中国科学技术大学

University of Science and Technology of China

人工智能讲义

机器学习中的线性代数

March 27, 2018

Outline

主要符号及其含义

$$\mathbf{A}_{n\times m}=[\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_m]=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}^T$$

解释说明

- $a_{ij}\in\mathcal{R}, i=1,2,\ldots,m, \quad j=1,2,\ldots,n$
- 向量都默认为“列向量”，小写黑体字母表示，矩阵是大写黑体表示。$(x_1,x_2,\ldots,x_n)$或$[x_1,x_2,\ldots,x_n]$ 表示行向量，上述矩阵中列向量 $\mathbf{a}_i=[a_{i1},a_{i2},\ldots,a_{in}]^T, i=1,2,\ldots,m$
- 简单写法：$\mathbf{A}=[a_{ij}], i=1,2,\ldots,m, \quad j=1,2,\ldots,n$
- 矩阵或向量的转置 $\mathbf{A}\rightarrow\mathbf{A}^T$
- 矩阵的逆 $\mathbf{A}\rightarrow\mathbf{A}^{-1}$ （若存在）
- 方阵的行列式 $\mathbf{A}\rightarrow|\mathbf{A}|$
- 单位阵 $\mathbf{I}$，空矩阵 $\mathbf{0}$
- 方阵的迹 $tr\mathbf{A}=\Sigma_{i=1}^{n}a_{ii}$，方阵对角线的和

矩阵的一些概念

- 矩阵对称：$\mathbf{A} = \mathbf{A}^T$
- Hermite/厄米特阵：$\mathbf{A}^* = \mathbf{A}$，即矩阵的共轭转置不变，对称矩阵也是 Hermite 的。特例：实对称阵
- 斜矩阵对称：$-\mathbf{A} = \mathbf{A}^T$
- 奇异矩阵：行列式为 0，或非满秩的矩阵
- 任给矩阵 $\mathbf{A}$，$\mathbf{A} = S(\mathbf{A}) + C(\mathbf{A})$，其中 $S(\mathbf{A}) = \frac{\mathbf{A}+\mathbf{A}^T}{2}$ 是对称的，$C(\mathbf{A}) = \frac{\mathbf{A}-\mathbf{A}^T}{2}$ 是斜对称的。

对称矩阵

$$\begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & 2 \\ -3 & 2 & 1 \end{bmatrix}$$

厄米特阵

$$\begin{bmatrix} 1 & 2+i & -3 \\ 2-i & 4 & 2 \\ -3 & 2 & 1 \end{bmatrix}$$

斜对称矩阵

$$\begin{bmatrix} 0 & 2 & -3 \\ -2 & 0 & -2 \\ 3 & 2 & 0 \end{bmatrix}$$

矩阵的一些概念

- 正交矩阵：$\mathbf{A}\mathbf{A}^T = \mathbf{I}$，转置就是逆
- 酉矩阵：$\mathbf{A}^*\mathbf{A} = \mathbf{I}$，任意不同两行表示的向量正交，正交矩阵是酉矩阵为实矩阵时的特例
- 正规矩阵（normal matrix）：$\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$，等价于通过酉变换实现 $\mathbf{A}$ 的对角化。正规矩阵的例子：实对称矩阵，实反对称，正交阵，厄米特阵，反厄米特阵，酉矩阵等。

正交矩阵

$$\begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$

酉矩阵

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

正规矩阵

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

一句话：通常是 $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$，但是正交矩阵把 $^{-1}$ 换成了 ^{T}。

酉矩阵的性质

- 定义：酉矩阵 $\mathbf{U}$, 满足 $\mathbf{U}\mathbf{U}^* = \mathbf{I}$
- 酉矩阵的行向量组构成标准正交向量组，可视为 $\mathbf{C}^n$ 的标准正交基；同样对列向量组成立；
- 酉矩阵行列式的值为 1，即 $|det(\mathbf{U})| = 1$
- 酉矩阵的逆矩阵还是酉矩阵，即 $\mathbf{U}^{-1} = \mathbf{U}^*$
- 酉矩阵的特征值长度为 1，即 $|\lambda(\mathbf{U})| = 1$
- 两个酉矩阵的乘积依然是酉矩阵

一句话：酉矩阵是定义在 $\mathcal{C}$ 上的标准正交基组

定义

- 输入：$\mathbf{A}: m \times n, \mathbf{B}: n \times p$
- 输出：$\mathbf{C} = \mathbf{AB}, \mathbf{C}_{m \times p} = [c_{ij}], c_{ij} = \Sigma_{k=1}^{n} a_{ik} b_{kj}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, p$

转置与矩阵乘法

- 已知：$\mathbf{A} \in \mathcal{R}^{m \times n}, \mathbf{B} \in \mathcal{R}^{n \times p}$，若 $\mathbf{C} = \mathbf{AB}, c_{ij} = \Sigma_{k=1}^{n} a_{ik} b_{kj}$
- 则有：$\mathbf{C}^T = \mathbf{B}^T \mathbf{A}^T, c_{ji} = \Sigma_{k=1}^{n} a_{jk} b_{ki}$

计算时间复杂度

- $O(mnp)$，三次方量级

矩阵与向量相乘

- 输入 $\mathbf{A}: m \times n, \mathbf{x}: n \times 1$
- 输出 $\mathbf{z}_{m\times 1} = \mathbf{A}\mathbf{x}, z_i = \Sigma_{j=1}^{n} a_{ij}x_j, i = 1, 2, \ldots, m$
- 评述：右侧列向量乘完后得到列向量。 矩阵乘向量, $O(mn)$

- 输入 $\mathbf{A}: m \times n, \mathbf{x}: m \times 1$
- 输出 $\mathbf{z}_{n\times 1}^{T} = \mathbf{x}^{T}\mathbf{A}, z_i = \Sigma_{j=1}^{m} a_{ij}x_j, i = 1, 2, \ldots, n$
- 评述：左侧行向量乘完后得到行向量。 向量乘矩阵, $O(mn)$

- 输入：$\mathbf{A}, m \times n, \mathbf{x}: n \times 1, \mathbf{y}: m \times 1$
- 输出：$\alpha = \mathbf{y}^{T}\mathbf{A}\mathbf{x} = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n} y_i a_{ij} x_j$
- 评述：最终得到标量。 向量乘矩阵乘向量, $O(mn)$

矩阵相关的扩展概念

何为空间?

