 \quad 0 \quad 1]^T$$

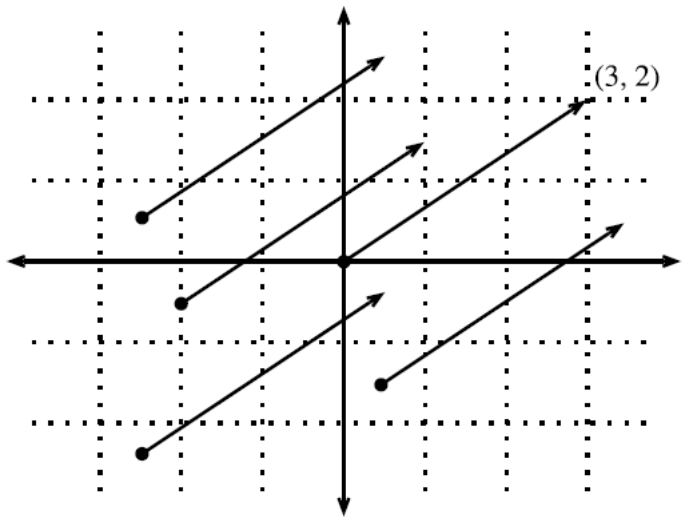
  - 一般形式的向量表示：对于一个具有 n 个元素的一般形式向量，该向量位于 n 维空间中 $x \in \mathbb{R}^n$

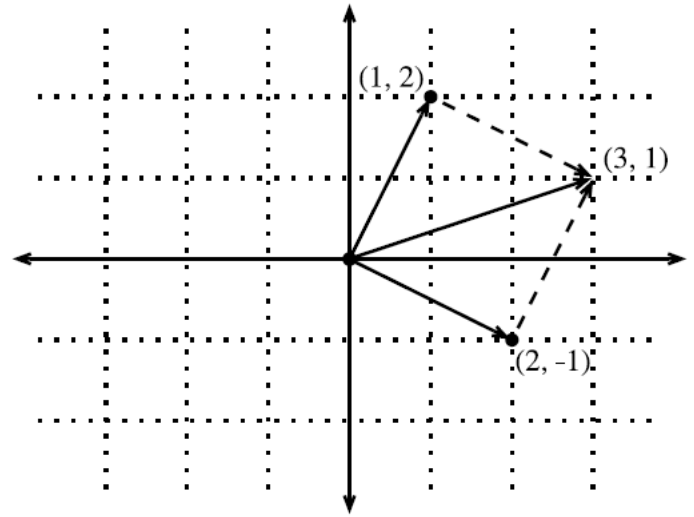
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

向量的几何含义

- **向量的第一种解释：空间中的点**
  - 例如，在二维空间中，我们可以将数据点相对于坐标原点进行可视化。
- **向量的第二种解释：空间中的方向**
  - 例如，向量 $\vec{v}=[3,2]^T$表示在二维空间向右移动3个单位，向上移动2个单位。
  - 有时使用符号 $\vec{v}$ 来表示向量具有方向。
- **向量加法**
  - 将坐标相加，并按照两个被加向量所指示的方向进行移动。



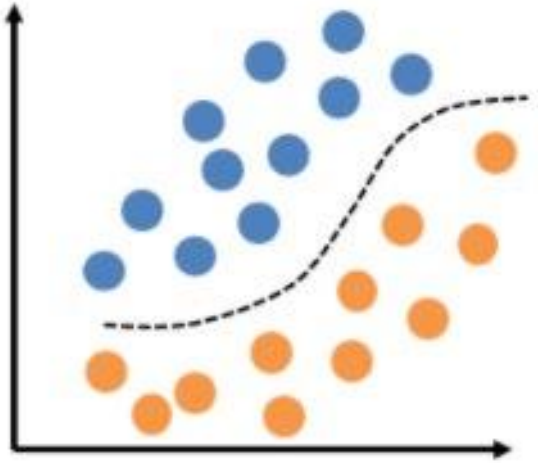


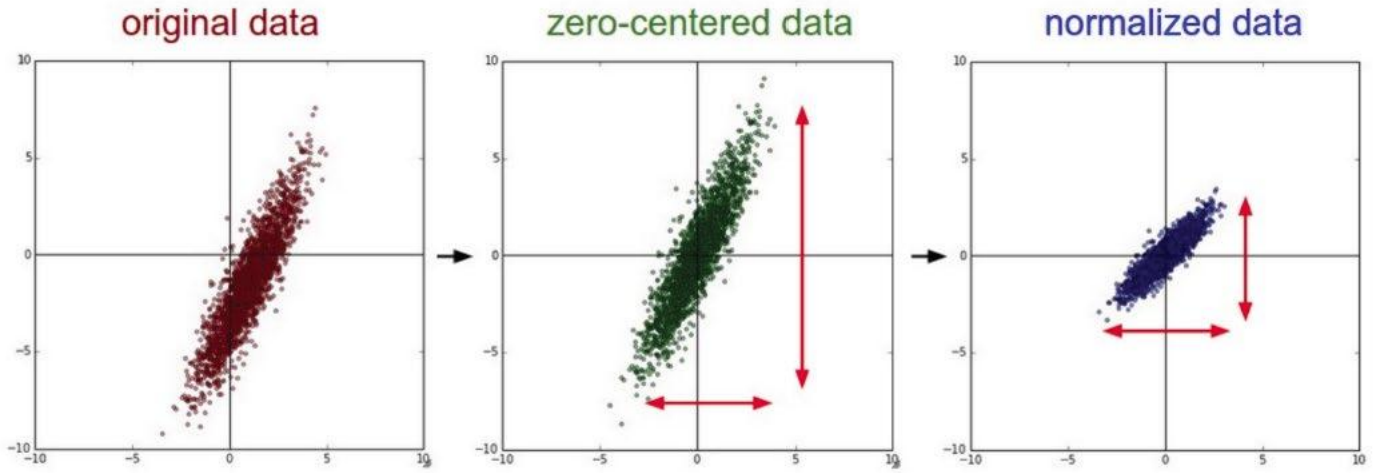


图片来自：http://d2l.ai/chapter_appendix-mathematics-for-deep-learning/geometry-linear-algebra-1.html#geometry-of-vectors

向量的几何含义

- **向量作为空间中的点使我们能够在机器学习中将输入示例的训练集视为空间中的一组点**
  - 因此，数据分类可以被视为发现如何分离属于不同类别的两个点簇（左图）
    - 例如，区分包含汽车、飞机、建筑物的图像
  - 或者，通过基变换，可以帮助可视化训练数据的零中心化和归一化（右图）





向量的点积

- **向量的点积，$u \cdot v = u^T v = \sum_i u_i v_i$**
  - 也被称为内积或标量积。
  - $u \cdot v$通常表示为$< u, v >$,点积是一个对称操作（满足交换律）。
- **点积的几何解释：与两个向量之间的角度相关，即向量 $u \cdot w$ 与向量的范数的比值是 cos(θ)**

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| cos(\theta) \qquad cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

  - 如果两个向量是正交的，θ=90°，即 $\cos(\theta)$=0，则 $u \cdot v$=0；
  - 在机器学习中，术语 $cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$ 有时被用作衡量两个向量或数据实例的接近程度，称为余弦相似度。

向量的范数

- 向量的范数是一个将向量映射到非负标量值的函数
  - 范数是向量大小的度量
- 范数 f 应满足以下性质：
  - 缩放：$f(\alpha\mathbf{x})=|\alpha|f(\mathbf{x})$；
  - 三角不等式：$f(\mathbf{x}+\mathbf{y})\leq f(\mathbf{x})+f(\mathbf{y})$；
  - 非负：$f(\mathbf{x})\geq 0$。
- 向量 x 的一般 ℓ_p **范数**计算如下：

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

常用向量范数

1. 当 p=2 时，有ℓ_2范数，

   - 也称为欧几里得范数（欧式范数），是最常用的范数；
   - ℓ_2 范数通常简写为 $\|\mathbf{x}\|$，省略下标 2。

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

2. 当 p=1 时，有ℓ_1 范数，

   - 使用元素的绝对值的和计算
   - 区分零元素和非零元素

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

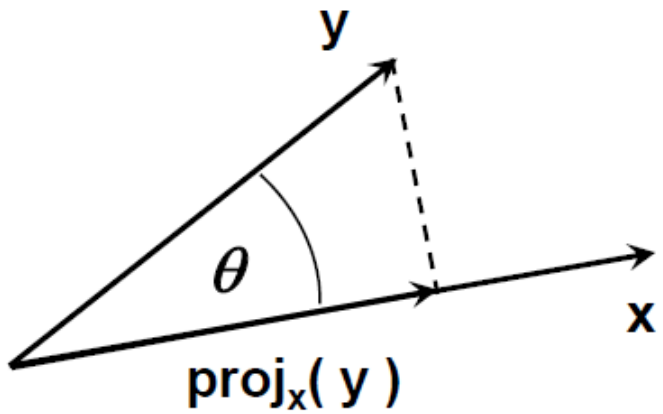
3. 当 p=∞ 时，有ℓ_∞范数，

   - 也称为无穷范数或最大值范数
   - 输出最大元素的绝对值

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

向量的投影

- **向量 y 在向量 x 上的正交投影**
  - 投影可以在任何维度 ≥2 的空间中进行
  - x 方向上的单位向量是$\frac{\mathbf{x}}{\|\mathbf{x}\|}$，单位向量的范数等于 1
  - 向量 y 在 x 上的投影长度是 $\|\mathbf{y}\|\cos(\theta)$，即标量投影
  - 正交投影是向量 $\mathbf{proj}_{\mathbf{x}}(\mathbf{y})$：



$$\mathrm{proj}_{\mathbf{x}}\,\mathbf{y} = \underbrace{\|\mathbf{y}\|\cos\theta}_{\text{标量投影}} \cdot \underbrace{\frac{\mathbf{x}}{\|\mathbf{x}\|}}_{\text{X方向上的单位向量}} = \frac{\|\mathbf{y}\|\cos\theta}{\|\mathbf{x}\|}\mathbf{x}.$$

