

1. 数学基础

- ①线性代数 ②概率统计 ③优化理论 ④图论

数学基础

零、高中数学

1.高中数学

圆: $(x_1 - c_1)^2 + (x_2 - c_2)^2 = r^2$, $(\mathbf{x} - \mathbf{c})^T (\mathbf{x} - \mathbf{c}) = r^2$, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

椭圆: $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1$, $\mathbf{x}^T \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \frac{1}{b^2} \end{bmatrix} \mathbf{x} = 1$, 逆时针旋转 θ 得

$$\frac{[x_1 \cos \theta + x_2 \sin \theta]^2}{a^2} + \frac{[x_1 \sin \theta - x_2 \cos \theta]^2}{b^2} = 1$$

算术平均数-几何平均数不等式(AM-GM Inequality): $A_n = \frac{a_1 + \dots + a_n}{n} \geq \sqrt[n]{a_1 \dots a_n} = G_n$

一、线性代数

1-0.引入

今有雉兔同笼，上有三十五头，下有九十四足，问雉兔各几何？

Q1: 方程求解 $\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 35 \\ 94 \end{bmatrix}$ 。

A1: $x = A^{-1}b$

Q2: 向量含义？

A2: $A = \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{matrix} H \\ F \\ C \\ R \end{matrix}$ (C,R:鸡,兔, H,F:头,脚), $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 是鸡兔数量, $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ 是头脚数量。

$Ax = b$ 实质是鸡兔坐标系 x 向头脚坐标系 b 的转换 $x \rightarrow b$, 反过来看, $x = A^{-1}b$ 是头脚坐标系 b 向鸡兔坐标系 x 的转换 $b \rightarrow x$ 。(注: 可视化时为了展示, 取 $b = [3 \ 8]^T$)

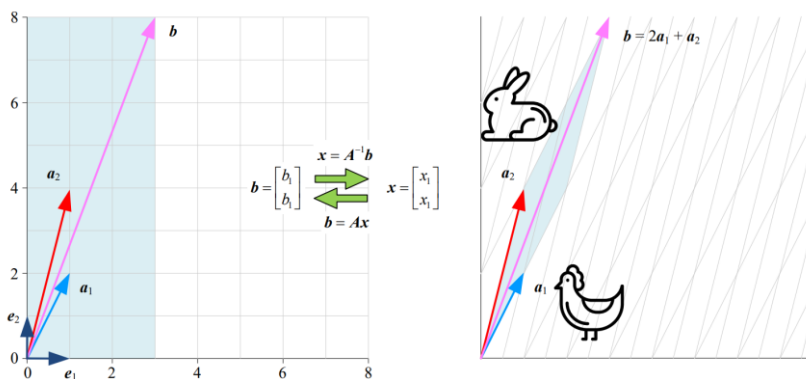


图 5. “头-脚系”和“鸡-兔系”相互转换

Q3: 线性组合。套餐捆绑销售: A 套餐 3 鸡 1 兔; B 套餐 1 鸡 2 兔。想买 10 鸡 10 兔。

A3: $x_1 w_1 + x_2 w_2 = x_1 \begin{bmatrix} 3 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$, 即 $\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$ 。 $x_1 w_1 + x_2 w_2$ 是线性组合, 将

基底 $\{w_1, w_2\}$ 混合得到 $a = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$ 。是鸡兔坐标系 x 向 AB 套餐坐标系 w 的转换 $x \rightarrow w$ 。

注: x_1, x_2, \dots 的所有线性组合称为 x_1, x_2, \dots 的张成(span), 记作 $\text{span}(x_1, x_2, \dots)$ 。

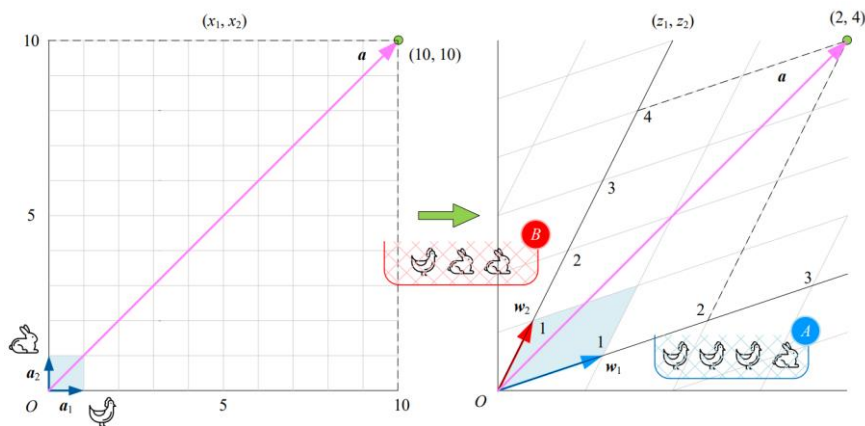


图 9. 坐标系转换, “鸡-兔系”到“A-B套餐系”

无论是 $b \rightarrow x$ 还是 $x \rightarrow w$ ，都叫做**基底变换**。对同一个向量 a ，在基底 $\{a_1, a_2\}$ 和 $\{w_1, w_2\}$

$$\text{下分别是: } a = \begin{cases} x = Ix = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 a_1 + x_2 a_2 \quad (I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}) \\ Wz = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1 w_1 + z_2 w_2 \end{cases}, \text{ 因此 } x = Wz, \text{ 新坐标系 } z \text{ 通}$$

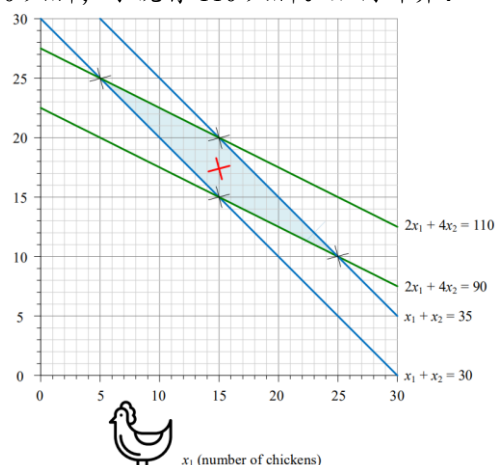
过 $z = W^{-1}x$ 转换网格形状(平面还是那个平面)。

如果新推出套餐 C,D, 记为 V , 那么 $a = Vs$, 得 $Wz = Vs$, 即 $s = V^{-1}Wz$, 也就是向量 a 从基底 $\{w_1, w_2\}$ 到 $\{v_1, v_2\}$ 。

Q4: 甲说有 30 个头, 乙说有 35 个头; 丙说有 90 只脚, 丁说有 110 只脚。如何计算?

A4: 这是**超定方程组**(overdetermined system):

$$\text{也就是 } \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 4 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 30 \\ 35 \\ 90 \\ 110 \end{bmatrix}, \text{ 对于此方程, } \begin{matrix} \text{rabbit} \\ x_2 \text{ (number of rabbits)} \end{matrix}$$



因为 A 不是方阵, 明显不可逆,

因此转化成 $A^T A x = A^T b$,

得: $x = (A^T A)^{-1} A^T b = \begin{bmatrix} 15 \\ 17.5 \end{bmatrix}$, 恰好是

平行四边形的中心位置(右图是线性规划图)。

$(A^T A)^{-1} A^T$ 常被称作广义逆(generalized inverse), 伪逆(pseudoinverse)。如果 $A^T A$ 非满秩则不可逆, 这时需要摩尔-彭若斯广义逆(Moore-Penrose inverse), 用 `numpy.linalg.pinv()`。

Q5: 线性回归中使用超定方程组。

A5: 线性回归也是一个 $Ax = b$ 问题($b = x_1 a_1 + \dots + x_D a_D$, 此线性组合中 x_i 是比例, a_i 是列向量)。线性回归中常用的是 $y = \theta^T X$, 类比可得 $\theta = (X^T X)^{-1} X^T y$ 。

下面以一元一次函数为例:

$y = ax + b$ 化成线性代数的表示方法即 $y = ax + bI = \begin{bmatrix} I & x \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix}$, 令 $X = \begin{bmatrix} I & x \end{bmatrix}$, 方程化为:

$$y = X \begin{bmatrix} b \\ a \end{bmatrix}, \text{ 解得 } \begin{bmatrix} b \\ a \end{bmatrix} = (X^T X)^{-1} X^T y.$$

误差 $\varepsilon = y - \hat{y} = y - (ax + bI)$, 显然 $\varepsilon \perp x, \varepsilon \perp I$,

$$\text{因此 } \begin{cases} I^T \varepsilon = 0 \\ x^T \varepsilon = 0 \end{cases} \Rightarrow \begin{cases} I^T (y - (ax + bI)) = 0 \\ x^T (y - (ax + bI)) = 0 \end{cases},$$

$$\text{也就是 } \begin{bmatrix} I & x \end{bmatrix}^T \left[y - \begin{bmatrix} I & x \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \right] = 0, \text{ 即}$$

$$X^T \left(y - X \begin{bmatrix} b \\ a \end{bmatrix} \right) = 0, \text{ 即 } X^T y = X^T X \begin{bmatrix} b \\ a \end{bmatrix},$$

$$\text{因此也有 } \begin{bmatrix} b \\ a \end{bmatrix} = (X^T X)^{-1} X^T y.$$

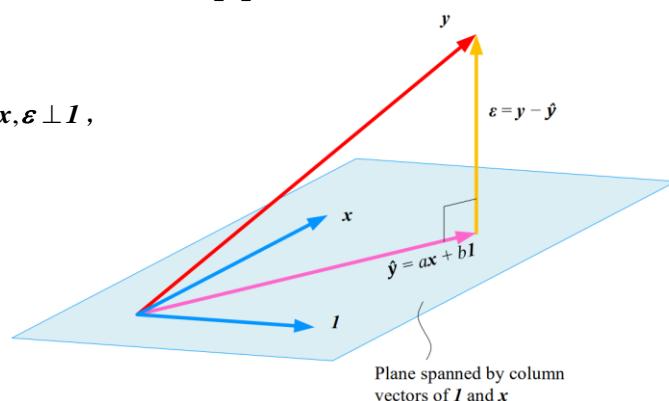


图 12. 几何角度解释一元最小二乘结果, 二维平面

此外, 还可以利用 $y = \rho_{x,y} \frac{\sigma_y}{\sigma_x} (x - \mu_x) + \mu_y = \rho_{x,y} \frac{\sigma_y}{\sigma_x} x + \left(-\rho_{x,y} \frac{\sigma_y}{\sigma_x} \mu_x + \mu_y \right)$ 计算 a, b 的值。

Q6: 每晚有 30% 的小鸡变成小兔，其他小鸡不变；与此同时，每晚有 20% 小兔变成小鸡，其余小兔不变。分析此 **马尔可夫过程** (Markov process)。

A6: 记第 k 天鸡兔的比例是 $\pi(k) = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}$ ，代表鸡兔分别占比 π_1, π_2 。变化过程是：

$k \rightarrow k+1: T\pi(k) = \pi(k+1)$ ，式中 **转移矩阵** (transition matrix) $T = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}$ 相当于 $\begin{bmatrix} p & q \\ 1-p & 1-q \end{bmatrix}$ ， $\pi(k)$ 叫 **状态向量** (state vector)。横向看 T ， $[0.7 \quad 0.2]$ 对应着第 k 的鸡兔怎么转换成第 $k+1$ 天的鸡。纵向看 T ， $\begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix}$ 对应着第 k 天的鸡将要怎么变化成第 $k+1$ 天的对应鸡兔。

$$\text{求解平衡状态: } \begin{cases} \begin{bmatrix} p & q \\ 1-p & 1-q \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} \Rightarrow (1-p)\pi_1 + q\pi_2 = 0 \\ \pi_1 + \pi_2 = 1 \end{cases} \Rightarrow \begin{cases} \pi_1 = \frac{q}{1-p+q} \\ \pi_2 = \frac{1-p}{1-p+q} \end{cases}$$

请注意 $T\pi = \pi$ 是特征值为 1 的 **特征值分解** (eigen decomposition)，所以还可以用特征值求解稳态，也就是特征值 $\lambda_1 = 1, \lambda_2 = \frac{1}{2}$ ，特征向量 $v_1 = \begin{bmatrix} -0.5547 \\ -0.8321 \end{bmatrix}, v_2 = \begin{bmatrix} -0.707 \\ 0.707 \end{bmatrix}$ ，注意 $\pi_1 + \pi_2 = 1$ ，

因此 v_1 才是符合题意的稳态，即 $\lim_{n \rightarrow \infty} \pi_n = \frac{1}{-0.5547-0.8321} \begin{bmatrix} -0.5547 \\ -0.8321 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$ 。

补充: 斐波拉契数列是二阶马尔可夫链。可以写成 $\begin{bmatrix} F_{k+1} \\ F_{k+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} F_k \\ F_{k+1} \end{bmatrix}$ ，转移矩阵 T 的

作用是剪切+镜像 $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ ，即：水平剪切角为 $\theta_S = \frac{\pi}{4}$ ，水平镜像角 $\theta_R = \frac{\pi}{4}$ 。易

得 T 的特征值是 $\lambda_1 = \frac{1-\sqrt{5}}{2}, \lambda_2 = \frac{1+\sqrt{5}}{2}$ 特征向量是 $v_1 = \begin{bmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_1 \end{bmatrix}, v_2 = \begin{bmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_2 \end{bmatrix}$ 。如

果对 T 进行特征值分解： $T = Q\Lambda Q^{-1} = \begin{bmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 & -1 \\ \lambda_1 & 1 \end{bmatrix}$ 。根据特征值分解的

结论，有 $T^n = Q\Lambda^n Q^{-1}$ ，因此 $\pi^n = T^n \pi_0 = Q\Lambda^n Q^{-1} \pi_0 = \begin{bmatrix} 1 & 1 \\ \lambda_1 & \lambda_2 \end{bmatrix} \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix} \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 & -1 \\ \lambda_1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$$= \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2^k - \lambda_1^k \\ \lambda_2^{k+1} - \lambda_1^{k+1} \end{bmatrix}, \text{ 即 } F_n = \frac{1}{\lambda_2 - \lambda_1} \lambda_2^k - \lambda_1^k = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k}{\sqrt{5}}。$$

1-1. 向量

① 如无说明，默认列向量。

① 内积/标量积/点积 $\mathbf{a} \cdot \mathbf{b} = \langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta = \sum_{i=1}^n a_i b_i$ (θ 即 $\angle \mathbf{a}, \mathbf{b}$)

内积是 \mathbf{a} 在 \mathbf{b} 方向上的投影与 \mathbf{b} 的乘积。

$\mathbf{a} \cdot \mathbf{b} = 0$ 时称 \mathbf{a} 和 \mathbf{b} 正交。

向量投影 $\text{Proj}_{\mathbf{a}} \mathbf{b} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|} = \mathbf{b} \cos \theta$ 。

由于 $|\cos \theta| \leq 1$ ，可得柯西-施瓦茨不等式 $(\mathbf{a} \cdot \mathbf{b})^2 \leq \|\mathbf{a}\|^2 \|\mathbf{b}\|^2$ ，即 $|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\| \|\mathbf{b}\|$ 。在 \mathbb{R}^n 空间中，上式等价于 $\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right)$ 。

类比 $\cos \theta$ 的计算公式，定义余弦相似度 $k(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \in [-1, 1]$ ，

定义余弦距离 $d(\mathbf{x}, \mathbf{y}) = 1 - k(\mathbf{x}, \mathbf{y}) \in [0, 2]$ 值越小越正相似。

`numpy.dot()` 计算向量内积/矩阵乘积(相当于矩阵运算符@)