- 空间：在集合概念的基础上，对元素定义了位置点及其相对关系，位置点之间可定义变换/运动关系；
- 线性空间：向量空间；增加长度概念，得到赋范线性空间；增加长度和角度概念，得到内积空间；线性空间中的任何一个点对象都可以表示为一个“向量”；距离空间，定义了点之间的距离。

何为变换?

- 所谓变换/运动，把一个线性空间中的对象（点、线、形状、子空间等）变换到另一个位置；
- 也可以看成是空间中对象未动，坐标系的变换，即理解为标准基向量组进行了旋转和平移。

何为矩阵?

- 是线性空间的基向量组；
- 是变换/运动，线性空间中选定基之后，向量刻画点对象，矩阵刻画对象的运动，用矩阵和向量的乘积表示对向量描述的对象施加运动

线性空间/向量空间的基

- n 维线性空间 $\mathbf{V}$ 的任意 n 个线性无关向量都是空间 $\mathbf{V}$ 的基.
- 点的坐标是向量。给定基向量组为 $\alpha_1, \alpha_2, \ldots, \alpha_n$，任意点 β 可以表示成基的线性组合：$\beta = x_1\alpha_1 + x_2\alpha_2 + \ldots + x_n\alpha_n$，向量 $(x_1, x_2, \ldots, x_n)$ 被称为点对象的“坐标”，该坐标的参照物是基向量组 $\alpha_1, \alpha_2, \ldots, \alpha_n$. 基向量组的每个分量定义了不同的方向以及每个方向上的单位长度。
- 用例子理解基向量组：在标准二维平面中任意画两个从原点出发的箭头/矢量（不同向，也不反向），以它们为基向量；平面中任何其它点/原点发出的箭头，在基向量上的投影相对于基向量的长度就是坐标。
- 完整描述内积空间的一个点对象/列向量/坐标 β，其矩阵运算形式有三种：
  - 为 $\beta^T = (x_1, x_2, \ldots, x_n)(\alpha_1, \alpha_2, \ldots, \alpha_n)^T$，向量乘矩阵/向量，右乘矩阵的行构成的基向量组
  - $\beta = (\alpha_1, \alpha_2, \ldots, \alpha_n)(x_1, x_2, \ldots, x_n)^T$，向量/矩阵乘向量，左乘矩阵的列构成基向量组
  - $\beta = (x_1, x_2, \ldots, x_n) \cdot (\alpha_1, \alpha_2, \ldots, \alpha_n)$ 内积形式
- 内积空间的标准正交基：任何两个基向量内积为 0；任何基向量的范数（长度）都是 1.
- 标准正交基的例子：$\mathbf{I} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \ldots \quad \mathbf{e}_n]$, 其中 $\mathbf{e}_1 = (1, 0, \ldots, 0)^T, \mathbf{e}_2 = (0, 1, 0, \ldots, 0)^T, \ldots, \mathbf{e}_n = (0, 0, \ldots, 0, 1)^T$.
- 若给定点对象的坐标，而省略参照的基向量组，那么默认基向量组是单位矩阵 $\mathbf{I}$。
- 矩阵每一行是个点对象，因此矩阵也有参照的基向量组。

$$\begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$

标准基向量组

矩阵与向量相乘的意义

理解一：向量在线性空间中某个特定基向量组下的表示

- $\mathbf{Ax}$：向量/点对象在 $\mathbf{A}$ 的列构成的基向量组中的坐标是 $\mathbf{x}$
- $\mathbf{x}^T\mathbf{A}$：向量/点对象在 $\mathbf{A}$ 的行构成的基向量组中的坐标是 $\mathbf{x}$
- 上述两个式子运算结果分别是列向量和行向量，可以理解为在单位矩阵 $\mathbf{I}$ 描述的线性空间中点对象的坐标
- 因此矩阵乘向量或者向量乘矩阵，都可以看成是“坐标变换”，特定基向量组下的坐标变换成“标准坐标”（单位阵 $\mathbf{I}$ 下）

从例子理解：

- 计算两个向量 $\mathbf{x}, \mathbf{y}$ 的相似性的“核矩阵方法”：$k(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbf{Ax}$
- 其中“核矩阵” $\mathbf{A}$ 是正定矩阵 $(\forall \mathbf{z} \neq \mathbf{0}, \mathbf{z}^T\mathbf{Az} > 0)$，存在唯一对角线都是正数下三角阵 $\mathbf{L}$，满足 $\mathbf{A} = \mathbf{LL}^T$，即 Cholesky/乔列斯基分解
- $k(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbf{Ax} = \mathbf{y}^T\mathbf{LL}^T\mathbf{x} = (\mathbf{y}^T\mathbf{L})(\mathbf{L}^T\mathbf{x})$，也就是说 $\mathbf{x}, \mathbf{y}$ 是在 $\mathbf{L}$ 行为基的线性空间中的坐标；将两个点对象同时转换到单位矩阵 $\mathbf{I}$ 描述的线性空间中，计算它们的相似性；这就是计算两个向量相似性的核矩阵方法隐藏在深处的“理由”
- 深度思考：行为基向量组的 $\mathbf{L}$，基向量和此基向量组下的坐标 $\mathbf{x}$ 的意义是什么？

矩阵是基向量组，描述线性空间

过渡矩阵：线性空间可用矩阵来描述，描述方法的变换

- $\forall \zeta \in \mathbf{V}, \mathbf{V}$ 是一 n 维线性空间，设 $\alpha_1, \alpha_2, \ldots, \alpha_n$ 和 $\beta_1, \beta_2, \ldots, \beta_n$ 是 $\mathbf{V}$ 的两组不同的基，α_i, β_i 分别是 n 维向量，在此两组基下，ζ 的坐标分别为 $x_1, x_2, \ldots, x_n$ 和 $y_1, y_2, \ldots, y_n$
- 过渡矩阵，即，存在方阵

$$\mathbf{A}_{n\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

使得 $[\beta_1, \beta_2, \ldots, \beta_n] = [\alpha_1, \alpha_2, \ldots, \alpha_n]\mathbf{A}$ 成立，即存在可逆矩阵 $\mathbf{A}$ 实现两组基之间的转换。