注意：$\cos\theta = \dfrac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$,

代入得：$\mathrm{proj}_{\mathbf{x}}(\mathbf{y}) = \|\mathbf{y}\| \cdot \dfrac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \cdot \dfrac{\mathbf{x}}{\|\mathbf{x}\|} = \dfrac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2}\mathbf{x}$

矩阵

- **矩阵的定义：一个由实数值标量组成的矩形数组，排列在 m 行和 n 列中**
  - 每个元素 a_{ij} 标识第 i 行、第 j 列的元素；
  - 元素表示为 a_{ij} ，或 A_{ij} ，或 $[\mathrm{A}]_{ij}$ ，或 $\mathrm{A}(i,j)$ ；

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

  - 对于矩阵 $\mathrm{A} \in \mathbb{R}^{m \times n}$，其大小（维度）表示为 $m \times n$ 或 (m,n)。

矩阵的运算

- **矩阵的加减法** $\left(\mathbf{A}\pm\mathbf{B}\right)_{i,j}=\mathbf{A}_{i,j}\pm\mathbf{B}_{i,j}$
  - 示例：
$$\begin{bmatrix}1&3&1\\1&0&0\end{bmatrix}+\begin{bmatrix}0&0&5\\7&5&0\end{bmatrix}=\begin{bmatrix}1+0&3+0&1+5\\1+7&0+5&0+0\end{bmatrix}=\begin{bmatrix}1&3&6\\8&5&0\end{bmatrix}$$
- **标量乘法** $\left(c\mathbf{A}\right)_{i,j}=c\cdot\mathbf{A}_{i,j}$
  - 示例：
$$2\cdot\begin{bmatrix}1&8&-3\\4&-2&5\end{bmatrix}=\begin{bmatrix}2\cdot1&2\cdot8&2\cdot-3\\2\cdot4&2\cdot-2&2\cdot5\end{bmatrix}=\begin{bmatrix}2&16&-6\\8&-4&10\end{bmatrix}$$
- **矩阵乘法** $\left(\mathbf{AB}\right)_{i,j}=\mathbf{A}_{i,1}\mathbf{B}_{1,j}+\mathbf{A}_{i,2}\mathbf{B}_{2,j}+\cdots+\mathbf{A}_{i,n}\mathbf{B}_{n,j}$
  - 仅当左矩阵的列数等于右矩阵的行数时定义
  - 注意 AB≠BA。
  - 示例：
$$\begin{bmatrix}2&3&4\\1&0&0\end{bmatrix}\begin{bmatrix}0&1000\\1&100\\0&10\end{bmatrix}=\begin{bmatrix}3&2340\\0&1000\end{bmatrix}$$

矩阵运算

- **矩阵的转置（Transpose，行列互换）**

$$\left(\mathbf{A}^T\right)_{i,j} = \mathbf{A}_{j,i} \qquad \begin{bmatrix} 1 & 2 & 3 \\ 0 & -6 & 7 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 1 & 0 \\ 2 & -6 \\ 3 & 7 \end{bmatrix}$$

- **矩阵运算的一些性质：**

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$
$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \boldsymbol{B}^T \qquad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$$
$$\left(\mathbf{A}^T\right)^T = \mathbf{A} \qquad (\mathbf{AB})^T = \boldsymbol{B}^T\mathbf{A}^T$$

- **方阵和单位矩阵**
  - 行数和列数相同的矩阵。
  - 主对角线为1，其余为0的方阵：

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

矩阵的运算（续）

- **矩阵的行列式：行列式记作det(A)，或|A|，是一个实数值标量**
  - 对于一般矩阵，行列式的计算公式为：

$$\det(\mathbf{A}) = \sum_j a_{ij}(-1)^{i+j} det\left(\mathbf{A}_{(i,j)}\right)$$

  - 在上述公式中，$A_{(i,j)}$ 是通过移除与索引 i 和 j 相关的行和列得到的矩阵的子式（Minor）。
  - 回顾：对 2×2矩阵，行列式的计算公式是？矩阵的代数余子式是？
- **矩阵的迹是所有主对角线元素的和**

$$\mathrm{Tr}(\mathbf{A}) = \sum_i a_{ii}$$

- **一个矩阵如果满足 $A=A^T$ ，则称为对称矩阵。**

矩阵-向量积

- **考虑矩阵 $\mathbf{A} \in \mathbb{R}^{m \times n}$ 和一个向量 $\mathbf{x} \in \mathbb{R}^{n}$，**
  - 矩阵可以用它的行向量表示：

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix}$$

- **矩阵-向量乘积是一个长度为 m 的列向量，其第 i 个元素是点积 $\mathbf{a}_i^T \mathbf{x}$:**

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_m^\top \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{x} \\ \mathbf{a}_2^\top \mathbf{x} \\ \vdots \\ \mathbf{a}_m^\top \mathbf{x} \end{bmatrix}$$

- **矩阵维度：$\mathbf{A}_{(m \times n)} \cdot \mathbf{x}_{(n \times 1)} = \mathbf{A}\mathbf{x}_{(m \times 1)}$**

矩阵-矩阵积

- 对矩阵$\mathbf{A}\in\mathbb{R}^{n\times k}$和$\mathbf{B}\in\mathbb{R}^{k\times m}$：

$$\mathbf{C}=\mathbf{AB}=\begin{bmatrix}\mathbf{a}_1^\top\\ \mathbf{a}_2^\top\\ \vdots\\ \mathbf{a}_n^\top\end{bmatrix}\begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m\end{bmatrix}=\begin{bmatrix}\mathbf{a}_1^\top\mathbf{b}_1 & \mathbf{a}_1^\top\mathbf{b}_2 & \cdots & \mathbf{a}_1^\top\mathbf{b}_m\\ \mathbf{a}_2^\top\mathbf{b}_1 & \mathbf{a}_2^\top\mathbf{b}_2 & \cdots & \mathbf{a}_2^\top\mathbf{b}_m\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{a}_n^\top\mathbf{b}_1 & \mathbf{a}_n^\top\mathbf{b}_2 & \cdots & \mathbf{a}_n^\top\mathbf{b}_m\end{bmatrix}$$

- 矩阵维度：$\mathbf{A}_{(n\times k)}\cdot\mathbf{B}_{(k\times m)}=\mathbf{Ax}_{(n\times m)}$

矩阵的Hadamard积

- **两个矩阵 A 和 B 的逐元素相乘称为哈达玛积（Hadamard product）或对位元素积（Elementwise Product）。**
  - 使用符号⊙表示。
  - 操作：

    每个元素是 **A** 和 **B** 对应位置元素的乘积。

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \cdots & a_{mn}b_{mn} \end{bmatrix}$$

性质	数学表达式
交换律	$\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$
结合律	$(\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C} = \mathbf{A} \circ (\mathbf{B} \circ \mathbf{C})$
分配律 (对矩阵加法)	$\mathbf{A} \circ (\mathbf{B} + \mathbf{C}) = \mathbf{A} \circ \mathbf{B} + \mathbf{A} \circ \mathbf{C}$
与普通乘法混合	若 **D** 为对角矩阵，则 $(\mathbf{DA}) \circ \mathbf{B} = \mathbf{D}(\mathbf{A} \circ \mathbf{B})$ 且 $\mathbf{A} \circ (\mathbf{DB}) = (\mathbf{A} \circ \mathbf{B})\mathbf{D}$) 当 **D** 右乘时)
正定性保持	若 **A** 和 **B** 是（半）正定矩阵，则 **A** ∘ **B** 也是（半）正定矩阵 (**Schur积定理**)。
迹的关系	$\operatorname{tr}\left(\mathbf{A}^T(\mathbf{B} \circ \mathbf{C})\right) = \operatorname{tr}\left((\mathbf{A} \circ \mathbf{B})^T\mathbf{C}\right)$

矩阵的Kronecker积

- **矩阵A的维度为 $m\times n$，矩阵B的维度为 $p\times q$，则它们的Kronecker Product（克罗内克积）将A、 B组合成一个更大的$mp\times nq$维的分块矩阵**
  - 使用符号⊗ 表示：

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

  - 操作：

    用 **A** 的每个元素 a_{ij} 作为标量去乘整个矩阵 **B**，并按 **A** 的结构排列这些块。

性质	数学表达式
双线性/分配律	$(\mathbf{A}+\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes\mathbf{C}+\mathbf{B}\otimes\mathbf{C}$ $\mathbf{A}\otimes(\mathbf{B}+\mathbf{C})=\mathbf{A}\otimes\mathbf{B}+\mathbf{A}\otimes\mathbf{C}$
结合律	$(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C})$
混合积	$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=(\mathbf{AC})\otimes(\mathbf{BD})$
转置	$(\mathbf{A}\otimes\mathbf{B})^T=\mathbf{A}^T\otimes\mathbf{B}^T$
逆 (当可逆时)	$(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$
迹 (当为方阵时)	$\mathrm{tr}(\mathbf{A}\otimes\mathbf{B})=\mathrm{tr}(\mathbf{A})\cdot\mathrm{tr}(\mathbf{B})$
秩	$\mathrm{rank}(\mathbf{A}\otimes\mathbf{B})=\mathrm{rank}(\mathbf{A})\cdot\mathrm{rank}(\mathbf{B})$