② 外积/向量积/叉乘 $\mathbf{a} \times \mathbf{b}$ 与 \mathbf{a}, \mathbf{b} 都垂直

$$\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\| \|\mathbf{b}\| \sin \theta = \begin{vmatrix} i & j & k \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix} = S_{\square} = 2S_{\triangle} \quad (\vec{a}, \vec{b} \text{ 构成的 } \square \text{ 和 } \triangle)$$

外积得到的是 \mathbf{a} 和 \mathbf{b} 所在平面的法线向量，模是平行四边形的面积。

③ 哈达玛积 Hadamard product / 逐项积 piecewise product / 元素乘积

对于向量 $\mathbf{a} = [a_1 \ \cdots \ a_n]^T$, $\mathbf{b} = [b_1 \ \cdots \ b_n]^T$, $\mathbf{a} \odot \mathbf{b} = [a_1 b_1 \ \cdots \ a_n b_n]^T$

对于同型矩阵 $\mathbf{A} = (a_{ij})_{m \times n}$, $\mathbf{B} = (b_{ij})_{m \times n}$, 有 $\mathbf{C} = \mathbf{A} \odot \mathbf{B} = (c_{ij})_{m \times n}$, 其中 $c_{ij} = a_{ij} \times b_{ij}$ 。内积是“向量→标量”，而哈达玛积是“向量→向量”，“矩阵→矩阵”。

$$\text{如 } \mathbf{A} = \begin{pmatrix} 1 & 3 & -2 \\ 4 & 2 & 5 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} -2 & 7 & 0 \\ 4 & -3 & 1 \end{pmatrix}, \mathbf{A} \odot \mathbf{B} = \begin{pmatrix} -2 & 21 & 0 \\ 16 & -6 & 5 \end{pmatrix}$$

④ 克罗内克积 Kronecker product / 张量积 tensor product

对于向量 $a = [a_1 \ \dots \ a_n]^T, b = [b_1 \ \dots \ b_m]^T$, $a \otimes b = ab^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \dots & a_1 b_m \\ a_2 b_1 & a_2 b_2 & \dots & a_2 b_m \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \dots & a_n b_m \end{bmatrix}_{n \times m}$, 也可

以写成 $[b_1 a \ b_2 a \ \dots \ b_n a]$ 或 $\begin{bmatrix} a_1 b^T \\ a_2 b^T \\ \vdots \\ a_n b^T \end{bmatrix}$ 。

张量积 $a \otimes b$ 是 a, b 张起的一个网格面, 也就是(图右上部分红线)对于某个 a_i 值以 b 这个系数来放大, (图右下部分蓝线)对于某个 b_j 值以 a 这个系数来放大:

注: 离散随机变量独立条件下, 联合概率

$P_{X,Y}(x,y) = P_X(x) \cdot P_Y(y)$ 和这个图是类似的。

对于任意大小的矩阵 $A_{m \times n}, B_{p \times q}$, 有 $A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}$ 图 29. 从几何角度解释向量张量积

可以看做是 A 的每个元素对 B 进行缩放。

$$\text{如 } A = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, B = \begin{pmatrix} 2 & 0 \\ 4 & 1 \end{pmatrix}, A \otimes B = \begin{pmatrix} 1 \cdot 2 & 1 \cdot 0 & 2 \cdot 2 & 2 \cdot 0 \\ 1 \cdot 4 & 1 \cdot 1 & 2 \cdot 4 & 2 \cdot 1 \\ 3 \cdot 2 & 3 \cdot 0 & 1 \cdot 2 & 1 \cdot 0 \\ 3 \cdot 4 & 3 \cdot 1 & 1 \cdot 4 & 1 \cdot 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 4 & 0 \\ 4 & 1 & 8 & 2 \\ 6 & 0 & 2 & 0 \\ 12 & 3 & 4 & 1 \end{pmatrix}$$

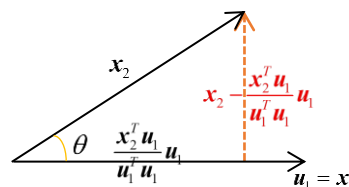
⑤ 标准正交基: 基向量相互正交且长度均为 1, 即: $u_i^T u_j = 0, i \neq j$ 且 $u_i^T u_i = 1$ 。

给定一组线性无关的向量 $x_1 \sim x_n$, 可用格拉姆-施密特正交化构造出标准正交基 $u_1 \sim u_n$ 。下面以 x_1, x_2 为例:

先令 $u_1 = x_1$, 再取 $u_2 = x_2 - \alpha_{21} u_1$, 要使 u_1, u_2 正交, 需

$$(x_2 - \alpha_{21} u_1)^T u_1 = 0 \Rightarrow \alpha_{21} = \frac{x_2^T u_1}{u_1^T u_1} = \frac{\|x_2\| \|u_1\| \cos \theta}{\|u_1\|^2} = \frac{\|x_2\| \cos \theta}{\|u_1\|}。 \text{因}$$

此 $\alpha_{21} u_1 = \|x_2\| \cos \theta \frac{u_1}{\|u_1\|}$ 就是 x_2 向 x_1 的投影。那么 $x_2 - \alpha_{21} u_1$ 自然是垂直于 x_1 的。



由此, $u_k = x_k - \sum_{i=1}^{k-1} a_{ki} u_i, a_{ki} = \frac{x_k^T u_i}{u_i^T u_i}$, 最后依次通过 $\frac{u_k}{\|u_k\|}$ 即可得到标准正交基。

简洁地说就是 $u_2 = u_1 - \text{proj}_{u_1}(x_2), \dots, u_n = u_{n-1} - \sum_{i=1}^{n-1} \text{proj}_{u_i}(x_n)$ 。只要保证 u_n 减去了在 $u_{1 \sim n-1}$ 上的投影, 那就会呈现垂直的结果。

⑥ 超平面 Hyperplane: $w^T x + b = 0$ 。显然 $w^T (x_1 - x_2) = 0$, 因此法向量 w 与平面内任意两点之间的连线 $x_1 x_2$ 正交。

中垂线: 点 μ_1, μ_2 的中垂线 x 满足 $(\mu_2 - \mu_1)^T x - \frac{1}{2}(\mu_2 - \mu_1)^T (\mu_2 + \mu_1) = 0$ 。这由 μ_1, μ_2 的

直线 $\mu_2 - \mu_1$ 与 x 和中点 $\frac{1}{2}(\mu_2 + \mu_1)$ 的直线垂直: $(\mu_2 - \mu_1)^T \left(x - \frac{1}{2}(\mu_2 + \mu_1) \right) = 0$ 可证。

注意在 K-Means 中取其中的 2 个特征与欧氏距离, 此时决策边界就是质心的中垂线。

向量距离: 点 q 到 $w^T x + b = 0$ 的距离是 $d = \frac{w^T x + b}{\|w\|}$ 。这是因为点 q 到超平面上任一点 x 的距离满足 $d = \|q - x\| \cos \langle q - x, w \rangle_\theta = \|q - x\| \frac{|w^T (q - x)|}{\|w\| \|q - x\|} = \frac{|w^T (q - x)|}{\|w\|} \stackrel{w^T x = -b}{=} \frac{w^T x + b}{\|w\|}$ 。

正交投影坐标: $x_q = q - \frac{w^T x + b}{\|w\|^2} w = q - \frac{(w^T x + b)}{w^T w} w$ 。

平行平面距离: 类似向量距离, 取两平面上的点 x_1, x_2 , $d_{\parallel} = \frac{|w^T (x_1 - x_2)|}{\|w\|} \stackrel{\substack{w^T x_1 = -b_1 \\ w^T x_2 = -b_2}}{=} \frac{|b_2 - b_1|}{\|w\|}$ 。

1-2. 矩阵

①将矩阵记为 $X_{n \times D}$ 以便与机器学习对应(n 个样本点, D 维)。

$$\text{记 } X_{n \times D} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_D] = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix}, \text{ 其中的 } \mathbf{x} \text{ 分别代表每个特征和每个样本。}$$

①张量 tensor 可以看作是一个多维数组。标量可以看作是 0 维张量, 向量可以看作 1 维张量, 矩阵可以看作是 2 维张量。

②各种矩阵

单位矩阵 I 是对角线元素全 1 其余全 0 的 $n \times n$ 方阵。

对角矩阵 若 $k_i \neq 0$, 则逆矩阵存在且 $\Lambda^{-1} = \text{diag}(k_1^{-1}, \dots, k_n^{-1})$ 。求幂: $\Lambda^N = \text{diag}(k_1^N, \dots, k_n^N)$ 。
 对角矩阵, 上/下三角矩阵的主对角线元素是特征值。对角矩阵的作用是缩放, 比如:

$$X_{n \times D} \Lambda_{D \times D} = [\lambda_1 \mathbf{x}_1 \quad \cdots \quad \lambda_D \mathbf{x}_D], \Lambda_{n \times n} X_{n \times D} = [\lambda_1 \mathbf{x}^{(1)} \quad \cdots \quad \lambda_D \mathbf{x}^{(n)}]_{n \times 1}^T$$

副对角矩阵可以将元素逆序, 即 $[x_1 \quad \cdots \quad x_n]_{1 \times D} \begin{bmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{bmatrix}_{D \times D} = [x_D \quad \cdots \quad x_1]_{1 \times D}$ 。一般地:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} a_{1,3} & a_{1,1} & a_{1,4} & a_{1,2} \\ a_{2,3} & a_{2,1} & a_{2,4} & a_{2,2} \\ a_{3,3} & a_{3,1} & a_{3,4} & a_{3,2} \\ a_{4,3} & a_{4,1} & a_{4,4} & a_{4,2} \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} a^{(1)}_3 & a^{(1)}_1 & a^{(1)}_4 & a^{(1)}_2 \\ a^{(2)}_3 & a^{(2)}_1 & a^{(2)}_4 & a^{(2)}_2 \\ a^{(3)}_3 & a^{(3)}_1 & a^{(3)}_4 & a^{(3)}_2 \\ a^{(4)}_3 & a^{(4)}_1 & a^{(4)}_4 & a^{(4)}_2 \end{bmatrix}$$

矩阵可逆(invertible)也称非奇异(non-singular)。矩阵可逆 \Leftrightarrow 满秩 $\Leftrightarrow |A| \neq 0$ 。

$$\text{分块矩阵} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} H & -HBD^{-1} \\ -D^{-1}CH & D^{-1} + D^{-1}CHBD^{-1} \end{bmatrix}, \text{ 式中 } H = (A - BD^{-1}C)^{-1}。$$

矩阵的秩(rank) $r(A)$ 是矩阵线性无关的行向量或列向量的最大数量, 满足:

$$r(A) \leq \min(m, n) \quad r(A) = r(A^T) = r(AA^T) = r(A^T A) \quad r(A+B) \leq r(A) + r(B) \quad r(AB) \leq \min(r(A), r(B))$$

注 1: 因为矩阵的行秩等于列秩, 因此定义中使用的是“或”。

注 2: 矩阵的秩 $r(A)$ 的几何意义是: 矩阵的列/行向量组能张成 $r(A)$ 维空间。

伴随矩阵 A^* 的每个元素 A_{ij}^* 是原矩阵 A 对应元素 a_{ij} 的代数余子式。 $AA^* = \text{diag}(|A|, \dots, |A|)$,

因此 $AA^* = |A|I$ 。若 $|A| \neq 0$, 则 $I = \frac{AA^*}{|A|}$, 即 $A^{-1} = \frac{A^*}{|A|}$ 。

$$\text{格拉姆 Gram 矩阵: 一个向量组 } \mathbf{a}_1, \dots, \mathbf{a}_D \text{ 的格拉姆矩阵是 } G = A^T A = \begin{bmatrix} \mathbf{a}_1^T \mathbf{a}_1 & \cdots & \mathbf{a}_1^T \mathbf{a}_D \\ \vdots & \ddots & \vdots \\ \mathbf{a}_D^T \mathbf{a}_1 & \cdots & \mathbf{a}_D^T \mathbf{a}_D \end{bmatrix}_{D \times D}。$$

性质: 由于 $\mathbf{x}_i^T \mathbf{x}_j = \mathbf{x}_j^T \mathbf{x}_i$, 因此格拉姆矩阵是一个对称矩阵。因为 Gram 矩阵度量的是数据的“长度/距离”(长度指 $\mathbf{a}_i^T \mathbf{a}_i$, 距离指 $\mathbf{a}_i^T \mathbf{a}_j$), 而且是对称矩阵, 因此重要。

$$\text{trace}(A^T A) = \mathbf{a}_1 \cdot \mathbf{a}_1 + \cdots + \mathbf{a}_D \cdot \mathbf{a}_D = \sum_{i=1}^n x_{i,1}^2 + \cdots + \sum_{i=1}^n x_{i,D}^2 = \sum_{j=1}^D \sum_{i=1}^n x_{i,j}^2, \text{ 也就是 } A \text{ 所有元素的平方和。}$$

正交矩阵: 若 $A^T = A^{-1}$, 即 $A^T A = AA^T = I$, 相当于 $\mathbf{a}_i^T \mathbf{a}_i = 1, \mathbf{a}_i^T \mathbf{a}_j = 0 (i \neq j)$, 称方阵 A 是正交矩阵 orthogonal matrix。请注意 $A^T A$ 就是 Gram 矩阵。

性质: 正交矩阵的行(列)向量均为单位向量且相互正交。正交矩阵的乘积/逆矩阵/转置矩阵还是正交矩阵。正交矩阵的行列式为 ± 1 。正交矩阵的几何操作对应“旋转”、“镜像”、

“置换”，或者它们的组合。

两种理解方法（其实就是矩阵相乘的两种理解）：

$$\textcircled{1} A^T A = \begin{bmatrix} a_1^T \\ \vdots \\ a_D^T \end{bmatrix} \begin{bmatrix} a_1 & \cdots & a_D \end{bmatrix} = \begin{bmatrix} a_1^T a_1 & \cdots & a_1^T a_D \\ \vdots & \ddots & \vdots \\ a_D^T a_1 & \cdots & a_D^T a_D \end{bmatrix}_{D \times D} = I_{D \times D}$$

$$\textcircled{2} A A^T = \begin{bmatrix} a_1 & \cdots & a_D \end{bmatrix} \begin{bmatrix} a_1^T \\ \vdots \\ a_D^T \end{bmatrix} = a_1 a_1^T + \cdots + a_D a_D^T = I_{D \times D}$$

③公式

转置和可逆 $(A+B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, $(AB)^{-1} = B^{-1} A^{-1}$, $(A^T)^{-1} = (A^{-1})^T$

行列式 $\det(cA) = c^n \det(A)$, $\det(A^n) = (\det(A))^n$, $\det(A^T) = \det(A)$

一般情况下 $(A+B)^{-1} \neq A^{-1} + B^{-1}$, $\det(A+B) \neq \det(A) + \det(B)$, $AB \neq BA$,

$$A^2 = B^2 \not\Rightarrow A = \pm B \not\Rightarrow A^T A = B^T B$$

如同除数不能为 0 一样，要时刻记住矩阵可不可逆。

④特征值/本征值 Eigenvalue 和特征向量/本征向量 Eigenvector

特征向量是经过矩阵的线性变换后，仍然处于同一条直线上的向量，即 $Ax = \lambda x (x \neq 0)$ 。简单地说，就是 A 使特征向量 x 只缩放不旋转，伸缩的比例就是特征值。