- 基变换的例子：二进制数转换为五进制；
$(2^0, 2^1, \ldots, 2^k, \ldots,) \Longrightarrow (5^0, 5^1, \ldots, 5^k, \ldots)$
- 我们通常用十进制作为“桥梁”，掌握不同进制和十进制的相互转换即可。
$(2^0, 2^1, \ldots, 2^k, \ldots,) \Longrightarrow (10^0, 10^1, \ldots, 10^k, \ldots)$
$(10^0, 10^1, \ldots, 10^k, \ldots,) \Longrightarrow (5^0, 5^1, \ldots, 5^k, \ldots)$

基变换的特殊情形：单位矩阵描述的标准正交基，当作“桥梁”

- 前述两个基向量组中，我们假设 $[\beta_1, \beta_2, \ldots, \beta_n]$ 是标准单位基，即单位矩阵 $\mathbf{I}$(如十进制)，讨论基变换。
- 把基向量组 $[\alpha_1, \alpha_2, \ldots, \alpha_n]$ 变换为 $\mathbf{I}$，此时过渡矩阵就是逆矩阵 $[\alpha_1, \alpha_2, \ldots, \alpha_n]^{-1}$
- 反之，从标准正交基 $\mathbf{I}$ 转换到任意基向量组 $[\alpha_1, \alpha_2, \ldots, \alpha_n]$，有 $[\alpha_1, \alpha_2, \ldots, \alpha_n] = \mathbf{I}[\alpha_1, \alpha_2, \ldots, \alpha_n]$，即此时基变换对应的过渡矩阵就是该基向量组本身表示的矩阵 $[\alpha_1, \alpha_2, \ldots, \alpha_n]$。

过渡矩阵与基变换

- 依据坐标的含义，我们有 (对比 k 进制转换为十进制，不通基向量组下点转换到 $\mathbf{I}$ 下)

$$\zeta = [\alpha_1, \alpha_2, \ldots, \alpha_n][x_1, x_2, \ldots, x_n]^T = [\beta_1, \beta_2, \ldots, \beta_n][y_1, y_2, \ldots, y_n]^T$$
$$\because [\beta_1, \beta_2, \ldots, \beta_n] = [\alpha_1, \alpha_2, \ldots, \alpha_n]\mathbf{A},$$
$$\Rightarrow \zeta = [\alpha_1, \alpha_2, \ldots, \alpha_n]\mathbf{A}[y_1, y_2, \ldots, y_n]^T$$
$$\Rightarrow [x_1, x_2, \ldots, x_n]^T = \mathbf{A}[y_1, y_2, \ldots, y_n]^T$$

变换基的过渡矩阵 $\mathbf{A}$ 把不同基下的点的坐标进行了变换。

- 注意到细节：过渡矩阵 $\mathbf{A}$ 把基 α 变换为 β (方阵乘方阵)，但是坐标变换是把 β 下的坐标变换为 α 下 (方阵乘向量)；换句话说就是变换基的过渡矩阵的逆矩阵用来做对应的坐标变换：

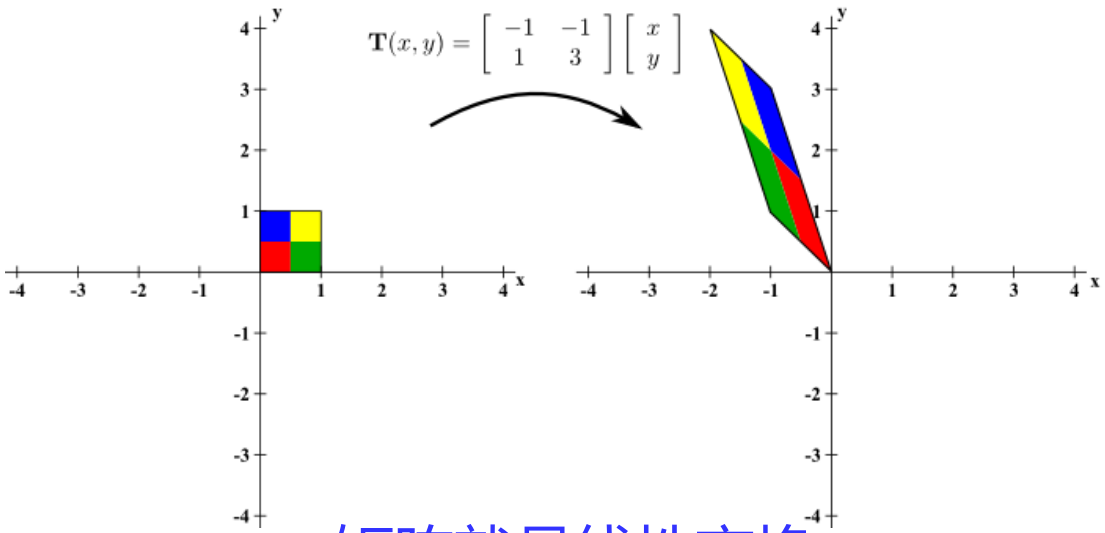
$$[y_1, y_2, \ldots, y_n]^T = \mathbf{A}^{-1}[x_1, x_2, \ldots, x_n]^T$$

过渡矩阵的逆矩阵作用在一个点对象的坐标向量上，实现点对象的坐标变换。

线性变换的概念和例子

理解线性变换：

- 把对象 $\mathbf{p}$ 用一次函数/线性函数变换为对象 $\mathbf{q}$，假设对象 $\mathbf{p}, \mathbf{q}$ 都可以用 $n-$ 维实数向量来描述；
- 对象 $\mathbf{p}, \mathbf{q}$ 在不同的线性空间中，变换就是“线性映射”，在同一个空间中就是“线性变换”
- 如图所示 2-D 的例子，平面上的任意一点对象被线性变换 $T(x, y)$ 变换为另一个对象；彩色方块用于展示部分点的变化
- 注意思考：哪些点保持不变？



矩阵就是线性变换

- 点对象坐标的观点：$\zeta = (\alpha_1, \ldots, \alpha_n)(x_1, \ldots, x_n)^T$ 是将给定点对象在基向量组 $(\alpha_1, \ldots, \alpha_n)$ 下的坐标 $\mathbf{x}$ 变换到单位矩阵 $\mathbf{I}$ 下的坐标 ζ；点对象未动
- 线性变换的观点：若基向量组 $\mathbf{I}$ 和 $(\alpha_1, \ldots, \alpha_n)$ 描述的线性空间是同构的，$(\alpha_1, \ldots, \alpha_n)$ 描述的线性空间中，坐标为 $(x_1, x_2, \ldots, x_n)^T$ 的点对象坐标要经线性变换，变为 $\mathbf{I}$ 描述的线性空间中的点 ζ；空间和参照坐标系不动，点对象动

线性变换与相似矩阵

- 设线性空间 $\mathbf{V}$ 有两基向量组 α, β，且有过渡矩阵 $\mathbf{P}$ 满足 $\beta = \alpha\mathbf{P}$，可知对线性空间中的点对象在不同基向量组下的坐标 $\mathbf{x}, \mathbf{y}$ 有关系 $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$
- 若同一个线性变换在两组不同的基 α, β 下分别为 $\mathbf{A}, \mathbf{B}$，则有 $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$

Proof.