矩阵的逆

- **对n×n的的方阵 A，如果其秩为 n，则 A^{-1}是它的逆矩阵，当且仅当它们的乘积是单位矩阵 I：** $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$
  - 逆矩阵的性质： $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$ $\left(\mathbf{AB}\right)^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
  - 如果一个矩阵的逆等于其转置，则该矩阵称为正交矩阵 $\mathbf{A}^{-1} = \mathbf{A}^{T}$
- **如果 det(A)=0（即 rank(A)<n），则逆矩阵不存在**
  - 不可逆的矩阵称为奇异矩阵。

注意：在数值计算中，求一个大矩阵的逆需要消耗昂贵的计算量，因此需要尽量避免；此外，它可能导致数值不稳定。

矩阵的伪逆

- 也称为Moore-Penrose伪逆；
- 对于非方阵，逆矩阵不存在，因此使用伪逆
  - 如果 m>n，则伪逆定义为 $\mathbf{A}^{\dagger} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{T}$ and $\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{I}$
  - 如果 m<n，则伪逆定义为 $\mathbf{A}^{\dagger} = \mathbf{A}^{T}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1}$ and $\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{I}$
- 举例：对于一个维度为$\mathbf{A}_{(2\times3)}$的矩阵，可以找到一个大小为 $\mathbf{A}^{\dagger}_{(3\times2)}$的伪逆，使得
  - $\mathbf{A}_{(2\times3)}\mathbf{A}^{\dagger}_{(3\times2)}=\mathbf{I}_{(2\times2)}$

“张量”

- **机器学习或深度学习语境下的张量是 n 维的标量数组**
  - 向量是一阶张量，$\mathbf{v} \in \mathbb{R}^n$；
  - 矩阵是二阶张量，$\mathbf{A} \in \mathbb{R}^{m \times n}$；
  - 以此类推，一个四阶张量是 $\mathbf{T} \in \mathbb{R}^{m \times n \times k \times l}$；
  - 张量用加粗字体的的大写字母表示（例如，**X,Y,Z**）。
- **在机器学习中，最常见的RGB 图像是三阶张量，即它们是三维数组**
  - 3 个轴分别对应宽度、高度和通道；
  - 例如，256×256×3；
  - 通道轴对应颜色通道（红、绿、蓝）。

实数域上的向量空间

- **向量空间 $\mathcal{V}=(\mathcal{V},+,\cdot)$ 是一个定义有两个运算的集合 $\mathcal{V}$：**
  - 向量加法 $+:\mathcal{V}\times\mathcal{V}\to\mathcal{V}$；
  - 标量乘法 $\cdot:\mathbb{R}\times\mathcal{V}\to\mathcal{V}$；
- **其中满足**
  - **1.** $(\mathcal{V},+,\cdot)$**是一个阿贝尔群，**
  - **2. 分配律**
    - 对于所有 $\lambda\in\mathbb{R}$，$x,y\in\mathcal{V}$，有 $\lambda\cdot(x+y)=\lambda\cdot x+\lambda y$
    - 对于所有 $\lambda,\psi\in\mathbb{R}$，$x\in\mathcal{V}$，有 $(\lambda+\psi)\cdot x=\lambda\cdot x+\psi\cdot x$
  - **3.结合律**：对于所有 $\lambda,\psi\in\mathbb{R}$，$x\in\mathcal{V}$，有 $\lambda\cdot(\psi\cdot x)=(\lambda\psi)\cdot x$
  - **4.单位元**：对于所有 $x\in\mathcal{V}$，有 $1\cdot x=x$

- **群：一个集合和其上的某运算⊙，$\mathcal{G}=(\mathcal{G},\odot)$ 当满足：**
  - **封闭性**：对于所有 $x,y\in\mathcal{G}$，$x\odot y\in\mathcal{G}$；
  - **结合律**：对于所有 $x,y,z\in\mathcal{G}$，$(x\odot y)\odot z=x\odot(y\odot z)$。
  - **单位元**：存在 $e\in\mathcal{G}$，对于所有 $x\in\mathcal{G}$，$x\odot e=x$ 且 $e\odot x=x$。
  - **单位逆元**：对于所有 $x\in\mathcal{G}$，存在 $y\in\mathcal{G}$，使得 $x\odot y=e$ 且 $y\odot x=e$。通常表示为 $x^{-1}=y$。

时称为群；

- **进一步地，当上述运算满足下述属性时，称为阿贝尔群：**
  - **交换性**：对于所有 $x,y\in\mathcal{G}$，$x\odot y=y\odot x$

线性相关与线性无关

- **线性组合：**
  - 对于向量空间 $\mathcal{V}$ 和向量$x_1 ,\ldots, x_n \in \mathcal{V}$，$\mathcal{V}$ 中的每个向量 v 可以表示为 $v=\lambda_1 x_1+\ldots+\lambda_k x_k$，其中 $\lambda_1 \ldots \lambda_k \in \mathbb{R}$，这种表示称为向量$x_1 ,\ldots, x_k$的线性组合。
- **线性相关/无关**
  - 如果存在非平凡的线性组合，使得 $0 = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i$，其中至少有一个$\lambda_i \neq 0$，则称向量$x_1 ,\ldots, x_k$是线性相关的。
  - 反之，如果只有平凡解，即对任意λ_i，$\lambda_i \neq 0$，则称向量$x_1 ,\ldots, x_k$是线性无关的。
- **线性表出和线性相关**
  - 如果向量集合$\{x_1 ,\ldots, x_k\}$中，至少有一个向量可以表示为其他向量的线性组合，则这些向量是线性相关的。

子空间中的基和坐标系

- **若$\mathcal{H}$是向量空间$\mathcal{V}$中的一个子空间，$\mathcal{V}$中的向量指标集$\mathcal{B}$称为$\mathcal{H}$的一个基，如果：**
  - $\mathcal{B}=\{\mathbf{b}_1,\ldots,\mathbf{b}_k\}$是一线性无关集；
  - 由$\mathcal{B}$生成的子空间**Span{$\mathcal{B}$}**（又称张成子空间）与$\mathcal{H}$相同，即所有可以表示成$\{\mathbf{b}_1,\ldots,\mathbf{b}_k\}$的线性组合的向量都在$\mathcal{V}$中，且与$\mathcal{H}$相同。
- **矩阵的列空间**
  - 矩阵A的列空间是A的各列向量的线性组合的集合，$m\times n$的矩阵的列空间是$\mathbb{R}^n$的子空间。
  - 矩阵A的零空间是$\mathbf{Ax}=\mathbf{0}$的所有解的集合。
- **线性空间中的坐标系**
  - 如果$\mathcal{B}=\{\mathbf{b}_1,\ldots,\mathbf{b}_k\}$是向量空间$\mathcal{V}$的一个基，$\mathbf{x}$在$\mathcal{V}$中，$\mathbf{x}$相对基$\mathcal{B}$的坐标（称$\mathbf{x}$的$\mathcal{B}$-坐标）是使$\mathbf{x}=c_1\mathbf{b}_1+\ldots+c_k\mathbf{b}_k$成立的权$[c_1\ldots c_k]^T$。
  - 注意：对$\mathcal{V}$中每一个向量$\mathbf{x}$，存在唯一一组标量数$\{c_1\ldots c_k\}$使得$\mathbf{x}=c_1\mathbf{b}_1+\ldots+c_k\mathbf{b}_k$成立。

坐标变换和基变换

- **标准基到基$\mathcal{B}=\{\mathbf{b}_1,\ldots,\mathbf{b}_k\}$的坐标变换：**
  - $\mathbb{R}^n$的标准基是单位矩阵$\mathbf{I}_n$的列的集合；
  - $\mathbf{x}=[x_1,\ldots x_n]^T$是$\mathbf{x}$相对于标准基$\{\mathbf{e}_1,\ldots\mathbf{e}_n\}$的坐标，则坐标变换 $x \to [x]_B$ 可由以下非齐次方程（向量方程）求得：

$$\mathbf{x} = P_B[\mathrm{x}]_\mathrm{B}, \text{其中，令} \mathbf{P}_\mathrm{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]^T$$

向量不变，参照系变了

- **基变换**
  - 在应用中，一个问题常常开始是由某个基$\mathcal{B}$描述，可能会通过将$\mathcal{B}$变为一个新的基$\mathcal{C}$得到进展。
- **基变换定理（基本公式）**
  - 设$\mathcal{B}=\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}$和$\mathcal{C}=\{\mathbf{c}_1,\ldots,\mathbf{c}_n\}$是向量空间$\mathcal{V}$的两个基，则存在一个$n\times n$矩阵$\mathbf{P}_{C\leftarrow B}$，使得$[\mathrm{x}]_C = P_\mathrm{B}[\mathrm{x}]_\mathrm{B}$。其中$P_\mathrm{B}$的列是基$\mathcal{B}$中的向量的$\mathcal{C}$-坐标向量，即

$$P_\mathrm{B} = [[b_1]_C, [b_2]_\mathrm{C}, \ldots, [b_n]_\mathrm{C}]$$

P_B称为坐标变换矩阵（Coordinate Transformation Matrix）描述如何将向量在旧基下的坐标转换为新基下的坐标。$P_{C\leftarrow B} = P_B^{-1}$则称为过渡矩阵（Change of Basis Matrix），描述新基向量在旧基下的线性组合关系。