可以看出对于多个 x ，只有特征向量在乘 A 之后方向不变，并且 Ax 最终形成一个椭圆。

并且半长轴和半短轴为 $\sqrt{\frac{1}{|\lambda_2|}}$, $\sqrt{\frac{1}{|\lambda_1|}}$ 。

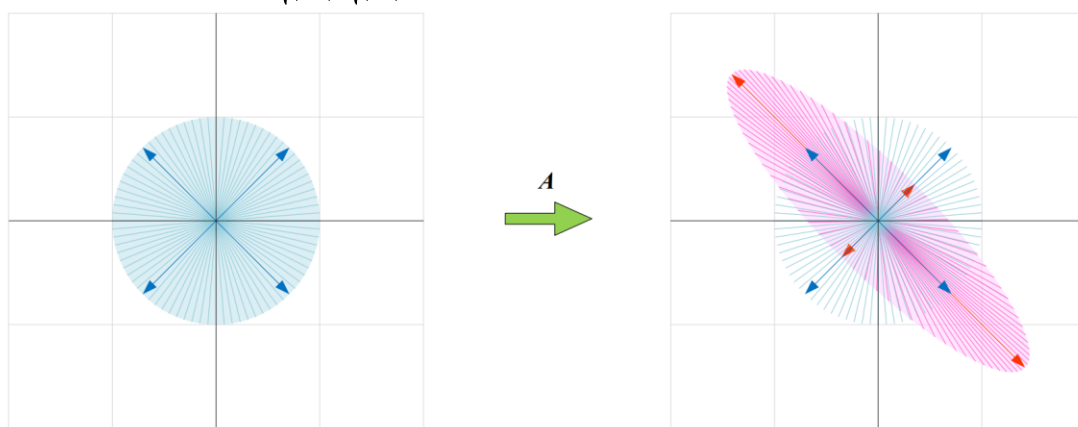


图 4. 矩阵 A 对一系列向量的映射结果

要使 $(A - \lambda I)x = 0$ 有非 0 解，则特征矩阵的行列式 $|A - \lambda I| = 0$ 。解得的 λ_i 称为**特征值**，特征值构成的集合称为矩阵的**谱**(Spectrum)，矩阵的**谱半径** $\rho(A)$ 是特征值绝对值的最大值。

矩阵 A 和 A^T 具有相同的特征值，即 $|A - \lambda I| = |A^T - \lambda I|$ 。 $f(A)$ 的特征值是 $f(\lambda)$ 。

关于同一个特征值 λ 的特征向量的非 0 线性组合仍是关于 λ 的特征向量。关于不同特征值的特征向量线性无关。

实对称矩阵的特征值均为实数且不同特征值的特征向量正交。

设特征值 λ_i ，称 $\sigma_i = \sqrt{\lambda_i} (i=1, 2, \dots, n)$ 为矩阵 A 的**奇异值**。

特征值之**和**是矩阵的迹 $\sum_{i=1}^n \lambda_i = \text{tr}(A)$ ，特征值之**积**是矩阵的行列式的值 $\prod_{i=1}^n \lambda_i = |A|$ 。

广义特征值: $Ax = \lambda Bx$ ，相当于求解 $|A - \lambda B| = 0$ 。如果 B 可逆，则等价于 $B^{-1}Ax = \lambda x$ 。

矩阵的迹 trace: n 阶方阵主对角线上的元素之和 $\text{tr}(A)$ 。 $\text{tr}(A^T) = \text{tr}(A)$, $\text{tr}(AB) = \text{tr}(BA)$ 。

如果 \mathbf{x}, \mathbf{y} 行数相同, 则 $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \text{tr}(\mathbf{xy}^T) = \text{tr}(\mathbf{yx}^T) = \text{tr}(\mathbf{x} \otimes \mathbf{y})$ 。

⑤二次型 Quadric Form

二次型: 由纯二次项构成的函数(不含一次项和常数的二次函数)。如 $x^2 - 2xy + z^2$ 。二次型可以写成矩阵形式 $\mathbf{x}^T \mathbf{A} \mathbf{x}$, 展开后为 $\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$, 满足 $a_{ij} = a_{ji}$ (二次型对应的矩阵是实对称矩阵, 平方项对应矩阵的对角线元素)。

正定二次型: 平方项是非负的, 也就是 $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$ 。如 $f(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + x_3^2$ 。

正定矩阵: 若对任意向量 $\mathbf{x} = (x_1, \dots, x_n)^T \neq \mathbf{0}$, 对称矩阵 \mathbf{A} 均有 $\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$, 则称 \mathbf{A} 为正定矩阵。若 $\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0$, 则称 \mathbf{A} 为半正定矩阵。若 $\mathbf{x}^T \mathbf{A} \mathbf{x} < 0$, 则称 \mathbf{A} 为负定矩阵。若 $\mathbf{x}^T \mathbf{A} \mathbf{x} \leq 0$, 则称 \mathbf{A} 为半负定矩阵。

性质: 正定矩阵的特征值全为正数, 半正定矩阵的特征值全为非负数。

如果对称阵 \mathbf{A} 满足: ①矩阵的特征值均大于 0, ②存在可逆矩阵 \mathbf{P} 使 $\mathbf{A} = \mathbf{P}^T \mathbf{P}$, ③如果 \mathbf{A} 正定, 则 \mathbf{A}^{-1} 也是正定矩阵, ④ \mathbf{A} 的所有顺序主子式均为正。则 \mathbf{A} **正定**。(注: \mathbf{A} 的 k 阶顺序主子式其实就是 \mathbf{A} 的左上区域元素, 也就是行列式

$$\begin{vmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{vmatrix}$$

如果对称阵 \mathbf{A} 满足: ①矩阵的特征值均小于 0, ②存在可逆矩阵 \mathbf{P} 使 $\mathbf{A} = -\mathbf{P}^T \mathbf{P}$, ③ \mathbf{A} 的所有奇数阶顺序主子式均为负, 偶数阶顺序主子式均为正。则 \mathbf{A} **负正定**。

对任意矩阵 $\mathbf{A}_{n \times D}$, $\mathbf{A}^T \mathbf{A}$ 都是**对称半正定**矩阵, 因为 $(\mathbf{A}^T \mathbf{A})^T = \mathbf{A}^T \mathbf{A}$, $\mathbf{x}^T \mathbf{A}^T \mathbf{A} \mathbf{x} = (\mathbf{Ax})^T \mathbf{Ax} \geq 0$, 同样 \mathbf{AA}^T 也是对称半正定矩阵。

标准型: 由纯平方项构成的二次型。如 $x^2 + z^2$ 。标准型中正平方项的数量称为正惯性指数, 负平方项的数量称为负惯性指数。因为二次型矩阵为对称矩阵, 因此一定可以对角化。

以 $x_1^2 + 5x_2^2 + 5x_3^2 + 2x_1x_2 - 4x_1x_3$ 为例, 通过正交变换 $\mathbf{x} = \mathbf{Py}$ 可得标准型 $5y_1^2 + 6y_2^2$:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 5 & 0 \\ -2 & 0 & 5 \end{bmatrix}, |\mathbf{A} - \lambda \mathbf{I}| = (5 - \lambda)(\lambda^2 - 6\lambda) \Rightarrow \begin{cases} \lambda = 5, \mathbf{x}_1 = [0 & 2 & 1]^T \\ \lambda = 6, \mathbf{x}_2 = [1 & 1 & -2]^T \\ \lambda = 0, \mathbf{x}_3 = [5 & -1 & 2]^T \end{cases} \Rightarrow \mathbf{P} = \begin{bmatrix} 0 & \frac{1}{\sqrt{6}} & \frac{5}{\sqrt{30}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{30}} \\ \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{6}} & \frac{2}{\sqrt{30}} \end{bmatrix}$$

⑥瑞利商 Rayleigh Quotient

瑞利商 $R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$, 其中 \mathbf{A} 是对称矩阵, \mathbf{x} 是非 $\mathbf{0}$ 向量。满足① $R(\mathbf{A}, k\mathbf{x}) = R(\mathbf{A}, \mathbf{x})$,

② $\lambda_{\min} \leq R(\mathbf{A}, \mathbf{x}) \leq \lambda_{\max}$ (λ 是特征值, 当 \mathbf{x} 是最小最大特征值对应的特征向量时取等)。

式①说明瑞利商存在冗余, 因此限定 \mathbf{x} 是单位向量 $\mathbf{x}^T \mathbf{x} = 1$, 即 $R(\mathbf{A}, \mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$ 。构造拉格朗日乘子函数 $L(\mathbf{x}, \lambda) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \lambda(\mathbf{x}^T \mathbf{x} - 1)$ 求导得 $2\mathbf{Ax} + 2\lambda\mathbf{x} = \mathbf{0}$, 这就是特征值的定义。而且,

代入特征值 λ_i 有: $R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^T (\lambda_i \mathbf{x})}{\mathbf{x}^T \mathbf{x}} = \frac{\lambda_i \mathbf{x}^T \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \lambda_i$ 。

广义瑞利商 $R(\mathbf{A}, \mathbf{B}, \mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{B} \mathbf{x}}$, 其中 \mathbf{A}, \mathbf{B} 是对称矩阵, \mathbf{x} 是非 $\mathbf{0}$ 向量。同样存在冗余, 加上限定条件 $\mathbf{x}^T \mathbf{B} \mathbf{x} = 1$, 即 $R(\mathbf{A}, \mathbf{B}, \mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x}$ 。同样构造 $L(\mathbf{x}, \lambda) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \lambda(\mathbf{x}^T \mathbf{B} \mathbf{x} - 1)$ 求导得

$2\mathbf{Ax} + 2\lambda\mathbf{Bx} = \mathbf{0}$, 这就是广义特征值的定义。代入特征值 λ_i 有: $R(\mathbf{A}, \mathbf{B}, \mathbf{x}) = \frac{\mathbf{x}^T (\lambda_i \mathbf{Bx})}{\mathbf{x}^T \mathbf{B} \mathbf{x}} = \lambda_i$ 。

广义瑞利商是线性判别分析 LDA 的目标函数, 分母是类内差异, 分子是类间差异。

⑦变换

对角化:

Part 1 相似变换:

若 $P^{-1}AP=B$, 则称 A, B **相似**, P 为相似变换矩阵。相似矩阵有相同的特征值。

设特征值 $\lambda_{1 \sim n}$ 对应特征向量 $x_{1 \sim n}$, 令 $P=[x_1 \cdots x_n]$, 对角矩阵 $\Lambda=\text{diag}(\lambda_1, \cdots, \lambda_n)$, 则 $AP=[Ax_1 \cdots Ax_n]=[\lambda_1 x_1 \cdots \lambda_n x_n]=P\Lambda$, 因此 $P^{-1}AP=\Lambda$ 。说明只要 A 有 n 个线性无关的特征向量, 使得 P 可逆, 就能以 A 的特征向量为列构造一个矩阵 P 。

Part 2 正交变换:

n 阶实对称矩阵 A 有 n 个线性无关的特征向量且实对称矩阵的不同特征值的特征向量正交。于是可以通过格拉姆-施密特方法将同一个特征值的所有特征向量正交化得到一组标准正交基 $p_{1 \sim n}$, 并以此构造相似变换矩阵 P , 则 P 是正交矩阵。通过正交变换 $P^T AP=\Lambda$ 可以将矩阵化为对角阵。事实上因为 P 是正交矩阵, 那么就有 $P^T=P^{-1}$, 因此还是有 $P^{-1}AP=\Lambda$, 所以正交变换是一种特殊的相似变换。

$$\text{以 } A=\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \text{ 为例, } |A-\lambda I|=- (\lambda-2)(\lambda+1)^2 \Rightarrow \begin{cases} \lambda=2 \Rightarrow x_1=[1 & 1 & 1]^T \\ \lambda=-1 \Rightarrow \begin{cases} x_2=[-1 & 1 & 0]^T \\ x_3=[-1 & 0 & 1]^T \end{cases} \end{cases} \text{ 正交单位}$$
$$\text{化得 } p_{1,2,3}, P=[p_1 \ p_2 \ p_3]=\begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix}, \text{ 有 } P^{-1}AP=P^TAP=\begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}。$$

如果 A 是对称矩阵, P 是正交矩阵, 则 $B=P^TAP$ 仍是对称矩阵。

线性变换 Linear Transformation:

与线性映射不同的是, 线性变换不是不可逆的。就像将三维物体投影到二维平面(线性映射, 从一个空间映射到另一个空间)是不可逆的, 线性变换是在当前空间下变化坐标系(数据点还是原来的数据点, 只是因为基底变了, 所以坐标变了)。

线性变换和线性映射的本质区别在于变换的矩阵是不是可逆的, 线性映射的变换矩阵有全 0 的维度, 也就是**不可逆**, 也就是压缩了某一维度的信息(比如三维物体投影到二维平面就了一个维度的信息, 这个维度对应的变换矩阵的那一行/列是全 0 的)。

详细的各种变化可见 [Visualize-ML/Book4_Ch08/P5~P7](#)

注意: 因为平移改变原点, 所以不是线性变换(是仿射变换)。

Part 1 旋转变换

$$\text{设 } x=[x_1 \ x_2]^T \text{ 在极坐标下是 } [r \ \theta]^T, \text{ 则 } x_1=r\cos\theta, x_2=r\sin\theta, \text{ 逆时针 } \alpha \text{ 是 } [r \ (\theta+\alpha)]^T,$$
$$x'=[r\cos(\alpha+\theta) \ r\sin(\alpha+\theta)]^T=[r\cos\alpha\cos\theta-r\sin\alpha\sin\theta \ r\sin\alpha\cos\theta+r\cos\alpha\sin\theta]^T$$
$$=[x_1\cos\theta-x_2\sin\theta \ x_2\cos\theta+x_1\sin\theta]^T=\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

旋转变换矩阵 $T=\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ 是一个正交矩阵, 对应的变换是正交变换。如果一个线性变换能保持向量之间的角度以及向量的长度不变, 则称为**正交变换**。因为 $\det(T)=1$, 所以旋转前后面积不变。

只有 $\det(T)=1$ 的正交矩阵才是旋转矩阵, 一般的正交矩阵是“旋转+镜像”。

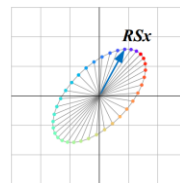
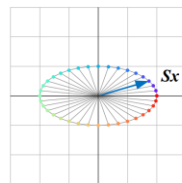
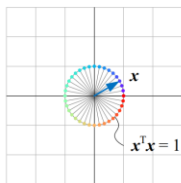
Part 2 缩放变换

矩阵 A 左乘对角矩阵 $\Lambda=\text{diag}(k_1, \cdots, k_n)$ 相当于第 i 行元素都乘 k_i (拉伸系数), 矩阵 A 右乘

对角矩阵 $\Lambda = \text{diag}(k_1, \dots, k_n)$ 相当于第 i 列元素都乘 k_i 。

比如 $\begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} x$ 如右图:

Rotate Scale



对于矩阵 $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \sqrt{a^2+b^2} \begin{bmatrix} a/\sqrt{a^2+b^2} & -b/\sqrt{a^2+b^2} \\ b/\sqrt{a^2+b^2} & a/\sqrt{a^2+b^2} \end{bmatrix} = \sqrt{a^2+b^2} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$

$= \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}_R \begin{bmatrix} \sqrt{a^2+b^2} & 0 \\ 0 & \sqrt{a^2+b^2} \end{bmatrix}_S$ 意味着 A 是缩放与旋转的复合。

另外的, A 的特征值 $\lambda = a \pm bi$ 是复数。显然实数特征值代表着在对应的特征向量方向上的纯拉伸/压缩, 虚数特征值则表现的是旋转+等比放大/缩小, 证明如下:

设虚数特征值 $\lambda = bi$, 由 $e^{i\theta} = \cos \theta + i \sin \theta$ 得 $e^{i\frac{\pi}{2}} = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} = i$, 所以 $\lambda = b \cdot e^{i\frac{\pi}{2}}$ 。也就是倍率为 b 的伸缩变换+逆时针 90° 的旋转变换 (也可以理解为 $i^2 = -1$ 是旋转 180° , 则 i 就是旋转 90°)。

Part3 镜像 reflection

关于通过原点、切向量为 $\tau = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}$ 的直线的镜像变换: $x' = \frac{1}{\|\tau\|^2} \begin{bmatrix} \tau_1^2 - \tau_2^2 & 2\tau_1\tau_2 \\ 2\tau_1\tau_2 & \tau_2^2 - \tau_1^2 \end{bmatrix} x$, 注