同一线性变换从不同基向量组将同一点对象变换到 $\mathbf{I}$ 下，故有
$\zeta = \alpha\mathbf{A}\mathbf{x} = \beta\mathbf{B}\mathbf{y}$
又 $\because \beta = \alpha\mathbf{P}, \mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$，代入上式，有
$\alpha\mathbf{A}\mathbf{x} = \alpha\mathbf{P}\mathbf{B}\mathbf{P}^{-1}\mathbf{x}$，即 $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ □

- 其中 $\mathbf{P}$ 是相似矩阵，也是两组基的过渡矩阵
- 相似的实用意义：$\Lambda = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$，若能把线性变换 $\mathbf{A}$ 相似成对角阵 Λ，那么可以对 Λ 做各种应该对 $\mathbf{A}$ 进行的运算，算完后再相似回去即可，对对角阵 Λ 的计算总是相对要简单很多。

向量函数的导数

- 已知：$\mathbf{x} \in \mathcal{R}^{n\times 1}, \mathbf{y} \in \mathcal{R}^{m\times 1}, \mathbf{y} = \Psi(\mathbf{x})$
- 则有：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \mathbf{x}} \\ \frac{\partial y_2}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \mathbf{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathbf{y}}{\partial x_1} \\ \frac{\partial \mathbf{y}}{\partial x_2} \\ \vdots \\ \frac{\partial \mathbf{y}}{\partial x_n} \end{bmatrix}^T = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- 评述：$\mathbf{y}$ 的每个分量对 $\mathbf{x}$ 的每个分量分别求偏导数，获得的 Jacobian 矩阵。$\mathbf{y}, \mathbf{x}$ 可以分别为标量，此时是上述情形的特例。

线性变换的导数：

- 已知：$\mathbf{y} \in \mathcal{R}^{m\times 1}, \mathbf{A} \in \mathcal{R}^{m\times n}, \mathbf{x} \in \mathcal{R}^{n\times 1}$，且 $\mathbf{y} = \mathbf{A}\mathbf{x}$
- 则有：$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$
- 证明：$y_i = \Sigma_{k=1}^{n} a_{ik}x_k, \Rightarrow \frac{\partial y_i}{\partial x_j} = a_{ij}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n.$
  $\Rightarrow \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$
- 评述：线性函数/线性变换的导数就是一次项/线性项的系数。

特例：

- 已知：当上述条件中 $\mathbf{A}$ 是 $1 \times n$ 的行向量时，
- 则有：$\mathbf{y}$ 为标量，$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$，导数和 $\mathbf{A}$ 一样，是行向量。

矩阵乘法与求导数

复合函数的导数：

- 已知：$\mathbf{y} \in \mathcal{R}^{m\times 1}, \mathbf{A} \in \mathcal{R}^{m\times n}, \mathbf{x} \in \mathcal{R}^{n\times 1}$，且 $\mathbf{y} = \mathbf{A}\mathbf{x}, \mathbf{x} = g(\mathbf{z})$
- 则有：$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \mathbf{z}} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}$

特例：

- 已知：$\alpha = \mathbf{y}^T\mathbf{A}\mathbf{x} = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n} y_i a_{ij} x_j$，

  则有：$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{y}^T\mathbf{A}$ $(\because \frac{\partial(\mathbf{y}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{y}^T\mathbf{A})$

  $\frac{\partial \alpha}{\partial \mathbf{y}} = \mathbf{x}^T\mathbf{A}^T$ $(\because \frac{\partial \alpha}{\partial \mathbf{y}} = \frac{\partial \alpha^T}{\partial \mathbf{y}} = \frac{\partial(\mathbf{x}^T\mathbf{A}^T\mathbf{y})}{\partial \mathbf{y}} = \mathbf{x}^T\mathbf{A}^T)$
- 若增加条件 $\mathbf{x} = \mathbf{y}$ 且 $\mathbf{A}$ 为方阵，则有 $\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$

Proof.

$\alpha = \mathbf{x}^T\mathbf{A}\mathbf{x} = \Sigma_{i=1}^{n}\Sigma_{j=1}^{n} x_i a_{aj} x_j$

$\Rightarrow \frac{\partial \alpha}{\partial x_k} = \Sigma_{j=1}^{n} a_{kj}x_j + \Sigma_{i=1}^{n} a_{ik}x_i, k = 1, 2, \ldots, n$

$\rightarrow \frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^T\mathbf{A}^T + \mathbf{x}^T\mathbf{A} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T)$ □

特征向量与特征值

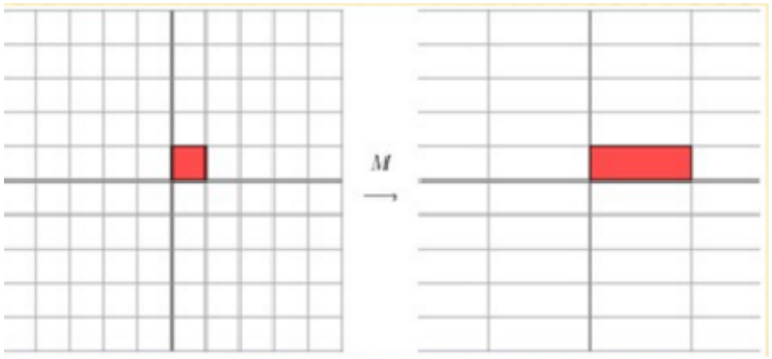
定义：

- 给定矩阵 $\mathbf{A}$，若有 $\mathbf{Ax} = \lambda\mathbf{x}$，则称标量 λ 为特征值，$\mathbf{x}$ 为 λ 对应的特征向量；