坐标变换和基变换（续）

- **过渡矩阵P：基变换中连接两个基的矩阵**
  - 设向量空间有两组基$\mathcal{B}=\{\mathbf{b}_1,\ldots,\mathbf{b}_n\}$和$\mathcal{C}=\{\mathbf{c}_1,\ldots,\mathbf{c}_n\}$分别组成矩阵B和C（注意，它们也是标准基到两个基的坐标变换矩阵），过渡矩阵P满足

$$\mathbf{P}_{C\leftarrow B}=\mathbf{P}_B^{-1}\mathbf{P}_C$$

其中P_B和P_C分别是两个基排列成的基矩阵

特性	过渡矩阵	广义坐标变换矩阵
应用范围	仅限基变换（线性、无平移）	广义坐标系变换（可含平移、仿射变换）
矩阵形式	方阵（$n\times n$）	若坐标系C′ 相对于C的变换包含线性变换部分A和平移部分t，则齐次坐标下的变换矩阵为：$\mathrm{T}=\begin{bmatrix}\mathrm{A} & \mathbf{t}\\ 0 & 1\end{bmatrix}$
可逆性	必为可逆矩阵（基变换的双射性）	可逆仅当变换为双射（如仿射变换可逆）

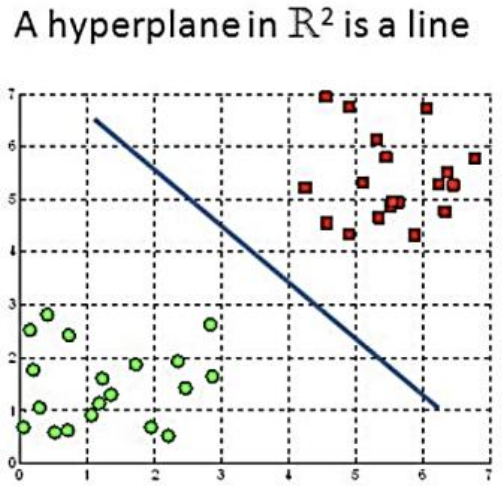
向量空间中的超平面

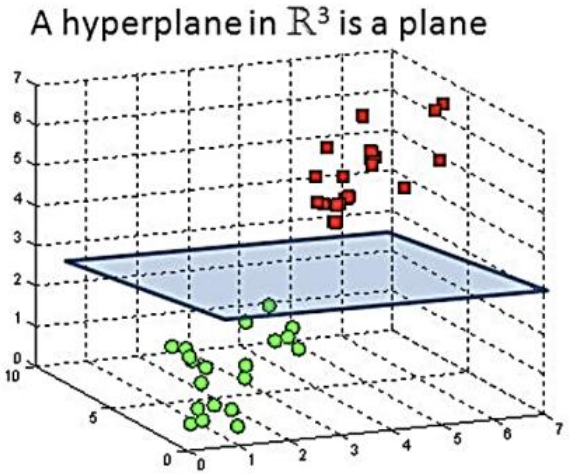
- **超平面是其维数比其所在空间少一维的子空间**
  - 在二维空间中，超平面是一条直线（即一维）；
  - 在三维空间中，超平面是一个平面（即二维）；
  - 在d维向量空间中，超平面有d-1维，并把空间分成两个半空间。
- **超平面是高维空间中平面概念的推广**
- **在机器学习中，超平面在线性可分类情况下被用于决策边界**
  - 如右图：落在超平面两侧的数据点被归属于不同的类别。

定义：超平面是由一个非齐次线性方程定义的点的集合，

$$H = \{\mathbf{x}|\mathbf{w} \cdot \mathbf{x} + b = 0\}$$

- 其中，$\mathbf{w} \in \mathbb{R}^n$是非零向量，称为法向量，决定超平面的方向；
- $b \in \mathbb{R}$是常数项，控制超平面与原点的偏移距离。





图片来源：https://kgpdag.wordpress.com/2015/08/12/svm-simplified/

超平面应用举例

- **例如，对于给定的数据点 $w=[2,1]^T$ ，我们可以使用点积来找到满足 $w \cdot v=1$ 的超平面**
  - 即，所有满足 $w \cdot v>1$ 的向量可以被归为一类，所有满足 $w \cdot v<1$ 的向量可以被归为另一类；
  - 解$w \cdot v=1$，得到：

$$\|\mathbf{v}\|\|\mathbf{w}\|\cos(\theta)=1 \iff \|\mathbf{v}\|\cos(\theta)=\frac{1}{\|\mathbf{w}\|}=\frac{1}{\sqrt{5}}$$

  - 即，以上解是满足 $w \cdot v=1$ 的点集，意味着这些点位于与向量 w 正交的直线上。即直线 $2x+y=1$；
  - v 在 w 上的正交投影是 $\|\mathbf{v}\|\cos(\theta)=\frac{1}{\sqrt{5}}$

小问题：在三维空间里，$w=[1,2,3]^T$,那么满足$w \cdot v=1$的超平面是什么样子的?

特征分解

- **特征分解是将矩阵分解为一组特征值和特征向量**
  - 方阵 A 的一个特征值是标量 λ，其对应的特征向量是非零向量 v满足 $\mathbf{Av} = \lambda\mathbf{v}$
  - 通过解以下**特征方程**找到特征值：$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$
  - 如果矩阵 A 有 n 个线性无关的特征向量$\{\mathbf{v}^1, \ldots, \mathbf{v}^n\}$与对应的特征值$\{\lambda_1, \ldots, \lambda_n\}$，则A的特征分解由下式给出：

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{-1}$$

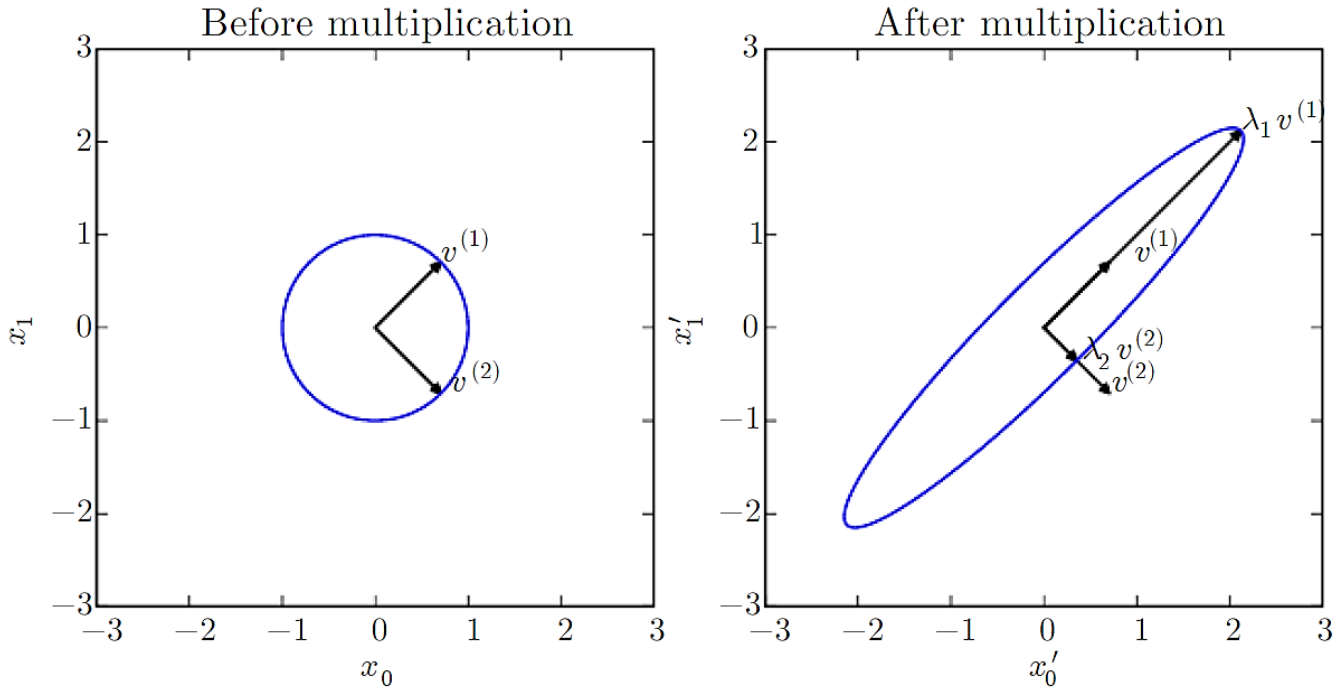
    - 矩阵 V 的列是特征向量，即 $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$
    - Λ 是特征值的对角矩阵，即 $\boldsymbol{\Lambda} = [\lambda_1, \ldots, \lambda_n]$
  - 矩阵 A 的逆可以使用下式表出：$\mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{-1}$
    - 这一方法避免了按行列式求逆的高额运算。
  - 当矩阵 A特征向量数不足n时，可以通过构造**广义特征向量基**，构造约当变换（近似于对角形的标准结构）。

特征分解

- **将矩阵分解为特征值和特征向量可以分析矩阵的某些属性**
  - 如果所有特征值都是正的，矩阵是正定的；
  - 如果所有特征值都是正的或零，矩阵是半正定的；
  - 如果所有特征值都是负的或零，矩阵是半负定的
    - 特性：半正定矩阵保证了对所有 x，都有 $\forall \mathbf{x}, \mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$
  - 矩阵 A 的行列式可以计算 $\det(\mathbf{A}) = \lambda_1 \cdot \lambda_2 \cdots \lambda_n$
  - 如果任何特征值为零，矩阵是奇异的（它没有逆矩阵）。
  - 注意：并非每个矩阵都可以对角化分解为特征值和特征向量
    - 在某些情况下，分解可能涉及复数；
    - 每个实对称矩阵都保证有一个特征分解。根据 $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$ ，其中 V 是一个正交矩阵，所有特征值为实数。