意线性变换矩阵的行列式值为 -1, 说明变换前后面积不变而图形翻转 (可以理解为图形的边的次序发生了调转, 这里的次序是指逆/顺时针去观察边的出现先后次序)。如果记该直线与水平轴的夹角为 θ , 则 $\tau = [\cos \theta \quad \sin \theta]^T$, 镜像变换矩阵 $T = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}$ 。

Part4 投影 projection/正交投影 orthogonal projection

将 x 投影到 τ 上的标量长度 $s = \frac{x^T \tau}{\|\tau\|} = x \tau = \tau^T x$, 向量投影 $\text{proj}_\tau(x) = s \frac{\tau}{\|\tau\|} = \frac{\tau^T x}{\|\tau\|^2} \tau$ 。

投影就是将平面的所有点 (平面网格) 坍塌成直线 $\tau = [\tau_1 \quad \tau_2]^T$, (x_1, x_2) 对应的投影点是

$(z_1, z_2): \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \frac{1}{\|\tau\|^2} \begin{bmatrix} \tau_1^2 & \tau_1\tau_2 \\ \tau_1\tau_2 & \tau_2^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ 。注意到 $\det(T) = 0$, 变换矩阵 $T = \frac{1}{\|\tau\|^2} \begin{bmatrix} \tau_1 & \tau_1 \\ \tau_2 & \tau_2 \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} =$

$\frac{1}{\|\tau\|^2} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} @ [\tau_1 \quad \tau_2] = \left(\frac{1}{\|\tau\|} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} \right) @ \left(\frac{1}{\|\tau\|} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} \right)^T = \tau \tau^T$, 式中 $\tau = \frac{1}{\|\tau\|} \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}$ 是 τ 的向量单位化, 称

$\tau \tau^T$ 为投影矩阵, 注意这里得到的 $\begin{bmatrix} z_1 & z_2 \end{bmatrix}^T$ 与我们上面的 $\text{proj}_\tau(x)$ 是一样的。另外, 下文为了简洁, 均认为 τ 是单位向量, 而不再使用 τ 。

镜像的公式可由投影推出，因为原始直线和镜像直线到投影直线的垂线向量之和是 $\mathbf{0}$ ，或者说 $\mathbf{z} = 2\mathbf{p} - \mathbf{x}$ 。因为 $\mathbf{p} = \boldsymbol{\tau}^T \boldsymbol{\tau} \mathbf{x} = \begin{bmatrix} \cos \theta \cos \theta & \cos \theta \sin \theta \\ \cos \theta \sin \theta & \sin \theta \sin \theta \end{bmatrix} \mathbf{x} =$

$$\begin{bmatrix} (\cos 2\theta + 1)/2 & \sin 2\theta/2 \\ \sin 2\theta/2 & (1 - \cos 2\theta)/2 \end{bmatrix} \mathbf{x}, \text{ 可以得出 } \mathbf{z} = \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \mathbf{x}$$

更进一步，定义 $\mathbf{v} \perp \boldsymbol{\tau}$ ，则 $[\boldsymbol{\tau}, \mathbf{v}]$ 是一组标准正交基，满足 $\mathbf{v}^T \mathbf{v} + \boldsymbol{\tau}^T \boldsymbol{\tau} = \mathbf{I}$ 。因此 $\mathbf{z} = 2\boldsymbol{\tau}^T \boldsymbol{\tau} \mathbf{x} - \mathbf{x} = (2\boldsymbol{\tau}^T \boldsymbol{\tau} - \mathbf{I})\mathbf{x} = (\mathbf{I} - \mathbf{v}^T \mathbf{v})\mathbf{x}$ ，式中 $\mathbf{H} = \mathbf{I} - \mathbf{v}^T \mathbf{v}$ 就是**豪斯霍尔德矩阵**，完成的转换叫做豪斯霍尔德反射 Householder reflection，也叫初等反射，向量 \mathbf{v} 的方向就是反射面所在的方向。

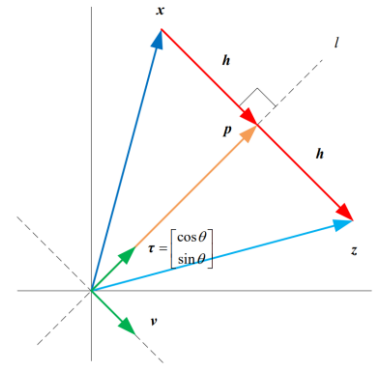
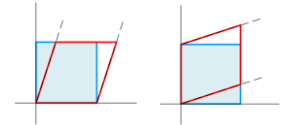


图 10. 投影视角看镜像

对于数据矩阵 $X_{n \times D}$ 和正交矩阵 $V_{D \times D}$ ，记 $Z = XV$ 。对于每个样本 $\mathbf{x}^{(i)}$ ，有 $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} V$ ，也就是 $\mathbf{x}^{(i)}$ 投影得到了**像(image)** $\mathbf{z}^{(i)}$ 。而 $z_{ij} = \mathbf{x}^{(i)} \mathbf{v}_j$ 则代表了 $\mathbf{x}^{(i)}$ 向 \mathbf{v}_j 投影后被压缩成了一个点。

Part 5 剪切 shear

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 & \cot \theta \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ 与 } \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \cot \theta & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ 分别代表沿水平/竖直方向}$$



剪切，剪切角为 θ ，如右图。

⑧矩阵元素求和、去均值——使用全 1 矩阵 \mathbf{I}

Part1 每列元素求和、去均值

$$\text{求和: } (\mathbf{I}_{n \times 1})^T X = \begin{bmatrix} \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{iD} \end{bmatrix}_{1 \times D}$$

$$\text{质心 centroid: } E(X) = \frac{(\mathbf{I}_{n \times 1})^T X}{n} = \begin{bmatrix} \frac{\sum_{i=1}^n x_{i1}}{n} & \cdots & \frac{\sum_{i=1}^n x_{iD}}{n} \end{bmatrix}_{1 \times D} = [\mu_1 \quad \cdots \quad \mu_D]$$

$$\text{质心: } \mu(X) = E(X)^T = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix} = \frac{X^T \mathbf{I}}{n}$$

行向量 $E(X)$ 一般常配合原始数据矩阵 X 一起出现，比如利用广播原则去均值。

列向量 $\mu(X)$ 多用在分布相关运算中，比如多元高斯分布。

$$\text{将 } E(X) \text{ 复制 } n \text{ 份: } \mathbf{I}_{n \times 1} @ E(X)_{1 \times D} = \frac{\mathbf{I}_{n \times 1} \mathbf{I}_{n \times 1}^T X}{n} = \begin{bmatrix} \mu_1 & \cdots & \mu_D \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_D \end{bmatrix}_{n \times D}, \text{ 相当于 } \frac{\mathbf{I}_{n \times 1} \otimes \mathbf{I}_{n \times 1} X}{n},$$

$\mathbf{I}_{n \times 1} \otimes \mathbf{I}_{n \times 1}$ 就是 $n \times n$ 的全 1 方阵，上式就相当于 X 向 \mathbf{I} 正交投影。

对 X 去均值(demean 或 centralize)，减去的是对应列方向上的均值：

$$X_c = X - \frac{\mathbf{I} \mathbf{I}^T X}{n} = \begin{bmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 & \cdots & x_{1D} - \mu_D \\ x_{21} - \mu_1 & x_{22} - \mu_2 & \cdots & x_{2D} - \mu_D \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \mu_1 & x_{n2} - \mu_2 & \cdots & x_{nD} - \mu_D \end{bmatrix}_{n \times D}, \text{ 即 } X_c = IX - \frac{\mathbf{I} \mathbf{I}^T X}{n} = \left(\mathbf{I} - \frac{\mathbf{I} \mathbf{I}^T}{n} \right) X$$

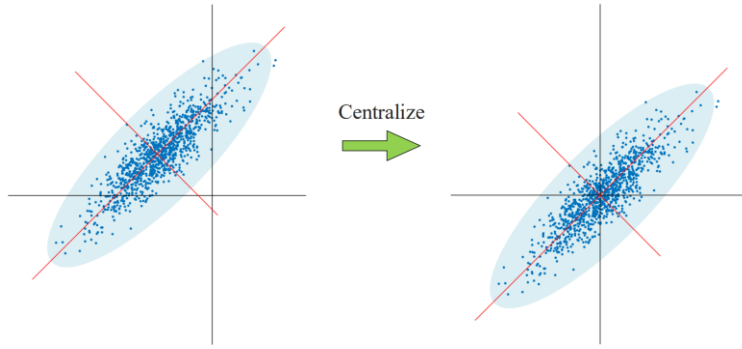


图 10. 去均值的几何视角

Part2 每行元素求和、去均值

$$\text{求和: } X I_{D \times 1} = \begin{bmatrix} \sum_{i=1}^D x_{1i} \\ \vdots \\ \sum_{i=1}^D x_{ni} \end{bmatrix}_{n \times 1}, \quad \text{均值: } \frac{X I_{D \times 1}}{D} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}_{n \times 1}, \quad \text{复制 } \frac{X I I^T}{D} = \begin{bmatrix} \mu_1 & \cdots & \mu_1 \\ \vdots & \ddots & \vdots \\ \mu_D & \cdots & \mu_D \end{bmatrix}_{n \times D}$$

$$\text{去均值: } X_c = X - \frac{X I I^T}{D} = \begin{bmatrix} x_{11} - \mu_1 & x_{12} - \mu_1 & \cdots & x_{1D} - \mu_1 \\ x_{21} - \mu_2 & x_{22} - \mu_2 & \cdots & x_{2D} - \mu_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \mu_D & x_{n2} - \mu_D & \cdots & x_{nD} - \mu_D \end{bmatrix}_{n \times D} = X \left(I - \frac{I I^T}{D} \right)$$

Part3 所有元素求和

$$\text{求和: } (I_{n \times 1})^T X I_{n \times 1} = \begin{bmatrix} \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{iD} \end{bmatrix}_{1 \times D} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \sum_j \sum_{i=1}^n x_{ij}, \quad \text{均值: } \frac{(I_{n \times 1})^T X I_{n \times 1}}{nD} = \frac{\sum_j \sum_{i=1}^n x_{ij}}{nD}$$

1-3.范数

①向量范数

p-范数 (L_p 范数): $\|\mathbf{x}\|_p = \left(|x_1|^p + |x_2|^p + \dots + |x_n|^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} (p \geq 1)$ 用于衡量向量的大小。

$p < 1$ 时不称为范数。

1-范数 (L_1 范数): $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \dots + |x_n|$

2-范数 (L_2 范数): $\|\mathbf{x}\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \dots + |x_n|^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\mathbf{x}^T \mathbf{x}}$, 也就是向量模、向量的长度。

欧氏距离就是 $\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ 。如果 $\|\mathbf{x}\|$ 下标省略, 默认是 $\|\mathbf{x}\|_2$ 。

以 $A\mathbf{x} = \mathbf{b}$ 为例, $\|\mathbf{b}\| = \sqrt{\mathbf{b}^T \mathbf{b}} = \sqrt{\mathbf{x}^T A^T A \mathbf{x}}$, $A^T A$ 叫 A 的格拉姆矩阵, $\mathbf{x}^T (A^T A) \mathbf{x}$ 就是二次型。

∞ -范数 (L_∞ 范数): $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_n|\}$

证明: 设最大值为 M , $\lim_{p \rightarrow +\infty} \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}} = \lim_{p \rightarrow +\infty} \left[M \left(\sum_{i=1}^n \frac{|x_i|^p}{M^p}\right)^{\frac{1}{p}}\right] = M \lim_{p \rightarrow +\infty} \left(\sum_{i=1}^n \frac{|x_i|^p}{M^p}\right)^{\frac{1}{p}}$, 显然 $1 \leq \sum_{i=1}^n \frac{|x_i|^p}{M^p} \leq n$, 因此 $1 = \lim_{p \rightarrow +\infty} 1^{\frac{1}{p}} \leq \lim_{p \rightarrow +\infty} \left(\sum_{i=1}^n \frac{|x_i|^p}{M^p}\right)^{\frac{1}{p}} \leq \lim_{p \rightarrow +\infty} n^{\frac{1}{p}} = 1$, 因此原式 $= M$ (注: $\lim_{n \rightarrow \infty} n^{\frac{1}{n}} = 1$)

②矩阵范数

函数 $\|\cdot\|: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ 称为一个矩阵范数, 是一种“向量→标量”的运算规则, 如果对任意 $m \times n$ 矩阵 A, B 及实数 a 满足:

- ①非负性: $\|A\| \geq 0$, 当且仅当 $A = O$ 时 $\|A\| = 0$ ②齐次性: $\|aA\| = |a| \|A\|$
③三角不等式/次可加性: $\|A + B\| \leq \|A\| + \|B\|$ ④相容性: $\|AB\| \leq \|A\| \|B\|$

满足①~③的范数称为矩阵上的向量范数 (Vector Norm on Matrix)。

诱导范数 Induced Norm: $\|A\|_p = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$, $p=2$ 时称为谱范数 Spectral Norm。

1-范数 满足 $\|A\|_1 = \max_{1 \leq j \leq D} \sum_{i=1}^n |a_{i,j}|$, 是列元素的绝对值之和的最大值。

∞ -范数 满足 $\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^D |a_{i,j}|$, 是行元素的绝对值之和的最大值。

谱范数 满足 $\|A\|_2 = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max(\sigma_1, \dots, \sigma_n)$, 其中 σ 是 A 的奇异值, 也就是 $A^T A$ 的特征值的

平方根。谱范数的平方 $\|A\|_2^2 = \max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T A^T A \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ 是瑞利商的极大值。

Frobenius 范数: 称 $\|A\|_F = \sqrt{\text{tr}(AA^T)} = \sqrt{\sum_{i=1}^n \sum_{j=1}^D |a_{ij}|^2}$ 为矩阵 A 的 F 范数, 也称 L_2 范数。

由柯西不等式, $\|A\mathbf{x}\| \leq \|A\|_F \cdot \|\mathbf{x}\|$, 因此 $\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|A\|_F$, 即 F 范数是谱范数的一个上界。

条件数 $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$, 如果 $\|\cdot\|$ 取谱范数, $\text{cond}(A) = \frac{\max(\sigma_1, \dots, \sigma_n)}{\min(\sigma_1, \dots, \sigma_n)}$ 。

谱归一化 Spectral Normalization: 利普希茨 Lipschitz 连续性 $|f(a) - f(b)| \leq K|b - a|$, 式

中 K 是 Lipschitz 常数, Lipschitz 连续性要求函数在区间上不能有超过线性的变化速度。

在神经网络中, 映射 $\mathbf{x} \rightarrow W\mathbf{x} + \mathbf{b}$ 满足: $\|W\mathbf{x}_1 + \mathbf{b} - W\mathbf{x}_2 + \mathbf{b}\| = \|W\mathbf{x}_1 - W\mathbf{x}_2\| \leq K\|\mathbf{x}_1 - \mathbf{x}_2\|$, 即 $\frac{\|W(\mathbf{x}_1 - \mathbf{x}_2)\|}{\|\mathbf{x}_1 - \mathbf{x}_2\|} \leq K$, 左边的极大值就是 W 的谱范数。因此 W 有较小的谱范数, 则 Lipschitz 常数也较小, 从而**保证输入值的较小改变不会导致输出值的突变**。

谱正则化 Spectral Regularization: 目标函数 $\frac{1}{N} \sum_{i=1}^N L(\mathbf{x}_i, \mathbf{y}_i) + \frac{\lambda}{2} \sum_{i=1}^l \sigma(W^{(i)})^2$, 式中 $L(\mathbf{x}_i, \mathbf{y}_i)$ 为损失函数, l 为神经网络层数, $W^{(i)}$ 为第 i 层的权重矩阵, λ 为正则化项的权重。可以防止权重矩阵出现大的谱范数, 从而保证神经网络的映射有较小的 Lipschitz 常数。