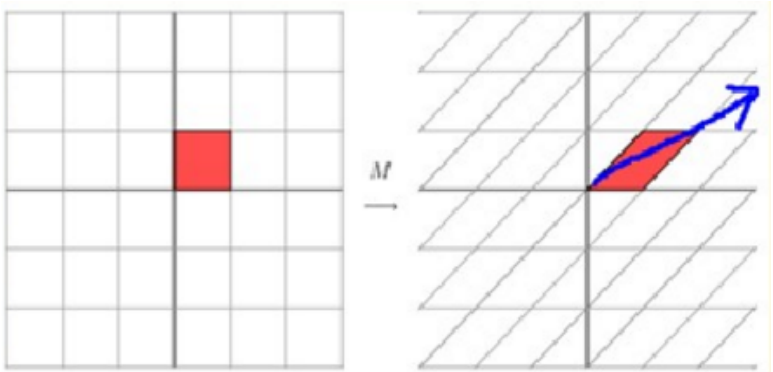
理解特征值与特征向量：

- 若 $\mathbf{A}$ 解释为基向量组 $\mathbf{I}$ 下的线性变换，则 $\mathbf{Ax}$ 把向量/点 $\mathbf{x}$ 变换到新的向量/点 $\lambda\mathbf{x}$，也就是变换前后，在线性空间 $\mathbf{I}$ 中，点对象/向量/箭头的表示（方向和长度等）仅仅是“缩放”效果；
- 若 $\mathbf{A}$ 的列解释为基向量组，线性空间在基向量组 $\mathbf{A}$ 下点对象的坐标为 $\mathbf{x}$，在标准基 $\mathbf{I}$ 下的坐标为 $\mathbf{Ax}$，正好等于 $\mathbf{x}$ 的 λ 倍，也就同一线性空间中两个坐标系，有一个向量的方向一致，只有缩放发生；
- 矩阵 $\mathbf{A}$ 的特征值就是矩阵 $\mathbf{A}$ 所对应的一元多次方程组的根；也是多项式 $|\mathbf{A} - \lambda\mathbf{I}| = 0$ 的根；
- 特征向量是线性变换 $\mathbf{A}$ 作用在空间中的点对象时，作用效果不发生“旋转”，只有缩放的那些向量；特征值是每个特征向量在线性变换时的缩放比例，$\lambda > 1$ 放大向量，反之缩小；$\lambda = 1$ 没有发生缩放。

特征值与特征向量的几何意义



$$M = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3x \\ y \end{bmatrix}$$



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

- 矩阵乘向量是对向量的伸缩和旋转变换；
- 对角矩阵是对向量在坐标轴方向（特征向量方向）上进行的拉伸（特征值大于 1），缩（特征值小于 1），无旋转；如左图; 右图是沿 45 度方向的一个拉伸；
- 表示线性变换的矩阵可能改变某些方向（旋转）及其伸缩，而不被变换矩阵改变的那些方向（尽管有可能改变了伸缩比），称为特征向量，伸缩比为特征值。
- 线性变换 $\mathbf{A}$ 有些特征：不变的方向和这些方向上的缩放比例，二者相互决定对方(满秩时)
- 特征分解的作用：近似，降维，用主要方向（特征值大）来表示变换的主要作用

谱的定义：

- 有限维向量空间上一个变换的谱是其所有特征值的集合；简单说，“谱”就是矩阵的所有不同特征值的集合。

谱定理/谱分解/特征分解

- 谱定理：一个线性变换可以用特征向量的线性组合表示，系数为特征值，代表特征向量方向上的能量或对应特征向量方向的重要性。
- 谱定理换句话说就是，可对角化的方阵 $\mathbf{A} = \mathbf{P}^{-1}\Lambda\mathbf{P}$，其中 Λ 是特征值构成的对角阵，相似变换矩阵 $\mathbf{P}$ 的第 i 列就是特征值 λ_i 对应的特征向量；
- 任何一个线性变换可以用若干个方向上的缩放来组合而成；在每个特征向量 $\mathbf{x}$ 的方向上缩放对应特征值 λ 倍，最后实现线性变换 $\mathbf{A}$

特征值与特征向量的性质

- 线性变换的特征向量是指在变换下不变或者简单地乘以一个缩放因子的非零向量；特征向量的特征值是它所乘的那个缩放因子；
- 方阵所有特征值之和等于方阵的迹；
- 方阵 $\mathbf{A}$ 的一个特征值为 λ，对应的特征向量为 $\mathbf{x}$，则有意义的方阵 $\mathbf{A}^k, k$ 为整数，其特征值是 λ^k，对应的特征向量 $\mathbf{x}$ 不变；
- 线性变换的主特征向量是对应特征值最大的特征向量；
- 特征值的几何重次是相应特征向量构成的空间的维数；
- 方阵可逆等价于方阵的特征值中没有 0；
- 方阵的特征值与其转置的特征值一样。
- 每个矩阵元素的绝对值列和小于等于 1 时，矩阵的特征值都小于等于 1（例如概率转移矩阵)；对矩阵元素的绝对值行和也有类似结论。

实对称矩阵的特征值与特征向量

- 实对称矩阵的特征值均为实数；
- 实对称矩阵的不同特征值对应的特征向量两两正交；可以一般化为：对称矩阵不同特征值对应的特征向量正交；
- 实对称矩阵一个特征值的重数（代数重数）与其对应的线性无关的特征向量的个数（几何重数）相等；
- 实对称矩阵必正交相似与一个对角矩阵。正交相似是指相似变换矩阵是正交的。
- 注意：可对角化的充要条件是方阵的所有特征向量线性无关。不一定是要求实对称（充分条件，非必要）
- 可对角化的充要条件是正规矩阵（定义是存在酉矩阵把它酉相似成对角阵)

行列式的定义

- 实数方阵的行列式，其值为实数；
- 当行列式执行行列变换时，若获得一个三角阵，其值为对角线的乘积；因此，这就是一种计算行列式的方法；
- 呼唤两行或两列，行列式值反号；
- 行列式转置，值不变；行列式有等比例的行或列，值为 0（因为行列变换可以造成全零的行或列）；
- 可逆矩阵的行列式不为 0，因为：$|\mathbf{A}\mathbf{A}^{-1}| = |\mathbf{I}| = 1$

行列式的意义

- 行向量或列向量构成多面体的棱；
- 行列式就该多面体的体积，有向体积，即有正有负；
- 方阵的逆矩阵的行列式为方阵行列式的倒数；
- 分块对角行列式的值为各块子行列式的乘积；