特征分解的几何性质

- **特征值和特征向量的几何解释是它们允许矩阵A对应的线性变换在特定方向上拉伸或压缩特征空间**
  - 下左图：展示了矩阵的两个单位特征向量v_1和 v_2（即，它们的长度为1）；
  - 下右图：向量v_1和v_2与特征值 λ_1 和 λ_2相乘。可以看到特征空间是如何在较大特征值 λ_1 的方向上被缩放的；
  - 这在主成分分析（PCA）中用于降维，其中对应于最大特征值的特征向量用于提取最重要的数据维度，而较小的特征值对应的特征向量则可以被忽略。



图示来源：《深度学习》（“花书”）

奇异值分解：Singular Value Decomposition

- **任意长方阵在内的所有矩阵都具有奇异值分解 (SVD)**
  - 基本形式：$A = U\Sigma V$
    - 其中，$A \in \mathbb{R}^{m\times n}$，$U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$，且U和V都是正交矩阵；
    - $\Sigma \in \mathbb{R}^{m\times n}$为对角矩阵，对角线元素为奇异值，$\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$
  - U和 V的来源
    - U：由 AA^T 的特征向量组成（标准正交基：$UU^T=I$）。
    - V：由 A^TA 的特征向量组成（标准正交基：$VV^T=I$）。
  - SVD可以表示为秩为1的矩阵的线性组合（如下例）：

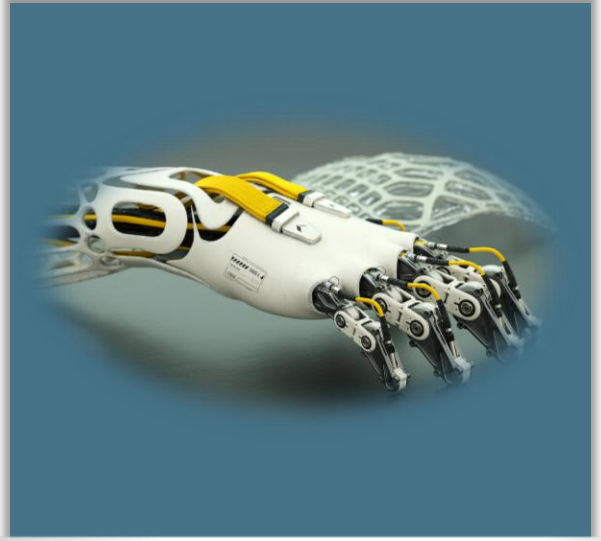
$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

几何意义：

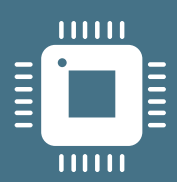
- 线性变换分解：SVD将任意线性变换分解为旋转（V^T）、缩放（Σ）、再旋转（U）。
- 主成分提取（PCA）：奇异值 σ_i 表示数据在对应主方向上的能量强度。

$$A = U\Sigma V^{\mathrm{T}} = \begin{bmatrix} | & | & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \boldsymbol{u}_3 \\ | & | & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \end{bmatrix} \begin{bmatrix} -\boldsymbol{v}_1^{\mathrm{T}}- \\ -\boldsymbol{v}_2^{\mathrm{T}}- \end{bmatrix} = \sigma_1 \begin{bmatrix} | \\ \boldsymbol{u}_1 \\ | \end{bmatrix} \begin{bmatrix} -\boldsymbol{v}_1^{\mathrm{T}}- \end{bmatrix} + \sigma_2 \begin{bmatrix} | \\ \boldsymbol{u}_2 \\ | \end{bmatrix} \begin{bmatrix} -\boldsymbol{v}_2^{\mathrm{T}}- \end{bmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$$



机器学习
导论




1. 机器学习中的常用符号
2. 矩阵运算回顾
3. 概率论回顾
4. 优化理论初步

随机变量和分布

- **随机变量 X 是一个依概率可以取不同值的变量**
  - 例子：X= 掷骰子
    - X 的可能值构成样本空间或结果空间，S={1,2,3,4,5,6}；
    - 我们将“看到5点”的事件表示为 {X=5} 或 X=5；
    - 该事件的概率是 P({X=5}) 或 P(X=5)；
    - 另外，P(5) 也可以用来表示 X 取值为5的概率。
- **概率分布描述了随机变量取每个可能状态的可能性**
  - 常用的简写表示法是，P(X) 是随机变量 X 上的概率分布；
  - 另外，记号 X~P(X) 可以用来表示随机变量 X 具有概率分布 P(X)。
- **随机变量可以是离散的或连续的**
  - 离散随机变量有有限个状态：例如，骰子的面；
  - 连续随机变量有无限个状态：例如，人的身高。

随机变量的定义和性质

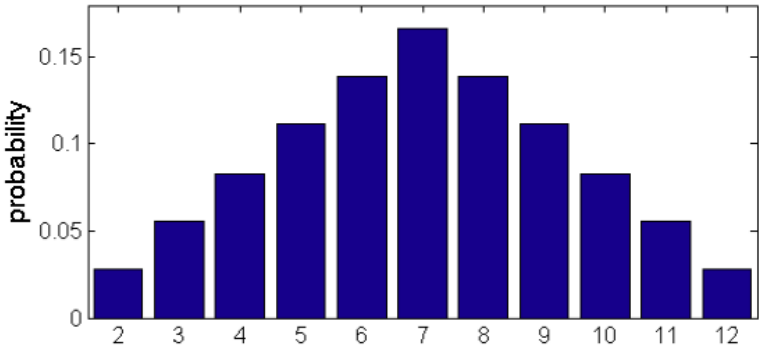
- **数学定义：**
  - 随机变量 X 是一个函数，它将样本空间 Ω 映射到实数集 $\mathbb{R}$。
  - 符号表示：X表示随机变量，x表示数值。
- **随机变量的多样性**
  - 不同的随机变量 $X, Y, \ldots$ 可以在相同的样本空间上定义。
- **事件关联：**
  - 对于固定值 x，我们可以关联一个事件，即随机变量 X 取值 x，即 $\{\omega \in \Omega | X(\omega) = x\}$
- **概率表示：**
  - 一般地，$P_X(\mathcal{S}) = P(X \in \mathcal{S}) = P(X^{-1}(\mathcal{S})) = P(\{\omega \in \Omega : X(\omega) \in \mathcal{S}\})$。

良定的概率分布需满足的公理

- **给定样本空间S中事件 $\mathscr{A}$的概率，记作 P($\mathscr{A}$)，必须满足以下性质：**
  - **非负性**：对于任何事件 $\mathscr{A}\in S$，P($\mathscr{A}$)≥0；
  - **整个样本空间的概率是 1**，P(S)=1；
  - **互斥事件的可加性**：对于所有互斥事件 $\mathscr{A}_1$,$\mathscr{A}_2$ $\in S$,两个事件同时发生的概率等于它们各自概率的和，P($\mathscr{A}_1\cup\mathscr{A}_2$)=P($\mathscr{A}_1$)+P($\mathscr{A}_2$);

离散和连续概率分布

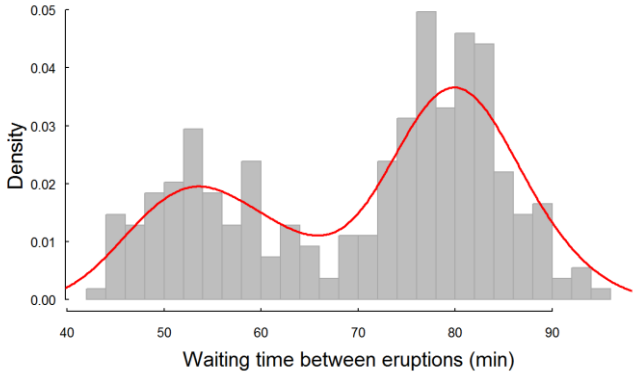
- 离散变量的概率分布可以使用概率质量函数（PMF，Probability Mass Function）来描述 $\mathbb{P}(X = x) := \mathbb{P}\left(\{\omega \in \Omega \mid X(w) = x\}\right)$
  - 例如，两个骰子的和（右图）；



- 连续变量的概率分布可以使用概率密度函数（PDF，Probability Density FUnction）来描述
  - 例如，某随机喷泉喷水的间隔时间（右图）；
  - PDF 给出了具有体积 δX 的无穷小区域的概率
  - 区间 [a,b] 上的概率，可以如通过将PDF进行如下积分：

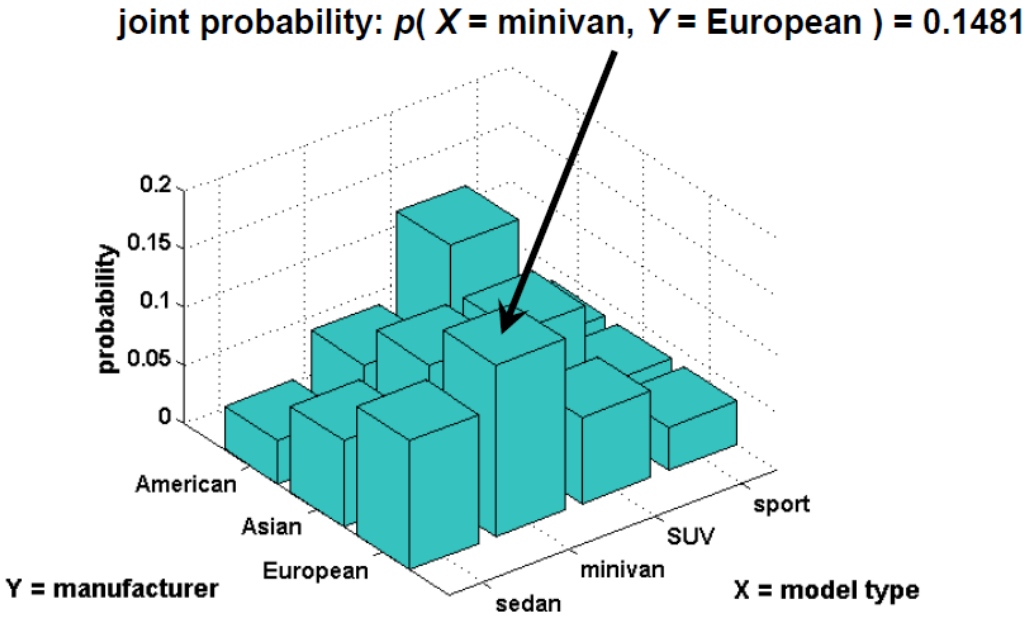
$$\mathbb{P}(X \in B) = \int_B f_X(x)dx$$