1-4. 矩阵分解

1. LU 分解

可逆矩阵 A 可分解为 $A = LU$, 其中 L, U 分别是下/上三角矩阵。

PLU 分解: $A = PLU$, 其中 P 为置换矩阵 permutation matrix, 任意一行或列只有一个 1, 剩余均为 0, 置换矩阵的作用是交换矩阵的行、列。

2. Cholesky 楚列斯基分解

对称半正定方阵 A 可分解为 $A = LL^T$ 或 $A = R^T R$, 其中 L, R 分别是下, 上三角矩阵。

如果 A 是实对称正定矩阵, 则此分解唯一。

矩阵 L 的元素的计算公式: $l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}$ $l_{ji} = \frac{1}{l_{ii}} \left(a_{ji} - \sum_{k=1}^{i-1} l_{ik} l_{jk} \right), j = i+1, \dots, n$

楚列斯基分解可以用来**判断矩阵的正定性**(能分解就是半正定矩阵), 还可以用来**解方程**: $A\mathbf{x} = \mathbf{b} \Rightarrow LL^T \mathbf{x} = \mathbf{b}$, 令 $\mathbf{y} = L^T \mathbf{x}$, 可先求解 $L\mathbf{y} = \mathbf{b}$ 再求解 $L^T \mathbf{x} = \mathbf{y}$ 。

对于列满秩的矩阵 X , 其格拉姆矩阵 $G = X^T X$ 正定, 就能 Cholesky 分解。

3. LDL 分解

$A = LDL^T$, 其中 L 是下三角矩阵且对角线元素均为 1, D 是对角矩阵, 起缩放作用。如果 D 对角线元素非负, 则 $A = LD^{\frac{1}{2}}(D^{\frac{1}{2}})^T L^T = LD^{\frac{1}{2}}(LD^{\frac{1}{2}})^T$, $LD^{\frac{1}{2}}$ 可以理解为平方根。

如果对 AA^T 进行 LDL 分解, 则 $AA^T = LD^{\frac{1}{2}}(LD^{\frac{1}{2}})^T$, 因此 $A \sim LD^{\frac{1}{2}}$ (但不能推出 $A = LD^{\frac{1}{2}}$),

这代表着经过 A 和 $LD^{\frac{1}{2}}$ 变换的过程不同, 但得到的结果的形状相同(二者变换前后的点的对应关系是不一样的)。

4. QR 分解(正交三角分解)——获取正交系

方阵 A 可分解为 $A = QR$, 其中 Q 是正交矩阵, R 是上三角矩阵。

如果 A 可逆且要求 R 的主对角元为正, 则此分解唯一。

非方阵的矩阵 $A_{n \times D}$ 可以分解为 $A = \begin{cases} Q \begin{bmatrix} R_D \\ 0_{(n-D) \times D} \end{bmatrix}, n > D \\ Q \begin{bmatrix} R_n & B_{n \times (D-n)} \end{bmatrix}, D > n \end{cases}$, 其中 Q 是 n 阶正交矩阵,

R_D, R_n 分别是 D, n 阶上三角矩阵。

QR 分解的方法(以 n 阶方阵为例): 由 $\mathbf{a}_n = \sum_{i=1}^n \mathbf{a}_n^T \mathbf{e}_i \mathbf{e}_i$, 其中 $\mathbf{e}_k = \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}$ 是格拉姆-施密特正交

化得到的 \mathbf{u}_k 进行单位化的结果。也就是 $\begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n \end{bmatrix} \begin{bmatrix} \mathbf{a}_1^T \mathbf{e}_1 & \mathbf{a}_2^T \mathbf{e}_1 & \mathbf{a}_3^T \mathbf{e}_1 & \cdots \\ 0 & \mathbf{a}_2^T \mathbf{e}_2 & \mathbf{a}_3^T \mathbf{e}_2 & \cdots \\ 0 & 0 & \mathbf{a}_3^T \mathbf{e}_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$ 。

5. 特征值分解

有 n 个线性无关的特征向量的 n 阶 **方阵** A 可分解为 $A = Q\Lambda Q^{-1}$ ，其中 $Q = [\mathbf{q}_1 \cdots \mathbf{q}_n]$ ， $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ 。这是因为特征值的定义就是 $AQ = Q\Lambda$ ，比如 $A\mathbf{q}_1 = \lambda_1\mathbf{q}_1, A\mathbf{q}_2 = \lambda_2\mathbf{q}_2 \Rightarrow A[\mathbf{q}_1 \ \mathbf{q}_2] = [\mathbf{q}_1 \ \mathbf{q}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ 。

$A\mathbf{x} = Q\Lambda Q^{-1}\mathbf{x}$ 意味着变换矩阵 A 可以理解为： Q^{-1} 将向量 \mathbf{x} 分解为特征向量的线性组合（因为 $\mathbf{Lx} = Q\mathbf{x}' \Rightarrow \mathbf{x}' = Q^{-1}\mathbf{Lx} = Q^{-1}\mathbf{x}$ ）， Λ 是根据特征值缩放， Q 重新将特征向量组合起来。因此，不需要完整的 A ，我们只需要知道特征值和特征向量就“复原”了 A 的“变换特征”。

另外，注意到我们前面提及 **可对角化** 的条件是 $P^{-1}AP = \Lambda$ ，这与 $A = Q\Lambda Q^{-1}$ 是等价的。

如果 A 可以特征值分解，那么 $\det(A) = \det(\Lambda)$ ，也就是特征值之积。这是因为： $\det(A) = \det(Q\Lambda Q^{-1}) = \det(Q)\det(\Lambda)\det(Q^{-1}) = \det(\Lambda)\det(QQ^{-1}) = \det(\Lambda)$ 。

如果 A 可以特征值分解，则 $A^p = Q\Lambda^p Q^{-1}$ ，因为 $(Q\Lambda^p Q^{-1})^{\frac{1}{p}} = \underbrace{Q\Lambda^p Q^{-1} Q\Lambda^p Q^{-1} \cdots Q\Lambda^p Q^{-1}}_{1/p} = Q\Lambda^p (Q^{-1}Q)\Lambda^p (Q^{-1} \cdots Q)\Lambda^p Q^{-1} = Q\left(\frac{\Lambda^p \cdots \Lambda^p}{1/p}\right)Q^{-1} = Q\Lambda^{p \cdot \frac{1}{p}}Q^{-1} = Q\Lambda Q^{-1} = A$ 。而我们知道 $e^x = 1 + x + \frac{x^2}{2!} + \cdots$ ，因此 $e^A = Q\left(I + \Lambda + \frac{\Lambda^2}{2!} + \cdots\right)Q^T = Qe^\Lambda Q^T = Q \begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_D} \end{bmatrix} Q^T$ 。这是因为 $e^\Lambda =$

$$I + \Lambda + \frac{\Lambda^2}{2!} + \cdots = \begin{bmatrix} 1 + \lambda_1 + \frac{\lambda_1^2}{2!} + \cdots & & \\ \vdots & \ddots & \vdots \\ \cdots & 1 + \lambda_D + \frac{\lambda_D^2}{2!} + \cdots \end{bmatrix} = \begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_D} \end{bmatrix}$$

对于多项式函数 $f(x) = a_n x^n + \cdots + a_1 x$ ，如果 $A = Q\Lambda Q^{-1}$ ，有 $f(A) = Qf(\Lambda)Q^{-1}$ 。

如果 A 是对称矩阵，那么 Q 是正交矩阵，此时也叫 **谱分解** spectral decomposition。而我们知道 Gram 矩阵是对称矩阵，那么有 $X^T X = Q\Lambda Q^{-1}$ ，即 $Q^{-1}X^T X Q = \Lambda$ ，因为 $Q^{-1} = Q^T$ ，则

$$\begin{bmatrix} \mathbf{q}_1^T X^T X \mathbf{q}_1 & \cdots & \mathbf{q}_1^T X^T X \mathbf{q}_D \\ \vdots & \ddots & \vdots \\ \mathbf{q}_D^T X^T X \mathbf{q}_1 & \cdots & \mathbf{q}_D^T X^T X \mathbf{q}_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix}, \text{ 令}$$

$\mathbf{y}_i = X\mathbf{q}_i$ ，显然 $\|\mathbf{y}_i\|^2 = \mathbf{y}_i^T \mathbf{y}_i = (X\mathbf{q}_i)^T X\mathbf{q}_i = \lambda_i$ 。几何意义如右图，对于数据矩阵 X 的第 i 行 $\mathbf{x}^{(i)}$ ，向 \mathbf{q}_j 投影的结果 $y_j^{(i)}$ 就是 $\mathbf{x}^{(i)}$ 在 $\text{span}(\mathbf{q}_j)$ 的坐标， $\|\mathbf{y}_j\|^2$ 是 $\mathbf{y}_j^{(i)}$ 到原点的距离的平方和之和，等于对应的 λ_j 。因此，最大的 λ_k 对应的 $\|\mathbf{y}_k\|^2$ 也是最大的。这昭示着对于 $X^T X$ ，最重要的是 $\lambda_k \mathbf{q}_k \mathbf{q}_k^T$ ，剩

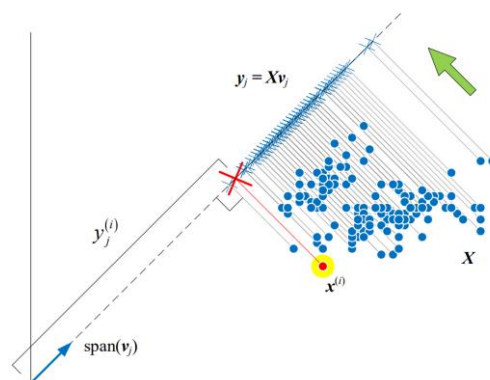


图 8. 数据矩阵 X 向 $\text{span}(\mathbf{v}_j)$ 投影结果为 \mathbf{y}_j ，几何视角

下的成分对 Gram 矩阵 $G = \sum_{i=1}^n \lambda_i \mathbf{q}_i \mathbf{q}_i^T$ 的影响较小, 另外 G 的这个式子是矩阵的第二视角, 也就是每个 $\lambda_i \mathbf{q}_i \mathbf{q}_i^T$ 都是同型矩阵, 最后再叠加。(图中使用的是 \mathbf{v} , 而本段使用的是 \mathbf{q})

6. 奇异值分解(SVD 矩阵分解)

如果特征值分解和奇异值分解的对象都是可对角化矩阵, 两个分解得到的结果等价, 但任何矩阵都能进行奇异值分解。

设矩阵 $X_{n \times D}$ 的秩 $r > 0$, 则存在 U 和 V 使得 $X = U \Sigma V^T$, 其中 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_D)$ (如果 Σ 的形状有缺口则填 0, 一般约定 $\sigma_1 > \sigma_2 > \dots > \sigma_D$), 为矩阵 X 的全部非零奇异值构成的对角矩阵, 称此为矩阵 X 的奇异值分解。左、右奇异向量 U, V 满足 $U^T U = I, V^T V = I$ 。

奇异值分解的几何含义: 对于任何一个矩阵, 我们要找到一组两两正交单位向量序列, 使得矩阵作用在此向量序列上后得到新的向量序列保持两两正交。奇异值的几何含义: 这组变换后的新的向量序列的长度。<https://zhuanlan.zhihu.com/p/31387766> <https://www.zhihu.com/tardis/bd/ans/225371236>

SVD 分解四大类。经济型是将 Σ 的 0 的部分去除, 紧凑型是当 $\text{rank}(X) = r < D$ 的时候 $r-1 \sim D$ 的位置的元素是 0, 截断型是一种近似(取前 $p(p < r)$ 个奇异值):

类型名	完全型	经济型	紧凑型	截断型
英文	full	economy-size, thin	compact	truncated
Σ 的形状	$n \times D$	$D \times D$	$r \times r$	$p \times p$
整体形状	$n \times n, n \times D, D \times D$	$n \times D, D \times D, D \times D$	$n \times r, r \times r, D \times r$	$n \times p, p \times p, D \times p$

如果对 XX^T 进行 SVD 分解, 则 $XX^T = U \Sigma V^T (U \Sigma V^T)^T = U \Sigma V^T V \Sigma^T U^T = U \Sigma \Sigma^T U^T$, 而我们之前提及 XX^T 的特征值分解是 $X^T X = Q \Lambda Q^{-1}$, 这说明 U 的列向量是 XX^T 的特征向量, $\Sigma \Sigma^T$ 是 XX^T 的特征值矩阵, 故**特征值与奇异值的关系是** $\lambda_i = s_i^2 (i = 1 \sim D), \lambda_i = 0 (i = D+1 \sim n)$ 。

当 U 是 $n \times n$ 方阵时, 对比 QR 分解 $A = QR$ 与 SVD 分解 $A = U (\Sigma V^T)$, 二者都含有正交矩阵, 但 Q 的 \mathbf{q}_1 与 \mathbf{x}_1 平行, U 则是在优化 Σ (使 Σ 满足 $\sigma_1 > \sigma_2 > \dots > \sigma_D$)。

当 Σ 是**经济型**时, $XV = U \Sigma \Rightarrow X_{n \times D} [\mathbf{v}_1 \ \dots \ \mathbf{v}_D] = [\mathbf{u}_1 \ \dots \ \mathbf{u}_D] \text{diag}[s_1, \dots, s_D]$, 因此 $[X\mathbf{v}_1 \ \dots \ X\mathbf{v}_D] = [s_1 \mathbf{u}_1 \ \dots \ s_D \mathbf{u}_D]$, 即 $X\mathbf{v}_i = s_i \mathbf{u}_i$, 即 $\|X\mathbf{v}_i\| = \|s_i \mathbf{u}_i\| = \|s_i\|$ 。也能通过 $X = U \Sigma V^T = [\mathbf{u}_1 \ \dots \ \mathbf{u}_D] \text{diag}[s_1, \dots, s_D] [\mathbf{v}_1^T \ \dots \ \mathbf{v}_D^T]^T = s_1 \mathbf{u}_1 \mathbf{v}_1^T + \dots + s_D \mathbf{u}_D \mathbf{v}_D^T$, 其中 $s_1 \mathbf{u}_1 \mathbf{v}_1^T$ 最能还原原矩阵 (奇异值 s_i 的大小决定了成分的重要性, 而 $\mathbf{u}_i, \mathbf{v}_i$ 决定了投影方向), 又因为 $s_1 \mathbf{u}_1 \mathbf{v}_1^T$ 的列向量的基础是 \mathbf{v}_1^T , 则列向量之间存在倍数关系, 因此 $\text{rank}(s_1 \mathbf{u}_1 \mathbf{v}_1^T) = 1$ 。简单地说, 奇异值往往对应着矩阵中隐含的重要信息, 且重要性和奇异值大小正相关。每个矩阵都可以表示为一系列秩为 1 的“小矩阵”之和, 而奇异值则衡量了这些“小矩阵”对于该矩阵的权重。因此可以用于数据压缩(只使用奇异值较大的部分)、噪声处理(认为噪声的奇异值较小)。

如, 求 $A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ 的 SVD 分解:





记 $B = A^T A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, 特征方程 $0 = |\lambda E - B| = \begin{vmatrix} \lambda - 1 & 0 & -1 \\ 0 & \lambda - 1 & -1 \\ -1 & -1 & \lambda - 2 \end{vmatrix} = \lambda(\lambda - 3)(\lambda - 1)$, 得 B 特征