矩阵分解：特征值分解

矩阵分解

- 将一个矩阵表示成两个或多个矩阵的乘积
- 类似于整数的因子分解
- 将矩阵视为线性变换，乘积矩阵代表连续若干个线性变换

特征分解

- 特征值 λ 和特征向量 $\mathbf{x}$：$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$
- 特征值分解：$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$，其中 $\mathbf{Q}$ 是特征向量构成的矩阵。$\mathbf{Q}$ 是正交阵
- 要求条件：满秩的方阵（实对称？正定，特征值大于 0）

矩阵分解：三角分解/LU 分解

三角分解法：

- 将方阵分解为上三角阵和一个下三角阵的乘积：$\mathbf{A} = \mathbf{LU}$
- 用途：简化行列式计算，求矩阵的逆和求解线性方程组等
- 分解不唯一

矩阵分解：QR 分解

QR 分解/正交三角分解：

- $\mathbf{A} = \mathbf{QR}$
- 其中 $\mathbf{Q}$ 是酉矩阵或正交矩阵，$\mathbf{R}$ 是上三角阵
- 分解不唯一，但要求 $\mathbf{R}$ 的对角线元素都为正时，分解唯一
- 用途：求矩阵特征值

矩阵分解：Cholesky 分解

Cholesky 分解：

- 实对称矩阵 $\mathbf{A}$ 是正定的，那么存在唯一分解：$\mathbf{A} = \mathbf{L}\mathbf{L}^T$
- 其中 $\mathbf{L}$ 是下三角矩阵，对角线元素为正
- 当 $\mathbf{A}$ 为复数域上的矩阵时，结论对厄米特矩阵也成立

矩阵分解：满秩分解

满秩分解：

- 若 $\mathbf{A}_{m\times n} \in \mathcal{C}_r^{m\times n}$ 为任意矩阵，秩为 r，则存在 $\mathbf{B}_{m\times r} \in \mathcal{C}_r^{m\times r}, \mathbf{B}_{m\times r} \in \mathcal{C}_r^{r\times n}$，使得 $\mathbf{A} = \mathbf{BC}$
- 其中 $\mathbf{B}, \mathbf{C}$ 分别是列秩和行秩为 r 的矩阵
- 分解不唯一

奇异值分解 (仅讨论实数向量的情形)：

- 奇异值分解：$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$
- $\mathbf{U}$ 是 m 阶的正交方阵，其列被称为左奇异向量，$\mathbf{V}$ 是 n 阶正交方阵，其列被称为右奇异向量
- Σ 是半正定的 $m \times n$ 阶对角阵
- 奇异值分解不唯一，是“对称矩阵正交相似于对角矩阵”的推广，“任意矩阵正交相似与（不全）的对角矩阵”

从特征分解到奇异值分解

- 实对称方阵 $\mathbf{A}$ 都有特征分解：$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \Lambda$，其中 $\mathbf{P}$ 是正交阵（转置就是逆）；
- 对于实的非对称方阵 $\mathbf{A}$，存在两个正交矩阵 $\mathbf{P}, \mathbf{Q}$ 满足 $\mathbf{P}^T\mathbf{A}\mathbf{Q} = \Sigma$

Proof.

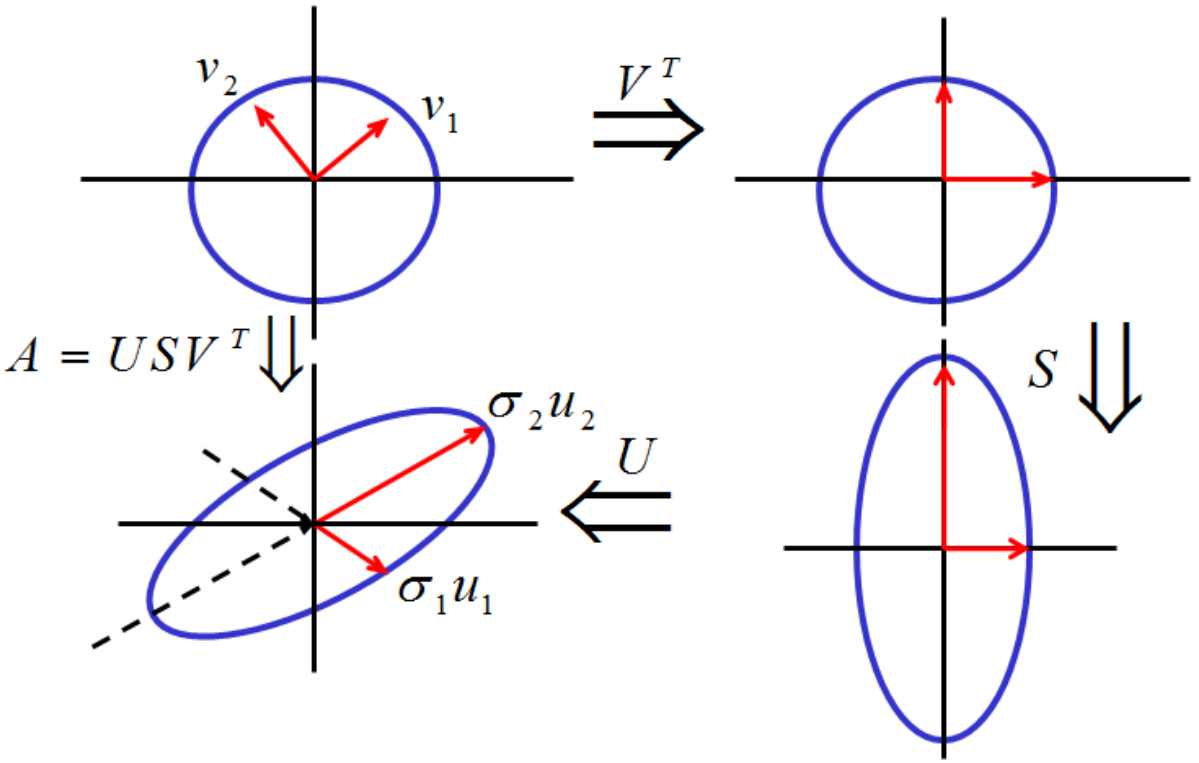
$\because \mathbf{A}$ 非奇异，故 $\mathbf{A}^T\mathbf{A}$ 实对称、正定矩阵；
即存在正交矩阵 $\mathbf{Q}$ 满足 $\mathbf{Q}^T(\mathbf{A}^T\mathbf{A})\mathbf{Q} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$
其中 $\lambda_i > 0$ 是 $\mathbf{A}^T\mathbf{A}$ 的特征值，令 $\alpha_i = \sqrt{\lambda_i}, \Sigma = diag(\alpha_1, \alpha_2, \ldots, \alpha_n)$
则有 $\mathbf{Q}^T(\mathbf{A}^T\mathbf{A})\mathbf{Q} = \Sigma^2$，改写为 $(\mathbf{A}\mathbf{Q}\Sigma^{-1})^T\mathbf{A}\mathbf{Q} = \Sigma$
令 $\mathbf{P} = \mathbf{A}\mathbf{Q}\Sigma^{-1}$ 得到 $\mathbf{P}^T\mathbf{A}\mathbf{Q} = \Sigma$，且有 $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ □

- 若 $\mathbf{A}_{m\times n} \in \mathcal{R}_r^{m\times n}$ 为任意秩为 r 的矩阵，则 $\mathbf{A}^T\mathbf{A}$ 是实对称、半正定的，不妨设 $\mathbf{A}^T\mathbf{A}$ 的特征值 $\mathbf{r} = (\lambda_1, \lambda_2, \ldots, \lambda_r, 0, 0, \ldots, 0)$，且满足 $\lambda_1 \geq lambda_2 \geq \ldots \geq \lambda_r > 0$。我们称 $\sigma_1 = \sqrt{\lambda_1}, \ldots, \sigma_r = \sqrt{\lambda_r}$ 为 $\mathbf{A}$ 的奇异值，记 $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$
- 存在 m 阶和 n 阶正交方阵 $\mathbf{U}, \mathbf{V}$，使得秩为 r 的矩阵 $\mathbf{A}_{m\times n} \in \mathcal{R}_r^{m\times n}$，满足 $\mathbf{U}^T\mathbf{A}\mathbf{V} = diag(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, 0, \ldots, 0)$，此为奇异值分解

理解奇异值分解

- 奇异值分解是实对称矩阵正交相似与对角阵的推广，任意矩阵都可以用两个正交矩阵分别左乘和右乘一个对角阵
- 若要分解的矩阵是 $\mathbf{A}$，其第一个右乘的正交矩阵的列向量其实是 $\mathbf{A}\mathbf{A}^T$ 的特征向量，而第三个左乘的正交向量的列向量其实是 $\mathbf{A}^T\mathbf{A}$ 的特征向量
- $\mathbf{A}\mathbf{A}^T$ 和 $\mathbf{A}^T\mathbf{A}$ 的非零特征值完全相同
- 矩阵的奇异值分解不唯一

V是正交矩阵，表示二维空间的一个旋转



S 维空间坐标平面上的椭圆将平面上的圆变换到三

U是正交矩阵，表示三维空间的一个旋转

奇异值分解的用途

- 奇异值分解可以降维：将原数据矩阵 $\mathbf{A}$ 用 $m \times r, n \times r$ 的两个矩阵和一个对角阵来替代存储；当 $r < \frac{mn}{m+n+1}$ 时，实现数据压缩存储；
- 奇异值对矩阵扰动不敏感：奇异值的变化之和不超过矩阵的变化之和；
- 奇异值比例不变性；旋转不变性（左乘一个正交矩阵表示旋转），即旋转前后，矩阵的奇异值不变；
- 可以用于计算矩阵的一个逼近，找一个低秩的矩阵 $(k \leq r)$ 来逼近原矩阵；用于除噪声，提取主要信息等；
- 矩阵的秩 k 逼近：设奇异值 $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$，对应的特征向量为 $\mathbf{u}_1\mathbf{v}_1^t, \mathbf{u}_2\mathbf{v}_2^t, \ldots, \mathbf{u}_r\mathbf{v}_r^t$，故有 $\mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^t + \sigma_2\mathbf{u}_2\mathbf{v}_2^t + \ldots + \sigma_r\mathbf{u}_r\mathbf{v}_r^t$，舍去权值小的项，剩下 k 项仍能很好地近似原矩阵。

数据压缩表示

- 大数据时代，数据矩阵 $\mathbf{A}_{n\times m}$，m 是数据个个数，n 是数据的维度，数据维度 n 和量 m 都非常大，存储和处理需要压缩 $\mathbf{A}$
- 数据压缩的作用：去噪声，选出主要或有用的特征和数据
- 不妨设 $\mathbf{A}$ 是原始数据矩阵经过非线性变换提取到的特征数据矩阵，接下来的机器学习过程就是寻找最优“线性变换”的过程

主要技术分类

- 列降维：即选择数据子集。方法包括：随机 (均匀) (有回放/无回放) 采样，代表性数据点 (Landmarks) 选择，非原始数据集中代表性数据点生成等。
- 行降维：即选择特征子集或生成新的特征及其集合。方法包括：主成分分析/PCA，流形学习/Manifold Learning，
- 行列同时降维：

协方差矩阵与核矩阵

已知条件：数据矩阵 $\mathbf{A}$

- 计算数据矩阵因内积运算带来的两个矩阵

协方差矩阵

- 任何两行之间做内积，得到实对称矩阵
- 描述任意两个特征之间的相关性

核矩阵

- 任何两列之间做内积，得到实对称矩阵
- 描述任意两个数据之间的相似性

问题描述

- 已知：数据矩阵 $\mathbf{X}_{n\times m}$，m 个 n 维的数据；（假设数据已经中心化，即减去了期望）
- 正交矩阵 $\mathbf{V}_{n\times q}, q < n$ 线性变换/降维变换，且令 $\mathbf{P}_{n\times n} = \mathbf{V}\mathbf{V}^T$，因为 $\mathbf{V}$ 是正交阵，故有 $\mathbf{V}^T\mathbf{V} = \mathbf{I}$
- 求：$arg\min_{\mathbf{V}} \sum_i^m ||\mathbf{x}_i - \mathbf{V}\mathbf{V}^T\mathbf{x}_i||^2 = ||\mathbf{X} - \mathbf{V}\mathbf{V}^T\mathbf{X}||^2$