$$P(X \in [a,b]) = \int_a^b P(X)dX$$



联合概率分布

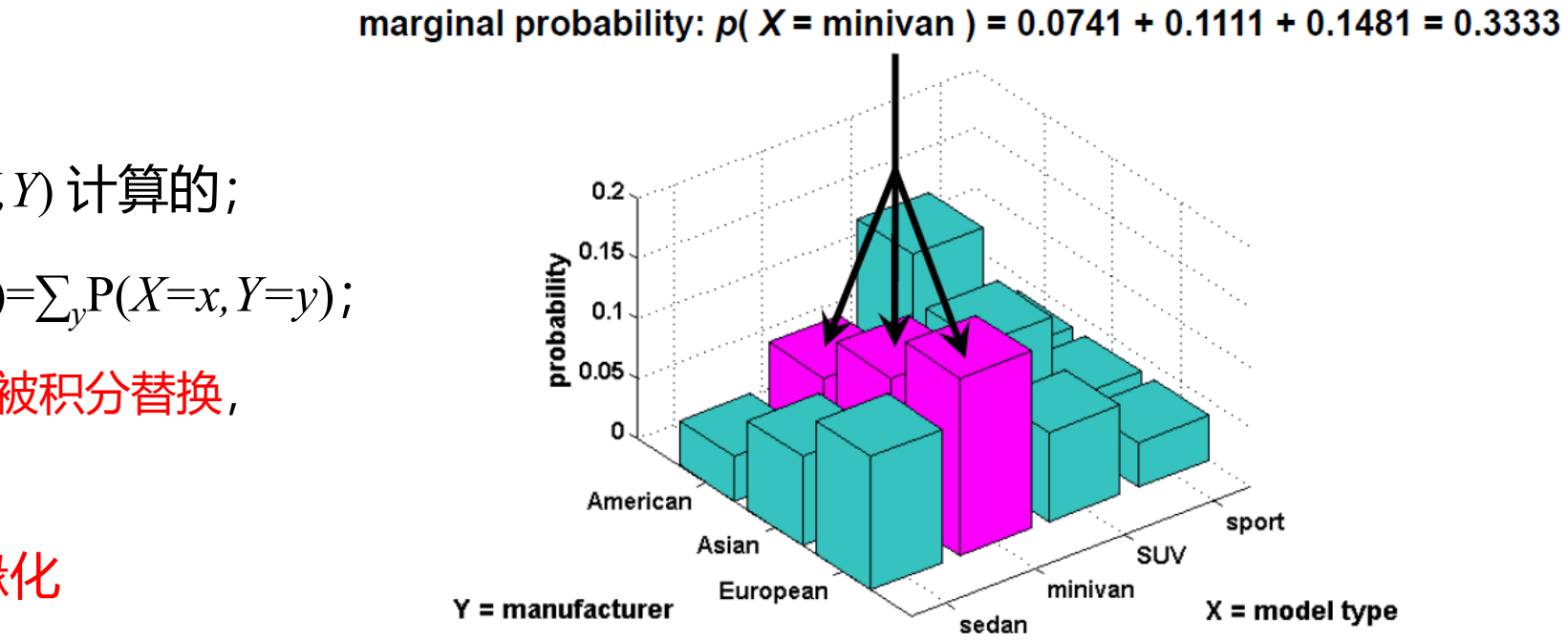
- 同时作用于多个变量的概率分布被称为联合概率分布。
- 给定两个随机变量X和Y的任意值 x 和 y，$X=x$ 和 $Y=y$ 同时发生的概率是多少?
  - $P(X=x,Y=y)$ 表示联合概率；
  - 为了简便，我们也可以将 P(x,y) 写作：$P(X,Y)$



例子：整车制造商品牌和车型的联合概率分布

边缘概率分布

- 边缘概率分布是多元分布条件下单个变量的概率分布：
  - 它是根据联合概率分布 $\mathrm{P}(X,Y)$ 计算的；
  - 即，使用求和规则：$\mathrm{P}(X=x)=\sum_y \mathrm{P}(X=x,Y=y)$；
    - 对于连续随机变量，求和被积分替换，$\mathrm{P}(X=x)=\int \mathrm{P}(X=x,Y=y)\mathrm{d}y$；
  - 求边缘概率的过程称为边缘化（marginalization）。



条件概率和Bayes定理

- **条件概率分布是指在另一个变量已经取某个特定值的条件下，一个变量的概率分布：**
  - 记作 P(X=x|Y=y)：绝对概率是不存在的，所有概率都取决于背景信息，在此，仅以Y为背景信息。
  - $P(X = x, Y = y) = \frac{P(X=x,Y=y)}{P(Y=y)}$。
- **当已知另一个变量的条件概率时，允许计算一个变量的条件概率**
  - $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$
  - 获得Bayes定理用到了如下联合分布的乘法规则：$P(X,Y) = P(Y|X)P(X)$;
  - 由对称性，我们也有$P(Y,X) = P(X,Y) = P(X|Y)P(Y)$;
  - P(X)：先验概率（prior probability），对 X 的初始信任程度；P(Y)：证据（evidence）；
  - P(X|Y)：后验概率（posterior probability），在结合 Y 的知识后对X情况的信任程度；
  - P(Y|X)：给定 X 时， Y 的似然性（likelihood）；

Bayes定理（续）+ 独立分布

- 贝叶斯定理：$\text{后验概率} = \frac{\text{似然性} \times \text{先验概率}}{\text{证据}}$。
- 独立分布（independent）：两个随机变量 X 和 Y 是独立的，如果 Y 的发生不透露关于 X 发生情况的任何信息
  - 例如，连续掷两次骰子是独立的。
- 我们可以写出：$P(X|Y)=P(X)$
  - 使用以下符号：$X \perp Y$;
  - 对于独立的随机变量：$P(X,Y)=P(X)P(Y)$。
- 在所有其他情况下，随机变量是相互依赖（dependent）的
  - 例如，随机喷泉连续喷发的持续时间;
  - 从一副牌中连续抽到King（抽出的牌不放回）。

联合概率下的链式法则和全概率公式

- 在背景B下，N个命题联合的概率即$P(A_1, A_2, \ldots, A_N|B)$，链式法则可写为
  - $P(A_1, A_2, \ldots, A_N|B) = P(A_1|A_2, A_3 \ldots, A_N, B)\underbrace{P(A_2|A_3 \ldots, A_N, B) \ldots P(A_N|B)}_{\text{剩余部分的联合概率}}$

递归“剥离”最左侧变量，逐步条件化

- 全概率公式：考虑样本空间中的一组由互斥事件组成的分区
  - $\Omega = B_1 \cup B_2 \cup \ldots \cup B_n$且$B_i \cap B_j = \emptyset$;
  - 考虑另一个事件$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)$，则有总概率公式

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

- 全概率公式下的贝叶斯定理
  - $P(B|A) = \frac{P(B)P(A|B)}{\sum_i P(A|B_i)P(B_i)}$。

期望和方差

- 离散变量的期望：$\mathbb{E}_{X\sim P}[f(X)] = \sum_X P(X)f(X)$

- 连续变量的期望：$\mathbb{E}_{X\sim P}[f(X)] = \int P(X)f(X)\,dX$
  - 例如，连续掷两次骰子是独立的。

- 期望给出了一个函数$f(x)$在概率分布$P(x)$下，自变量x采样所得的平均值（期望值）。

- 方差的定义式：$\mathrm{Var}\big(f(X)\big) = \mathbb{E}\big[(f(X) - \mathbb{E}[f(X)])^2\big]$

- 方差用来衡量函数$f(x)$在概率分布$P(x)$下，由自变量x采样所得值偏离期望的情况。
  - 方差通常用 σ^2 表示；
  - 上述方程类似于由函数 $f(X_i) = X_i - \mu$ 求原点二阶矩，
  - 对离散变量情况有 $\sigma^2 = \sum_i P(X_i)\cdot(X_i - \mu)^2$

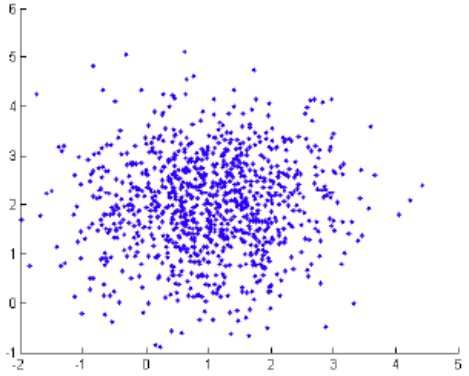
协方差（Covariance）

- 协方差衡量两个随机变量之间线性相关程度的大小：

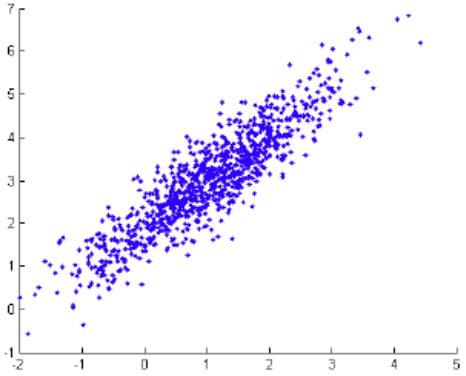
$$\mathrm{Cov}\big(f(X), g(Y)\big) = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(Y) - \mathbb{E}[g(Y)])]$$