值为 3,1,0，则 A 的奇异值为 $\sqrt{3},1,0$ 。 B 的特征值对应的特征向量解出来为 $\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$ 。

$$r(A)=2,\Sigma=\begin{pmatrix}\sqrt{3} & 0 \\ 0 & 1\end{pmatrix},\text{ }V\text{ 为归一化后的特征向量拼接成的矩阵 }V=\begin{pmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{3}}\end{pmatrix},$$

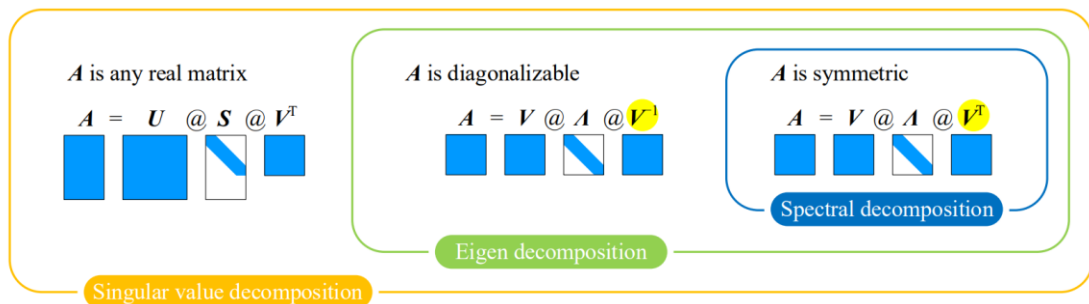
$$U_1=AV_1\Sigma^{-1}=\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0\end{pmatrix},U_2=\begin{pmatrix}0 \\ 0 \\ 1\end{pmatrix},U=(U_1:U_2)=\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1\end{pmatrix}$$

表 1. 四种常用矩阵分解

矩阵分解	QR 分解	Cholesky 分解	特征值分解	SVD 分解
前提	任何实数矩阵都可以 QR 分解	正定矩阵才能 Cholesky 分解	可对角化矩阵才能进行特征值分解	任何实数矩阵都可以 SVD 分解
示意图	$A = Q @ R$ 	$A = R^T @ R$ 	$A = V @ \Lambda @ V^{-1}$ 	$A = U @ S @ V^T$ 
公式	$A = QR$	$A = R^T R$ $A = LL^T$	$A = V\Lambda V^{-1}$ $A = V\Lambda V^T$ (A 为对称方阵时，其特征值分解又叫谱分解)	$A = USV^T$ (注意 V 的转置运算)
结果	Q 是正交矩阵 (完全型分解)，意味着 Q 是规范正交基 R 是上三角矩阵	L 为下三角方阵 R 为上三角方阵	Λ 为对角方阵，对角线元素为特征值 V 列向量为特征向量 如果 A 为对称方阵， V 为正交矩阵，即满足 $V^T V = V V^T = I$	U 为正交矩阵 (完全型分解)，它的列向量为左奇异向量 S 主对角线元素为奇异值 V 为正交矩阵 (完全型分解)，它的列向量为右奇异向量 U 和 V 都是规范正交基

几何视角	Q 代表旋转	写成 LDL^T 形式 (L 主对角线元素为 1) L 代表剪切 D 代表缩放	V 代表旋转 A 代表缩放	U 代表旋转 S 代表缩放 V 代表旋转
结果唯一?	A 列满秩, 且 R 的对角元素为正实数的情况下结果唯一	当限定 R 的对角元素为正时, 分解结果唯一	矩阵 V 不唯一 本书的特征向量都是单位向量, 特征向量一般差在正负符号上	矩阵 U 和 V 不唯一 本书左奇异向量、右奇异向量都是列向量
特殊类型	完全型 (Q 是正交矩阵) 经济型 (Q 是规范正交基, 但不是正交矩阵)	正定矩阵 埃尔米特矩阵 (不在本书讨论范围)	对称矩阵 正规矩阵 (不在本书讨论范围之内)	完全型 经济型 缩略型 截断型
向量空间	Q 的列向量为规范正交基, Q 的第一列向量 q_1 是 A 的第一列向量 a_1 的单位化 R 的列向量相当于坐标值	如果 $A = X^T X$ (即 Gram 矩阵) 正定, 对 A 进行 Cholesky 分解得到上三角矩阵 R , R 的列向量可以代表 X 列向量	如果 A 为对称方阵, V 为规范正交基 如果 $A = X^T X$ 且 X 列满秩, V 是 X 的行空间 $R(X)$	完全型 SVD 分解获得四个空间: 列空间 $C(X)$ 和左零空间 $\text{Null}(X^T)$, 行空间 $R(X)$ 和零空间 $\text{Null}(X)$ 完全型 SVD 分解相当于一次性完成两个特征值分解
优化视角			$\arg \max_v v^T A v$ 或 subject to: $v^T v = 1$ $\arg \max_{x \neq 0} \frac{x^T A x}{x^T x}$	$\arg \min_v \ A v\ $ 或 subject to: $\ v\ = 1$ $\arg \min_{x \neq 0} \frac{\ A x\ }{\ x\ }$
Numpy 函数	numpy.linalg.qr()	numpy.linalg.cholesky()	numpy.linalg.eig()	numpy.linalg.svd()
本章分解对象	原始数据矩阵 X	格拉姆矩阵 $G(X^T X)$ 协方差矩阵 Σ 相关性系数矩阵 P	格拉姆矩阵 $G(X^T X)$ 协方差矩阵 Σ 相关性系数矩阵 P	原始数据矩阵 X 中心化数据矩阵 X_c 标准化数据矩阵 Z_X
本系列丛书主要应用	解线性方程组 最小二乘回归 施密特正交化	蒙特卡罗模拟, 产生满足特定协方差矩阵要求的随机数 判断正定性	马尔科夫过程 主成分分析 瑞利商 矩阵范数	求解伪逆矩阵 矩阵范数 最小二乘回归 主成分分析 图像压缩

其中谱分解 \subset 特征值分解 \subset 奇异值分解:



二、高数

2-1.一元函数

泰勒公式: $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n$

弧长: $s = \int_a^b \sqrt{1+y'^2} dx = \int_a^b \sqrt{x'^2(t)+y'^2(t)} dt$

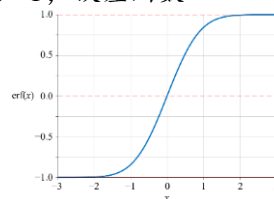
一阶线性微分方程 $y' + a(x)y = b(x) \Rightarrow y = e^{-\int a(x)dx} \left(e^{\int a(x)dx} b(x) dx + C \right)$

高斯函数 $f(x) = e^{-x^2}$ 的积分是 $\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$, 由此可得 $\int_{-\infty}^{+\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = 1$, 误差函数

errorfunction: $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ 。

高斯函数与误差函数有如下关系: $F(x) = \int_{-\infty}^x e^{-t^2} dt = \frac{\sqrt{\pi}}{2} \operatorname{erf}(x) + \frac{\sqrt{\pi}}{2}$ 。

二元高斯函数与高斯函数的关系: $\int_{-\infty}^{\infty} e^{-x^2-y^2} dy = \sqrt{\pi} e^{-x^2}$ (二元高斯函数对 y 偏积分变为关于 x 的高斯函数)



2-2.多元函数

泰勒公式: 若 $f(x_1, \dots, x_m)$ 在 $\mathbf{a} = (a_1, \dots, a_m)^T$ 点处 n 阶可导, 则该点处的泰勒公式是

$$f(\mathbf{x}) = \sum_{p=0}^n \frac{1}{p!} \left(\Delta x_1 \frac{\partial}{\partial x_1} + \dots + \Delta x_m \frac{\partial}{\partial x_m} \right)^p f(a_1, \dots, a_m) + o(\|\Delta \mathbf{x}\|^n)$$
$$= f(\mathbf{a}) + (\nabla f(\mathbf{a}))^T (\mathbf{x} - \mathbf{a}) + \frac{1}{2} (\mathbf{x} - \mathbf{a})^T H(\mathbf{x} - \mathbf{a}) + o(\|\mathbf{x} - \mathbf{a}\|^2) (n=2)$$

n 重积分: $\int_{\mathbb{R}^n} \exp(-\mathbf{x}^T \mathbf{x}) d\mathbf{x} = \pi^{\frac{n}{2}}$, 该结论可以用多维正态分布。

偏导: 如果偏导数连续, 则混合偏导数与求导次数无关。

梯度: $\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}^T$, 也就是函数 f 对所有的自变量 $x_i (i=1 \sim n)$

的偏导数组成的向量。

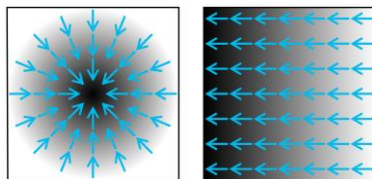
如右图, 标量场的值用灰度表示, 越暗表示越大的数值, 而其相应的梯度用蓝色箭头表示。

对 D 维向量 \mathbf{x} 和任意 $A \in \mathbb{R}^{n \times D}$ 有: $\nabla_{\mathbf{x}} A\mathbf{x} = A^T$

对 D 维向量 \mathbf{x} 和任意 $A \in \mathbb{R}^{D \times n}$ 有: $\nabla_{\mathbf{x}} \mathbf{x}^T A = A$

对 D 维向量 \mathbf{x} 和任意 $A \in \mathbb{R}^{D \times D}$ 有: $\nabla_{\mathbf{x}} \mathbf{x}^T A \mathbf{x} = (A + A^T) \mathbf{x}$ 。特别地, $\nabla_{\mathbf{x}} \|\mathbf{x}\|^2 = \nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{x} = 2\mathbf{x}$

对于任何矩阵 \mathbf{X} 也有: $\nabla_{\mathbf{x}} \|\mathbf{X}\|_F^2 = 2\mathbf{X}$



对 $\mathbf{y} = W\mathbf{x}$ 与损失函数 $f(\mathbf{y})$, 依次将 W, \mathbf{x} 视为常数, 有 $\nabla_{\mathbf{x}} f = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^T \nabla_{\mathbf{y}} f = W^T \nabla_{\mathbf{y}} f$ 和

$\nabla_W f = (\nabla_{\mathbf{y}} f) \mathbf{x}^T$ 。

偏导数在图像边缘检测的应用:

以灰度图为例, 因为图像的点是离散的, 因此需要用差分近似代替。

为了减少噪声的干扰, 我们不使用单侧差分公式 $f'(x) \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$ ($\Delta x \rightarrow 0$), 而是使用

中心差分公式 $f'(x) \approx \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x}$ ($\Delta x \rightarrow 0$)。也就是 Sobel 算子：

$$\begin{cases} f_x \approx f(x+1, y-1) - f(x-1, y-1) + 2f(x+1, y) - 2f(x-1, y) + f(x+1, y+1) - f(x-1, y+1) \\ f_y \approx f(x-1, y+1) - f(x-1, y-1) + 2f(x, y+1) - 2f(x, y-1) + f(x+1, y+1) - f(x+1, y-1) \end{cases}$$

相当于卷积核 $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ 和 $\begin{bmatrix} -1 & -2 & 1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ ，其中给中间一行/列赋的权重为 2。

如果要综合考虑水平和垂直方向的边缘强度，就需要使用梯度的模，这就是 HOG(梯度方向直方图 Histogram of Oriented Gradient)的思想： $M = \sqrt{f_x^2 + f_y^2}$, $\alpha = \arctan \frac{f_y}{f_x}$ 。因为两个相反的方向被认为是相同的，因此 $\alpha \in [0^\circ, 180^\circ]$ 。直方图可以看出哪些方向上的梯度权重很大，也就是边缘方向主要是该梯度的垂直方向。

雅可比矩阵：

将 $y_1 = f_1(x_1, \dots, x_D), \dots, y_n = f_n(x_1, \dots, x_D)$ 记作 $\mathbf{y} = f(\mathbf{x})$ ，称 $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_D} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_D} \end{bmatrix}_{n \times D}$ 为**雅**

可比矩阵，每一行对应一个多元函数的梯度。

考虑 $\mathbf{y} = A\mathbf{x}$ ，有 $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = A$ （因为 $\frac{\partial y_i}{\partial x_j} = a_{ij}$ ）。比如 $\begin{cases} x = r \cos \theta \\ y = r \sin \theta \end{cases} \Rightarrow \frac{\partial(x, y)}{\partial(r, \theta)} = \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix}$

称雅可比方阵的行列式为**雅可比行列式**。

如果 $\mathbf{y} = f(\mathbf{x})$ 满足 $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \neq 0$ ，则逆映射 $\mathbf{x} = f^{-1}(\mathbf{y})$ 存在且 $\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^{-1}$ 且 $\left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| = \left| \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right|^{-1}$ 。

对于 $z = f(y_1, \dots, y_n), y_j = g_j(x_1, \dots, x_D)$ ，有 $\frac{\partial z}{\partial x_i} = \sum_{j=1}^n \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}$ ，写成矩阵形式是：

$$\nabla_{\mathbf{x}} z = \begin{bmatrix} \frac{\partial z}{\partial x_1} \\ \vdots \\ \frac{\partial z}{\partial x_D} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_1} \\ \vdots \\ \sum_{j=1}^n \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_D} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_D} & \dots & \frac{\partial y_n}{\partial x_D} \end{bmatrix} \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \vdots \\ \frac{\partial z}{\partial y_n} \end{bmatrix} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^T \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \vdots \\ \frac{\partial z}{\partial y_n} \end{bmatrix} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^T \nabla_{\mathbf{y}} z。$$

多项式定理：

$$(u_1 + \dots + u_m)^n = \sum_{p_1 + \dots + p_m = n} \frac{n!}{p_1! \dots p_m!} (u_1^{p_1} \dots u_m^{p_m})$$

三、概率统计

3-1.四个公式

①条件概率 $P(B|A) = \frac{P(AB)}{P(A)}$ $P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1 \cdots A_{n-1})$

②全概率公式 $P(B) = \sum_{i=1}^n P(A_i)P(B|A_i)$

③贝叶斯公式 $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ ，因 prior A ，果 evidence B ，先验概率 prior $P(A)$ ，

后验概率 posterior $P(A|B)$ ，似然函数 likelihood $P(B|A)$ 。

结合②与③可得 $P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^n P(A_i)P(B|A_i)}$

④条件独立 $P(A|B,C) = P(A|C)$ ，在 C 发生的情况下 B 是否发生并不影响 A 。 A 和 B 关于 C 条件独立可以记为 $A \perp B|C$ 。“在 C 发生的情况下”相当于是更换了样本空间。

结合①和④可得 $P(AB|C) = P(A|BC)P(B|C) = P(A|C)P(B|C)$

注意**独立不能推出条件独立，条件独立不能推出独立**。比如投掷骰子 x,y ，事件 $A:x$ 奇数，事件 $B:y$ 奇数，事件 $C:x,y$ 奇偶相同， A,B 独立但不条件独立。事件 $A:x$ 是 1，事件 $B:x$ 是 1 或 2，事件 $C:x$ 是 1 或 2， $P(A) = \frac{1}{6}, P(B) = P(C) = \frac{1}{3}$ ，满足 $P(AB|C) = \frac{1}{2} = P(A|C)P(B|C)$ 但 $P(AB) = \frac{1}{3}$ ， A,B 条件独立但不独立。

3-2.随机变量

取值可变且每个取值对应一个概率的变量叫随机变量 Random Variable。

离散型随机变量：

概率质量函数/分布列 Probability Mass Function(PMF): $p(x_i) = P(X = x_i)$

累积分布函数 Cumulative Distribution Function(CDF): $p(X \leq x_j) = \sum_{i=1}^j p(x_i)$

连续性随机变量：

概率密度函数 Probability Density Function(PDF): $f(x) \geq 0, \int_{-\infty}^{+\infty} f(x)dx = 1$

累积分布函数 Cumulative Distribution Function(CDF): $F(y) = p(X \leq y) = \int_{-\infty}^y f(y)dy$