理解一

- $\mathbf{V}^T\mathbf{X}$ 表示将 n 维 m 个数据 $\mathbf{X}$ 变换到 q 维空间 $\mathbf{V}^T$ 中，$\mathbf{V}^T$ 每一列是长度为 q 的基向量
- $\mathbf{V}$ 为正交矩阵，所以 $\mathbf{V}^T = \mathbf{V}^{-1}$，故，$\mathbf{V}(\mathbf{V}^T\mathbf{X})$ 将 q 维空间中的 m 个数据逆变换回原 n 维空间
- 上述优化问题，就是求线性变换 $\mathbf{V}$ 把数据变化到一个低维空间后(可能损失了信息)，再恢复到原空间，若能完全恢复（尽可能恢复)，则最优。(对比自编码)

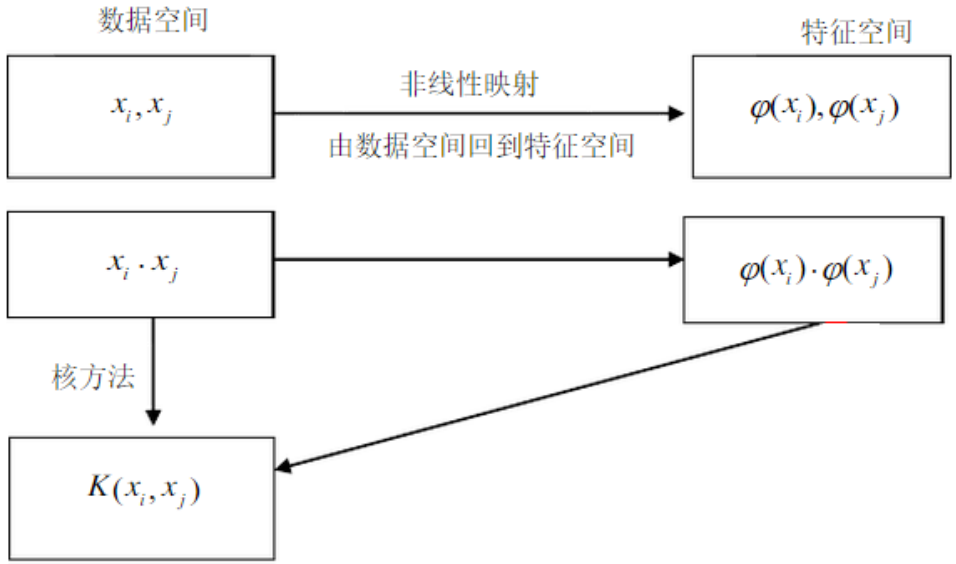
理解二

- 假设降维后数据为 $\mathbf{Y}_{q\times m}=\{\mathbf{y}_i, i=1,2,\ldots,m\}$，矩阵 $\mathbf{V}_{n\times q}$ 将降维后数据重建为原始数据，即我们希望最小化 $||\mathbf{X}-\mathbf{VY}||$,
- 求导 $\frac{d}{d\mathbf{y}_i}=0$，即令导数为 0 来求极值，可得 $\mathbf{y}_i=\mathbf{V}^T\mathbf{x}_i$，即得到最小化问题 $||\mathbf{X}-\mathbf{VV}^T\mathbf{X}||$

求解过程

- 将最小化 $\sum_i^m ||\mathbf{x}_i - \mathbf{V}\mathbf{V}^T\mathbf{x}_i||^2$ 等价于最小化：
  $\sum_i^m ||\mathbf{x}_i||^2 - \sum_i^m ||\mathbf{V}\mathbf{V}^T\mathbf{x}_i||^2$(原问题用内积表示 2-范数展开即得)，
- 原问题等价于最大化 $\sum_i^m ||\mathbf{V}\mathbf{V}^T\mathbf{x}_i||^2 = \sum_i^m ||\mathbf{P}\mathbf{x}_i||^2 = \sum_i^m \mathbf{x}_i^T\mathbf{P}\mathbf{x}_i = tr(\mathbf{X}^T\mathbf{P}\mathbf{X}) = tr(\mathbf{P}\mathbf{X}\mathbf{X}^T) = tr(\mathbf{P}\mathbf{S})$，其中 $\mathbf{S}$ 是数据的协方差矩阵（利用迹的性质 $tr(ABC) = tr(BCA) = tr(CAB)$）
- 最大化 $tr(\mathbf{P}\mathbf{S}) = tr(\mathbf{V}\mathbf{V}^T\mathbf{S}) = tr(\mathbf{V}^T\mathbf{S}\mathbf{V}) = \sum_i^q \mathbf{v}_i^T\mathbf{S}\mathbf{v}_i$，约束条件 $\mathbf{v}_i^T\mathbf{v}_i = 1$，利用拉格朗日乘子法优化
  $\mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_q, \alpha_i, \ldots, \alpha_q) = \sum_i^q \mathbf{v}_i\mathbf{S}\mathbf{v}_i - \alpha_i(\mathbf{v}_i^T\mathbf{v}_i - 1)$，得到
  $\mathbf{v}_i^T\mathbf{S} = \alpha_i\mathbf{v}_i^T$，即 $\mathbf{S}\mathbf{v}_i = \alpha_i\mathbf{v}_i$（因为 $\mathbf{S}$ 对称，等式两边取转置即得）

核方法



解释

- 核函数方法的基本原理是通过非线性函数把输入空间映射到高维空间，在特征空间中进行数据处理，其关键在于通过引入核函数，把非线性变换后的特征空间内积运算转换为原始空间的核函数计算，从而大大简化了计算量

数学描述

- 特征提取函数，包括降维、升维和维度不变，记为 $\mathbf{x} \longrightarrow \phi(\mathbf{x})$
- 两个数据对象 $\mathbf{x}, \mathbf{y}$，其特征向量为 $\phi(\mathbf{x}), \phi(\mathbf{y})$，特征向量的内积为 $\phi(\mathbf{x}) \cdot \phi(\mathbf{y})$
- 令 $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{y})$，核函数

从训练数据矩阵获得线性分类器

线性分类器

- 给定训练数据集 $\{\mathbf{D}_{m\times n}, \mathbf{y}_{m\times 1}\}$
- 线性分类器 $f(x) = \mathbf{W} \cdot \mathbf{x}$，寻找 $\mathbf{W}$ 使得 $\mathbf{y} = \mathbf{WD}$，解线性方程组，对于超定、欠定和其它无解情形时，采用最小二乘法来近似/拟合
- 优化错误率：$arg\min_{\mathbf{w}} \Sigma_i (y_i - \mathbf{W} \cdot \mathbf{d}_i)^2$