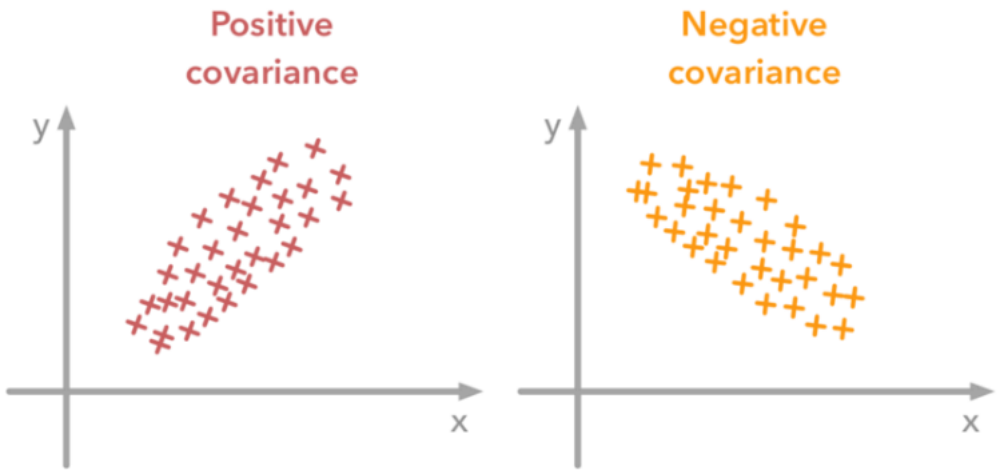
- 协方差衡量 X 和 Y 同时（或相反）偏离其均值的趋势。

协方差很小



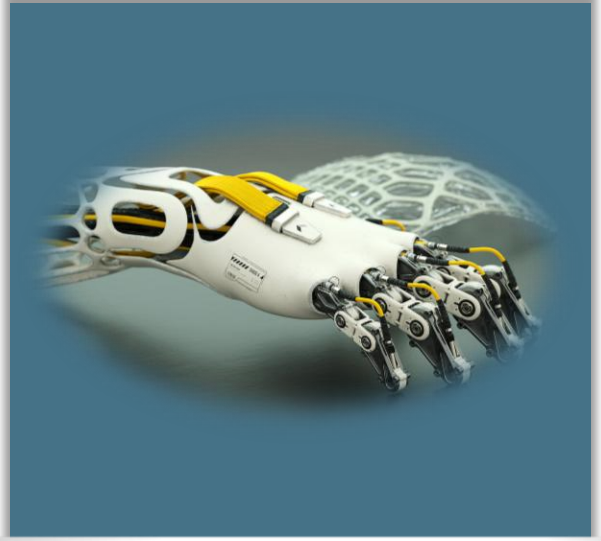
协方差很大



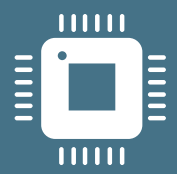


- 相关系数：是协方差除以两个变量的标准差得到的标准化值

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$



机器学习
导论




1. 机器学习中的常用符号
2. 矩阵运算回顾
3. 概率论回顾
4. 优化理论初步

梯度

- **梯度的定义：多元函数$f(\mathbf{x})$的梯度是由其所有输入变量的偏导数组成的向量。**
  - 输入向量$\mathbf{x} = [x_1, x_2, \ldots, x_n]$，梯度表示为：

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

    - 符号∇（读作Nabla）表示梯度，常简写为$\nabla f(\mathbf{x})$或$\nabla_{\mathbf{x}} f$。
    - 物理意义：梯度方向是函数值增长最快的方向，模长表示增长率。
- 初步思路：在机器学习中，梯度下降算法通过沿着所构建的损失函数$\mathcal{L}$对模型参数集θ的梯度$\nabla_{\theta}\mathcal{L}$的相反方向（负梯度方向）更新参数，以最小化损失函数。

优化问题概述

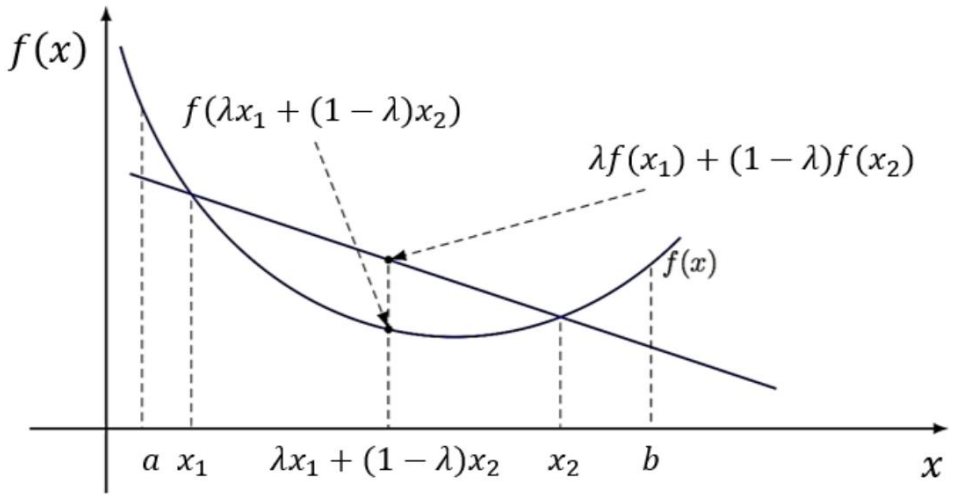
- **优化（Optimization）问题涉及**
  - 1.优化目标函数；
  - 2. 在有效参数空间中，寻找能够最小化或最大化该函数值的参数。
- **大多数优化算法以最小化多元变量函数 $f(\mathbf{x})$ 的形式表示**
  - 最大化目标函数可通过最小化其负值实现（例如，最小化 $-f(\mathbf{x})$）；
  - 在最小化问题中，目标函数通常被称为成本函数、损失函数或误差函数。
- **机器学习中的大多数优化问题都是非凸的（思考：什么是凸问题？）**
  - 这意味着损失函数不是凸函数。
  - 尽管如此，针对凸问题的算法设计与分析仍对机器学习领域的发展具有重要指导意义。

（补充）凸函数

- **数学上，某函数$f(x)$是凸函数，若对所有x_1，x_2，及对所有$\lambda \in [0,1]$都有**

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$$

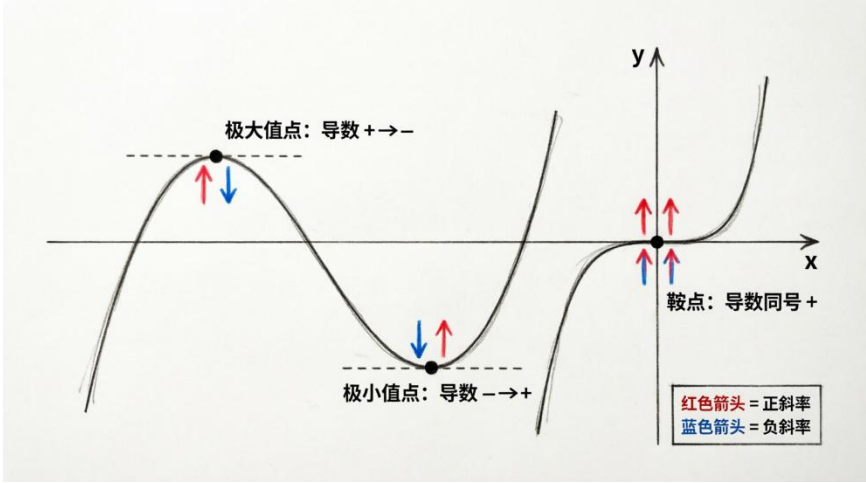
- 以单变量函数为例：



- 即：若连接函数图像上任意两点的线段在任何点都不位于函数图像下方，则称该单变量函数为凸函数。
- **凸函数必然存在全局最小值**：凸函数的任何局部最小值同时也是全局最小值。

函数优化中的极点

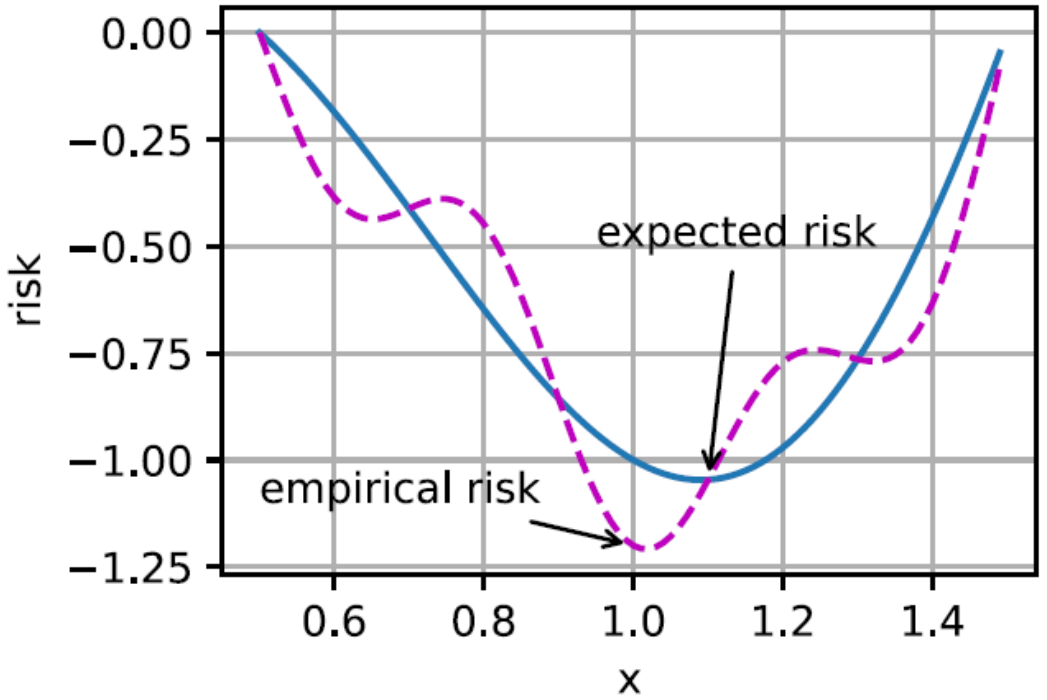
- **以单变量可微函数 $f(x)$为例，驻点（或临界点）是指它的导数为零的点，即：$f'(x)=0$。驻点按函数特性可分为（其中，极大值和极小值点统称极点）**
  - **极小值点**：导数从负变正的点。
  - 极大值点：导数从正变负的点。
  - **鞍点**：函数的导数在这一点的两侧保持同号（正或负）。