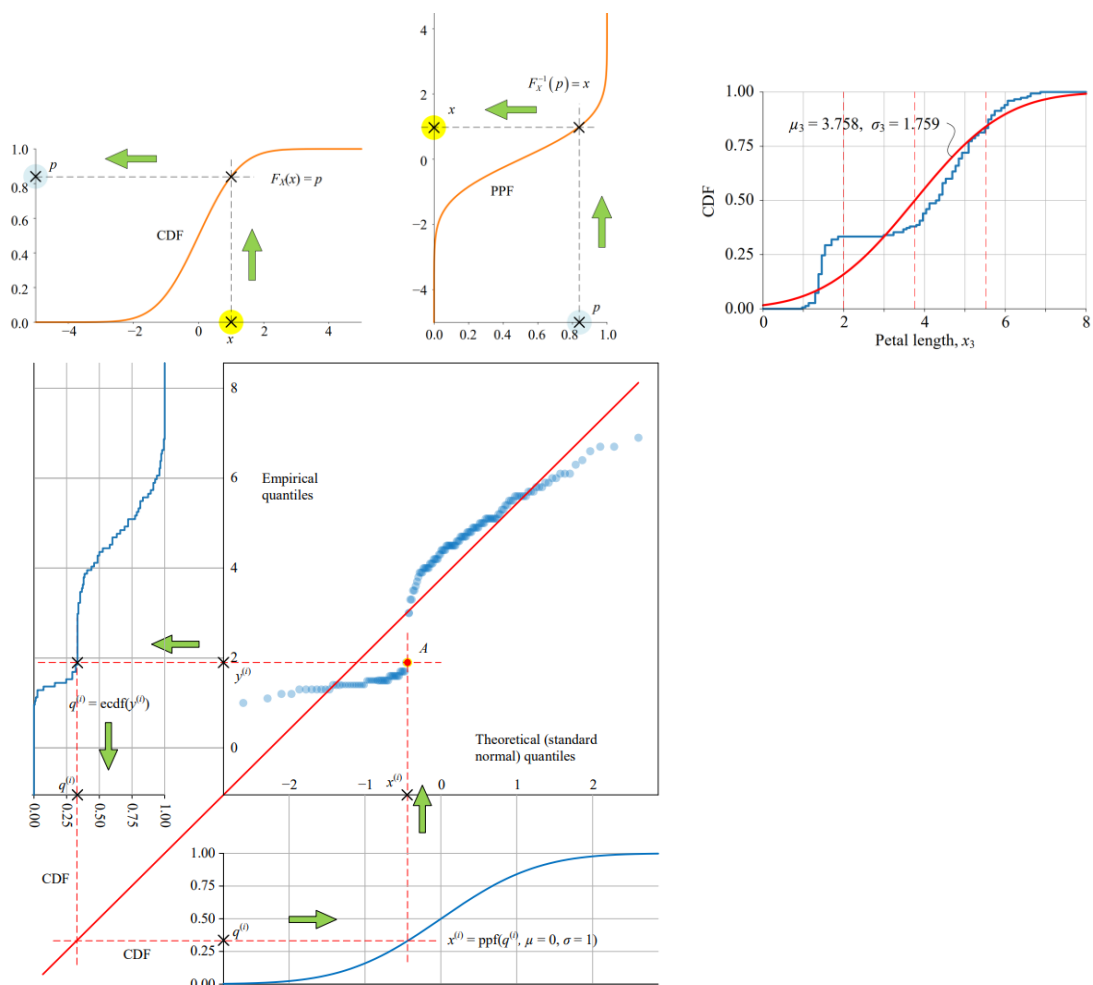
PDF 与 CDF 满足 $p(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1) \approx f(\xi)\Delta(x_2 - x_1)$ ，式中 $\xi \in [x_1, x_2]$

PMF 就是概率值，而 PDF 需要 n 重积分才能得到概率值。

Percent-Point Function (PPF,百分点函数)是 CDF 逆函数，如下左图是正态分布的 PPF。

经验分布函数(empirical cumulative distribution function, ECDF)用来描述一组样本数据分布。横坐标表示数据的取值，它的纵坐标则表示小于等于横坐标的数据比例。沿着数据取值的增加，每遇到一个新的样本，都跳跃 $1/n$ ，是一个阶跃函数，如下右图是 Iris 的特征 3。

QQ 图(quantile-quantile plot)中的 Q 代表分位数，用于检查数据是否符合某个分布的统计图形。QQ 图是散点图，横坐标一般为假定分布(比如标准正态分布)分位数(ECDF)，纵坐标为待检验样本的分位数。在对角线之上代表样本在此处的分布比 ECDF 假定的更多。如下下图是 Iris 的特征 3。



期望、方差、标准差：

$$EX = \begin{cases} \sum_i x_i p(x_i) \\ \int_{-\infty}^{+\infty} xf(x)dx \end{cases}, \quad DX = E[(X - EX)^2] = \begin{cases} \sum_i (x_i - EX)^2 p(x_i) \\ \int_{-\infty}^{+\infty} (x - EX)^2 f(x)dx \end{cases} = EX^2 - E^2X, \quad \sigma = \sqrt{DX}$$

$DX = E[X^2] - E[X]^2$ 中的 $E[X^2]$ 代表样本 X 以原点为基准的离散程度， $E[X]^2$ 代表整体 X 以原点为基准的离散程度，因为前者的平方是针对单个样本 X ，后者的平方是针对整体 EX 。当 $EX = 0$ 时，也就是 X 质心在原点，此时 $DX = E[X^2]$ 。

σ 的提出是为了统一量纲，比如样本单位是 cm ，方差单位就是 cm^2 ，标准差单位是 cm 。

Jensen 不等式：对凸函数，有 $E[f(x)] \geq f(E[x])$ ，这是因为凸函数满足 $f(\sum_{i=1}^n a_i x_i) \leq \sum_{i=1}^n f(a_i x_i)$ ，

其中 $\sum_{i=1}^n a_i = 1, a_i \geq 0$ ， $E[x] = \sum_{i=1}^n a_i x_i$ ， $E[f(x)] = \sum_{i=1}^n f(a_i x_i)$ 。

3-3. 常见概率分布

表：(分布图像可见 [Visualize-ML_Book5_Ch05_图 1](#)，关系可见 <https://www.math.wm.edu/~leemis/chart/UDR/UDR.html>)

		PMF / PDF	EX	DX
离散均匀分布(整数)		$P(X = k) = \frac{1}{b - a + 1}$	$\frac{a + b}{2}$	$\frac{(b - a + 2)(b - a)}{12}$

伯努利分布	$X \sim B(p)$	$P(X=1)=p, P(X=0)=1-p$	p	$p(1-p)$
二项分布	$X \sim B(n, p)$	$P(X=k) = C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
几何分布	$X \sim G(p)$	$P(X=k) = (1-p)^{k-1} p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
超几何分布		$P(X=k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}$	$n \frac{K}{N}$	$n \frac{K(N-K)(N-n)}{N^2(N-1)}$
泊松分布	$X \sim p(\lambda)$	$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!} (\lambda > 0)$	λ	λ
均匀分布	$X \sim U(a, b)$	$f(x) = \frac{1}{b-a} (a \leq x \leq b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布-连续的泊松	$X \sim E(\lambda)$	$f(x) = \lambda e^{-\lambda x} (x \geq 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
正态分布/高斯分布	$X \sim N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2
t 分布	$T \sim t(\nu)$	$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} (1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$	$0(\nu > 1)$	$\begin{cases} \frac{\nu}{\nu-2} (\nu > 2) \\ \infty \end{cases}$

泊松分布用于建模随机事件的发生次数。要求 n 非常大, p 非常小, np 存在有限的极限 λ 。用于描述在给定的时间段、距离、面积等范围内随机事件发生的概率, 如每小时走入商店的人数, 一定时间内机器出现故障的次数, 一定时间内交通事故发生的次数等等。

多项分布: 是二项分布的推广 $P(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n! p_1^{x_1} \dots p_k^{x_k}}{x_1! x_2! \dots x_k!}$, 需要 $\sum_{i=1}^k x_i = n$ 。含义是在 n 次独立重复的试验中, 每次试验有 k 个可能的结果中的一个发生的次数的概率分布。比如抽 5 次卡, 出 1 个斯卡蒂 2 个浊心斯卡蒂 2 个艾雅法拉的概率。

负二项分布: Binomial 关注的是 N 次实验中成功的次数 $k \in [0, N]$, Negative Binomial 关注的是 r 次实验失败时成功的次数 $k \in [0, +\infty)$, 因此负二项分布的 PMF 是 $P(X=k) = \binom{k+r-1}{k} p^k (1-p)^r = (-1)^k \binom{-r}{k} p^k (1-p)^r$ 。

对数正态分布: 考虑 $Z \sim N(0,1)$, 则 $X = a^{\mu+Z\sigma}$ 称为 Log-Normal Distribution, 有 $f(x) = \frac{\exp\left(-\frac{(\log_a x - \mu)^2}{2\sigma^2}\right)}{x \ln a \sqrt{2\pi}\sigma}$, 满足 $EX = e^{\mu+\frac{\sigma^2}{2}}$, $DX = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$, 图像呈现左偏。

为什么正态分布这么常见: 中心极限定理: 大量独立同分布随机变量的均值经适当标准化后依分布收敛于正态分布, 也就是**独立同分布、随机、相加**。以高尔顿钉板为例, **独立**就是每个弹珠不会受其他的弹珠影响, **同分布**就是顶上只有一处开口(弹珠的起始状态一致), **随机**就是撞到钉子后的路线是随机的, **相加**就是每次弹珠事件会为某个位置的弹珠数量加 1。另外, 钉板开口是否位于顶部中央是无所谓的, 开在别的位置, 分布形态不变(只是平移)。

为什么有些样本不是正态分布: ①属性 A 不仅直接作用于样本, 还可能影响属性 B 从而间接作用于样本, 分析 A 的时候并不一定考虑到了 B。②不够随机, 结果很可能会偏向某一边。③数据的定义域不是 $(-\infty, +\infty)$ 。③样本数据不满足简单的相加(比如癌细胞分裂是相乘, 使得癌细胞数量随着时间呈指数级增长)。

正态分布中的椭圆: 对于二元正态分布 $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) = N(\mu, \Sigma)$:

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}} \exp \left\{ -\frac{1}{2} \left(\frac{1}{1-\rho_{1,2}^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 - 2\rho_{1,2} \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 \right] \right) \right\}$$

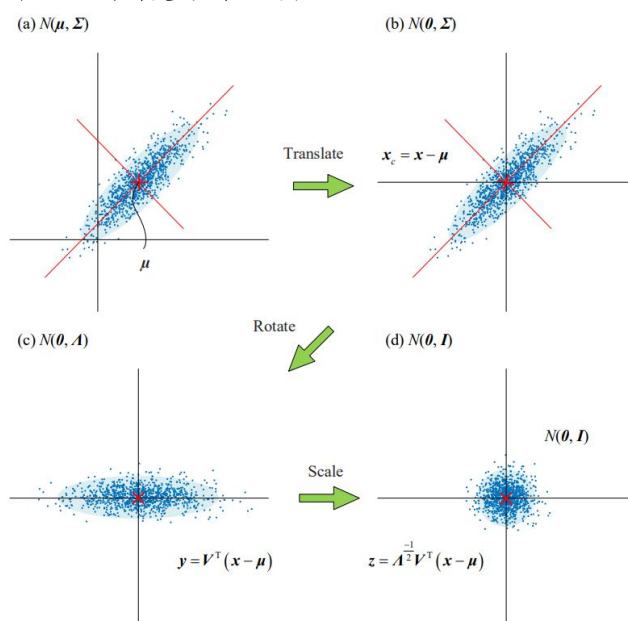
$$\text{对于多元: } f_{\chi}(\mathbf{x}) = \frac{\exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}}, \text{ 其中 exp 化马氏 distance 为 similarity,}$$

下面是归一化与缩放特征值。

协方差矩阵 $\boldsymbol{\Sigma}$ 半正定，如果其正定，则 $\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \Rightarrow \boldsymbol{\Sigma}^{-1} = \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T$ 。那么：

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) = \left[\boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) \right]^T \left[\boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) \right] = \left\| \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) \right\|_2^2$$

说明这个过程是将 \mathbf{x} (超椭圆分布) 通过平移 $\mathbf{x} - \boldsymbol{\mu}$ 、旋转/镜像 \mathbf{V}^T 、缩放 $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ 变成了 \mathbf{z}' (超球分布)，如下图是二维上的椭圆变为圆的过程。



Z-分数：也叫标准分数， $z = \frac{x - \mu}{\sigma}$ ，这个过程叫做数据的**标准化**(standardize)。不同的 z

值对应不同的区域面积，可见 Book5_Ch09_图 12 的表格，越大的 z 值表示越多比例的数据在 $-z \sim z$ 之间（用于 Z 检验的时候，比如有 95% 的数据位于 $(-z, z)$ 之间对应的 z 值是 1.96）。

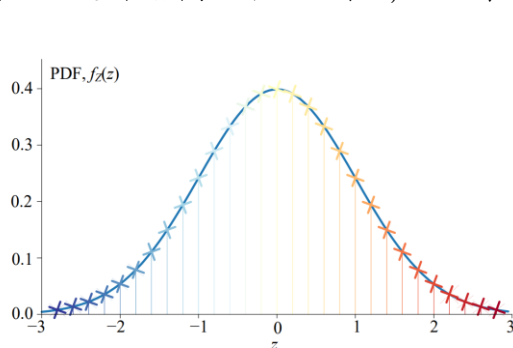


图 9. 标准正态分布 z 和 PDF 的对应关系

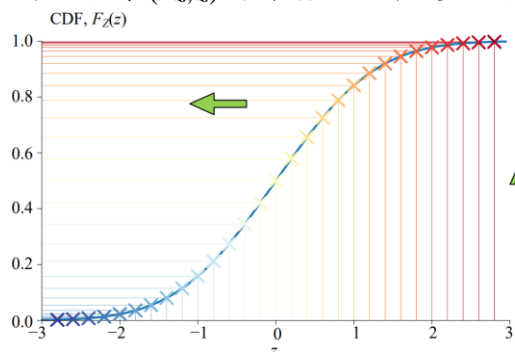


图 10. 标准正态分布 z 和 CDF 值的映射关系

请注意 **standardize** 和 **normalize**(归一化)的区别，后者表示将一组数据转化为 $[0, 1]$ 区间的数值，另外向量单位化(vector normalization)指的是将非零向量转化成 L^2 模为 1 的单位向量。

拉普拉斯分布 Laplace distribution: $f_X(x) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$, 和**高斯分布**类似, 而拉普拉斯分布的 PDF 图像在对称轴处存在尖点。 μ 决定概率密度分布位置, b 决定分布形状。满足 $EX = \mu, DX = 2b^2$ 。

逻辑分布: 一元 $f_X(x) = \frac{\exp\left[-\frac{(x-\mu)}{s}\right]}{s \left\{1 + \exp\left[-\frac{(x-\mu)}{s}\right]\right\}^2}, F_X(x) = \frac{1}{1 + \exp\left[-\frac{(x-\mu)}{s}\right]}$, 与高斯分布几

乎一样, 但是逻辑分布“**厚尾**”。

学生 t-分布是**厚尾分布**, 多应用于根据小样本数据来估计呈正态分布且方差未知的总体的均值。式中自由度 $\nu = n-1$ 。当 $\nu \rightarrow +\infty$ 时, t 分布趋近于正态分布。

多元 t-分布 $f_X(x) = \frac{\Gamma\left(\frac{\nu+D}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \nu^{\frac{D}{2}} \pi^{\frac{D}{2}} |\Sigma_t|^{\frac{1}{2}}} \left[1 + \frac{1}{\nu} (x-\mu)^T \Sigma_t^{-1} (x-\mu)\right]^{-\frac{\nu+D}{2}}$, 式中 $\Sigma_t = \frac{\nu}{\nu-2} \Sigma$,