- **驻点性质的二阶判据**
  - 若单变量函数$f''(x)>0$，该点为极小值；进一步地，全局最小值对应多变量函数的**Hessian矩阵 $\nabla^2 f(x)$正定**（凸函数）；
  - 类似地，若单变量函数$f''(x)<0$，该点为极大值；进一步地，全局最大值对应多变量函数的**Hessian矩阵 $\nabla^2 f(x)$负定**（凹函数）；
  - 若单变量函数$f''(x)=0$，结果不确定，可能是鞍点，也可能不是。进一步地若多变量函数的Hessian矩阵不定，则结果不确定。

机器学习模型训练与优化问题的关联

- **有监督机器学习的目标：找到合适的模型以预测数据**
  - 最小化目标函数：针对训练样本集设计目标函数，降低训练误差。
  - 同时，针对测试样本集，降低泛化误差。
- **经验函数与期望函数**
  - 经验函数g：通过优化算法寻找经验风险最小点。
  - 期望函数f：基于有限训练数据生成的期望风险曲线。
- **机器学习算法的核心任务**
  - 通过最小化测试样本集误差，寻找期望风险最小点。
  - 该最小点可能与训练集的最小点位置不同。
  - 形式上可能并非严格的最小值（如存在过拟合或欠拟合）。
  - 为此，需在训练目标函数的设计上引入约束等额外信息。



基于梯度的一阶优化技术简介

- **以任一有N维输入的函数为例，假其设在满足二阶最小值条件的前提下，在点v处有**

$$\frac{\partial}{\partial \omega_1} g(v) = 0$$

$$\frac{\partial}{\partial \omega_2} g(v) = 0$$

$$\vdots$$

$$\frac{\partial}{\partial \omega_N} g(v) = 0$$

  - 则其梯度形式 $\nabla g(v) = \mathbf{0}_{N\times 1}$ 形成一阶最优性条件。

- **实际应用中的难点**
  - 实际上，几乎不可能以代数方式求解上述方程组的闭合解。
  - 因此，需要引入迭代算法求解。

基于梯度方向的参数更新

梯度和目标函数最速下降方向

- **一个多输入函数$g(\omega)$在给定参数点$\boldsymbol{\omega}^0$处可由一个超平面$h(\omega)$局部逼近：**

$$h(\omega) = g(\omega^0) + \nabla g^T(\omega^0)(\omega - \omega^0)$$

  - 上式可重写为 $h(\boldsymbol{\omega}) = a + \boldsymbol{b}^T\omega$，其中

$$a = g\left(\boldsymbol{\omega}^0\right) - \nabla g\left(\boldsymbol{\omega}^0\right)^T \boldsymbol{\omega}^0 \text{和} \boldsymbol{b} = \nabla g\left(\boldsymbol{\omega}^0\right)$$

  - 这一超平面为$g(\cdot)$在$\boldsymbol{\omega}^0$处的泰勒级数逼近，也是$g(\cdot)$在这一点的正切值。

- **梯度下降**

  - 函数g(ω)在一个特定点的负梯度$-\nabla g(\omega)$总是定义了一个在该点的一个合理的下降方法。
  - 这类方法中所含的参数更新步骤可以表示为通式：$\omega^k = \omega^{k-1} - \alpha d^k$ 。
  - 单步参数更新具有以下形式：

$$\omega^k = \omega^{k-1} - \alpha \nabla g\left(\omega^{k-1}\right)$$

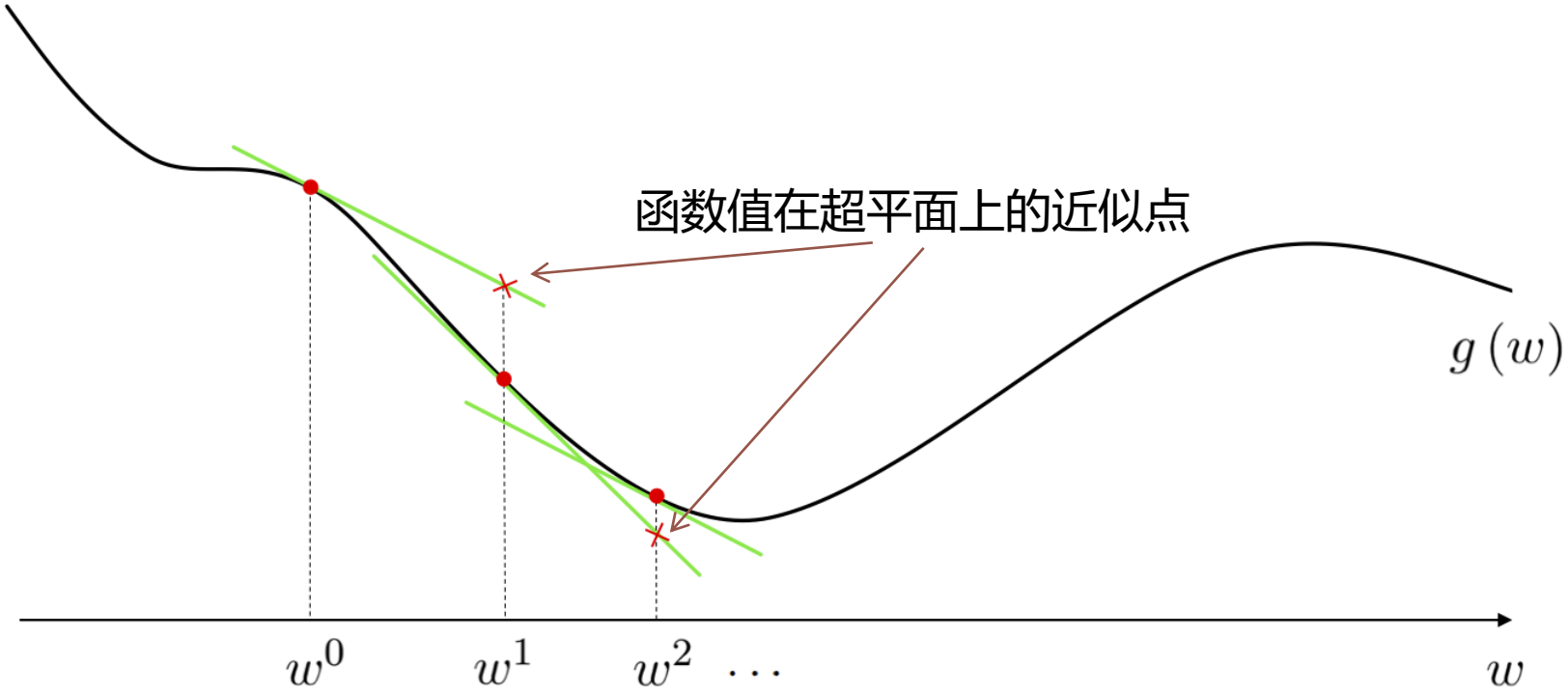
基于梯度方向的参数更新（续）

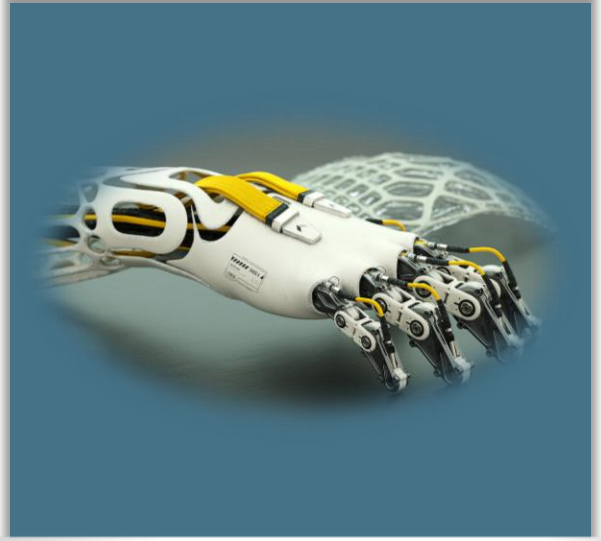
- **梯度下降标准公式**

$$\omega^k = \omega^{k-1} - \alpha \nabla g\left(\omega^{k-1}\right)$$

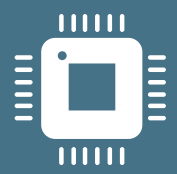
  - 其中，α为更新步长。

- **迭代执行此步骤，可以得到一个与函数g的局部最小值临近的解（见图示）**





机器学习
的数学基础




讨论