其中 Σ 是多元高斯分布的协方差矩阵。

伽玛函数是阶乘的推广: $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ 满足 $\Gamma(x+1) = x\Gamma(x)$,

有 $\Gamma(\nu) = (\nu-1)!, \nu \in \mathbb{N}$, $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\sqrt{\pi}$, 满足:

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} = \begin{cases} \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2} & (\nu \% 2 = 1) \\ \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3} & (\nu \% 2 = 0) \end{cases}$$

自由度 ν (degree of freedom): 当以样本的统计量来估计总体的参数时, 样本中独立或能自由变化的数据的个数, 称为该统计量的自由度。自由度等于独立变量数减掉其衍生量数。比如, 方差的定义是样本减平均值 (一个由样本决定的衍生量) 的平方之和, 因此对 N 个随机样本而言, 其自由度为 $N-1$ 。

其余见 Book5_7.7 及之后。

幂律分布: $P(X=k) = C \cdot k^{-\alpha}$ 或 $f(x) = C \cdot x^{-\alpha}$, 式中 $C = \left(\sum_{i=1}^n x_i^{-\alpha}\right)^{-1}$ 或 $\left(\int_{x_{\min}}^{x_{\max}} x^{-\alpha} dx\right)^{-1}$ 为归

一化常数。产生幂律分布的原因有①**偏好依附模型** preferential attachment model (马太效应): 以点赞为例, 每个人以 p 的概率点赞一篇新文章, 以 $1-p$ 的概率点赞某个现有的文章, 点赞某个特定的现有文章的概率是该特定文章的点赞数除以目前的全部点赞数。②**自组织临界模型**: 在系统中各个组件建立了相互依赖的关系, 产生了幂律分布, 当到达某个临界点状态后, 系统状态就会快速变化。比如雪崩 (超过临界点时, 事件发生的概率就急速上升)。

3-4. 分布变换

设 $Y = g(X)$, $g(x)$ 严格单调且存在反函数 $g^{-1}(y) = h(y)$, 则有 $f_Y(y) = f_X(h(y)) |h'(y)|$ 。

例如 $Z \sim N(0,1), X = \sigma Z + \mu$, 则 $Z = \frac{X-\mu}{\sigma}, f_X(x) = f_Z(h(x)) |h'(x)| =$

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x-\mu}{\sigma}\right)^2}{2}\right) \cdot \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = N(\mu, \sigma^2)。$$

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

3-5. 随机向量

将某个样本的 D 个属性的随机变量值组合成随机向量 \mathbf{x} 。

离散型随机向量：

边缘概率质量函数分别是： $p_X(x) = \sum_y p(x, y)$, $p_Y(y) = \sum_x p(x, y)$ 。

如果将 x 视为 n 个样本， y 视为 D 个属性，那么 $p(x) = p_X(x)$ 可以理解为按行求和，得到的大小是 $(n, 1)$ 。 $p(y) = p_Y(y)$ 可以理解为按列求和，得到的大小是 $(1, D)$ 。

连续型随机向量：

边缘概率密度函数分别是： $f_X(x) = \int_{-\infty}^{+\infty} f(x, y)dy$, $f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)dx$ 。

贝叶斯公式：

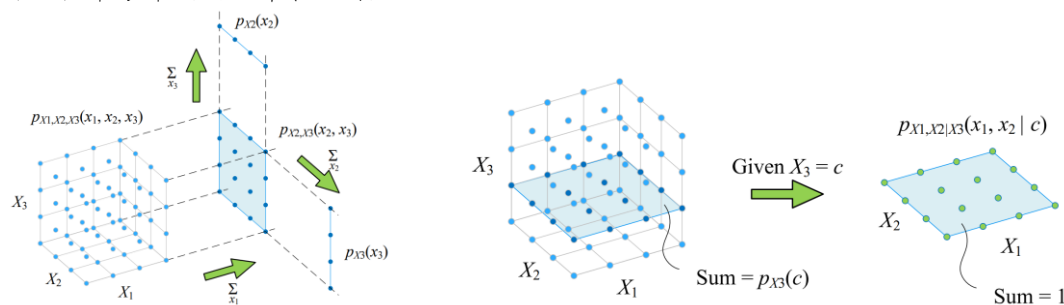
$$f_{Y|X}(y|x) = \frac{f(x|y)f(y)}{f(x)} = \frac{f(x|y)f(y)}{\int_{-\infty}^{+\infty} f(x, y)dy} = \frac{f_{X,Y}(x, y)}{f_X(x)}, f(y|x) \text{ 即为 } f_{Y|X}(y|x), \text{ 后同。也就是条件}$$

概率是联合概率除以边缘概率（注意 $f_X(x)$ 虽然是具体的值，但它是概率密度而不是概率）。

边缘概率相当于是条件概率的加权平均 $f(y) = \int_x f(y|x)f(x)dx$ 。

偏求和是个降维过程，如下左图，把立方体在不同维度上压扁。沿着哪个方向求和，就相当于完成了这个维度上的合并，该维度消失。

而**条件概率**是切片过程，如下右图，只考虑切片上的概率分布情况，而不考虑整个立方体的概率分布（缩小样本空间）。



另外由此可得后验概率公式是 $f(C_k | x_1, x_2) = \frac{f(x_1, x_2, C_k)}{f(x_1, x_2)}$ 。

独立同分布：

如果 $f(x, y) = f_X(x)f_Y(y)$ 几乎处处成立（不成立点是有限集或无限可数集），则它们相互独立。如果 X, Y 相互独立，则 $E[XY] = E[X]E[Y]$ 。如果一组随机变量**相互之间独立且服从同一种分布**，则称它们是**独立同分布**(IID, Independent And Identically Distributed)

如果样本集 $\mathbf{x}_i, i=1 \sim D$ 独立同分布且服从概率 $p(\mathbf{x})$ ，则它们的联合概率为 $p(\mathbf{x}_1, \dots, \mathbf{x}_D)$

$= \prod_{i=1}^D p(\mathbf{x}_i)$ 。该假设（样本服从独立同分布）常用于最大似然估计等。

协方差：

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)] = \begin{cases} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} (x_{i_1} - EX)(y_{i_2} - EY) p(x_{i_1}, y_{i_2}) \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - EX)(y - EY) f(x, y) dx dy \end{cases} = E[XY] - E[X]E[Y]$$

X, Y 不相关 $\Leftrightarrow X, Y$ 独立 $\Leftrightarrow f(x, y) = f(x)f(y) \Leftrightarrow \text{cov}(X, Y) = 0$

不相关：无线性关系。**独立**：无线性、非线性关系。也就是协方差衡量的是线性相关性。但对二维正态分布 (X, Y) ，独立和不相关等价。

方差： $D\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n DX_i + 2\sum_{i=1}^n \sum_{j=i+1}^n \text{cov}(X_i, X_j)$ $\xrightarrow{n=2} D[X+Y] = DX + DY + 2\text{cov}(X, Y)$ ，因

此当 X_1, \dots, X_n 相互独立时，和的方差等于方差的和 $D\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n DX_i$ 。并且 $D(a_1X_1 + a_2X_2)$

$$= a_1DX_1^2 + a_2DX_2^2 + 2\text{cov}(X_1, X_2) = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}^T \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathbf{a}^T \Sigma \mathbf{a} \text{ 是二次型。}$$

$$\text{协方差矩阵: } \Sigma = \begin{bmatrix} \text{cov}(x_1, x_1) & \cdots & \text{cov}(x_1, x_D) \\ \vdots & \ddots & \vdots \\ \text{cov}(x_D, x_1) & \cdots & \text{cov}(x_D, x_D) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{bmatrix} \text{ 是半}$$

正定矩阵。

线性相关系数/皮尔逊相关系数：

$$\rho_{X,Y} = \text{corr}(X, Y) = \text{cov}\left(\frac{X - EX}{\sqrt{DX}}, \frac{Y - EY}{\sqrt{DY}}\right) = \frac{\text{cov}(X, Y)}{\sqrt{DX}\sqrt{DY}} = \frac{EXY - EXEY}{\sqrt{EX^2 - E^2X}\sqrt{EY^2 - E^2Y}}$$

$|\rho| = 1$ 说明一定有 $Y = aX + b$ 。 $\rho = 0$ 说明 X, Y 不相关，也就是 $\text{cov}(X, Y) = 0$ 。

因为 $D[X+Y] = DX + DY + 2\text{cov}(X, Y)$ 可得 $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho_{X,Y}\sigma_X\sigma_Y$ ，这是余弦定理。
 $c^2 = a^2 + b^2 - 2ab\cos\theta$

因为 $\text{cov}(X, Y) = 2\rho_{X,Y}\sigma_X\sigma_Y$ ，因此协方差矩阵与相关系数矩阵的关系是 $\Sigma = DPD$ ，其中 $D = \text{diag}(\sigma_1, \dots, \sigma_D)$ 。

对于两列数据 \mathbf{x}, \mathbf{y} ， $\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\| \|\mathbf{y}_c\|}$ ，其中 $\mathbf{x}_c = \mathbf{x} - E\mathbf{x}$ 。这说明相关系数与余弦相似度

$$k(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \text{ 十分接近。}$$

惯性 inertia：

描述样本数据的紧密程度，就是总离差平方和 SSD(Sum of Squares for Deviations, SSD)：

$$\text{SSD}(X) = \sum_{i=1}^n \text{dist}(\mathbf{x}^{(i)}, E(X))^2 = \sum_{i=1}^n \|\mathbf{x}^{(i)} - E(X)\|_2^2 = \text{trace}(X_c^T X_c)，\text{式中 } X_c = X - E(X)。$$

矩：

X 的 k 阶原点矩 $\nu_k = EX^k$ ， k 阶原点绝对矩 $\alpha_k = E|X|^k$ 。

X 的 k 阶中心矩 $\mu_k = E(X - EX)^k$ ， k 阶中心绝对矩 $\beta_k = E|X - EX|^k$ 。(以 EX 为中心)

偏度、峰度：

偏度描述的是概率分布的偏离(非对称)程度：

$$\text{Skew}(X) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{\mu_3}{\beta_3} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right]^{\frac{3}{2}}}$$

$\text{Skew}(X) < 0$: 左偏(负偏, 众数 > 中位数 > 均值)。 $\text{Skew}(X) = 0$: 对称。 $\text{Skew}(X) > 0$: 右偏。
 因为是偏离(而不是偏向), 所以左偏是指离开了左边, 也就是峰值靠右, 左边长尾。

偏度为零不一定意味着分布对称, 可见 Visualize-ML_Book5_Ch02_图 35。

偏度检验: skewtest 函数返回值: ① statistic: The computed z-score for this test. 偏度检验的统计量, 它用来衡量数据的偏斜程度。 ② pvalue: The p-value for the hypothesis test. 偏度检验的 p 值, 它用来判断统计量的显著性, 也就是数据与某一假设模型一致程度的概率。

如果 z-score 的绝对值较大, 且 p 值小于显著性水平(通常设定为 0.05), 就可以拒绝原假设。

峰度描述的是概率分布曲线的陡峭程度, -3 的峰度称为超值峰度(excess kurtosis):

$$\text{Kurt}(X) = E\left(\frac{X - \mu}{\sigma}\right)^4 - 3 = \frac{E(X - EX)^4}{(E(X - EX)^2)^2} - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right]^2} - 3$$

$\text{Kurt}(X) < 0$: 平峰, 薄尾。 $\text{Kurt}(X) = 0$: 正态分布。 $\text{Kurt}(X) > 0$: 尖峰, 厚尾。

3-6. 极限定理

切比雪夫 Chebyshev 不等式: $p(|x - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$, 式中 $\mu = EX, \sigma^2 = DX, \varepsilon$ 是任一正数。

随机变量离期望越远, 则落入该区间的概率越小。

大数定律(LLN, Law of Large Numbers): 任取 $\varepsilon > 0$, 若恒有 $\lim_{n \rightarrow +\infty} p(|\bar{X}_n - E\bar{X}_n| < \varepsilon) = 1$,

式中 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, 则称 $\{X_n\}$ 服从(弱)大数定律。样本数量越多时, 样本的算术平均值有越大的概率接近其真实的概率分布的期望。

林德贝格-勒维 Lindeberg-Levy 中心极限定理(独立同分布中心极限定理):

$\lim_{n \rightarrow +\infty} p\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x) = N(0, 1)$ 。标准误(standard error) $SE = \frac{\sigma}{\sqrt{n}}$ 。多次地独立地从总体中

抽取样本, 并计算每次样本的平均值, 并用这些样本平均值去估算总体的期望。

3-7. 参数估计

最大似然估计(MLE, Maximum Likelihood Estimation): 就是找到让似然函数取得最大值的参数。似然函数是优化要估计的参数 θ 的函数 $L(\theta) = \prod_{i=1}^n p(x_i; \theta)$, 也就是联合概率。取

对数 $\ln L(\theta) = \sum_{i=1}^n \ln p(x_i; \theta)$ 。要让最大化样本出现的概率, 就是 $\theta = \arg \max_{\theta} \sum_{i=1}^n \ln p(x_i; \theta)$, 相当于求 $\ln L(\theta)$ 的极大值点。

以伯努利分布为例, n 个样本中有 α 个取值为 1, 则似然函数为 $\ln L(p) = \ln(p^\alpha (1-p)^{n-\alpha})$
 $= \alpha \ln p + (n-\alpha) \ln(1-p)$, $\frac{\partial \ln L(p)}{\partial p} = 0 \Rightarrow p = \frac{\alpha}{n}$ 。

区间估计(interval estimate): 区间估计在点估计的基础上附加误差限(margin of error)来构造置信区间(confidence interval), 置信区间对应的概率, 被称为置信度(confidence level)。

总体方差已知，估计均值： $\Pr\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$ ，式中 $z_{1-\alpha/2} = F_{N(0,1)}^{-1}\left(1 - \frac{\alpha}{2}\right) = -F_{N(0,1)}^{-1}\left(\frac{\alpha}{2}\right)$ 也就是 z 分数，其中 $F_{N(0,1)}^{-1}(\cdot)$ 就是标准正态分布的逆累计分布函数 ICDF。比如 95% 的置信概率： $\Pr\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$ 。

总体方差未知，估计均值： $\Pr\left(\bar{X} - t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}}\right) = 1 - \alpha$ ，式中样本均方差 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2}$ ，t 分布的自由度为 $n-1$ ，CDF 值为 $1 - \frac{\alpha}{2}$ 。

总体均值未知，估计方差： $\Pr\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)}\right) = 1 - \alpha$ ， χ 是卡方分布。

最大后验概率估计(MAP, Maximum A Posteriori Probability Estimate): 相较于 MLE, MAP 认为 θ 服从某种概率分布，此时求解的是 $\theta = \arg \max_{\theta} \sum_{i=1}^n p(x|\theta)p(\theta)$ ，也就是认为参数服从概率分布 $p(\theta)$ ，多了 $p(\theta)$ 这一项。

贝叶斯估计: 相较于 MAP，贝叶斯估计不是求出参数 θ 具体的值，而是求出其概率分布 $p(\theta|x)$ 。 $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int_{\theta} p(x|\theta)p(\theta)d\theta}$ ，参数的估计量是 $E[p(\theta|x)]$ 。

核密度估计(KDE, Kernel Density Estimation)/Parzen 窗技术: 无须求出概率密度函数的参数，而是用一组标准函数的叠加来表示概率密度函数。

概率密度函数估计值 $p(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^n k\left(\left\|\frac{x-x_i}{h}\right\|^2\right) = f_{h,k}(x)$ ，式中 h 是核函数的窗口半径/带宽/缩放系数(手动设置，过小不平滑，过大丢信息)，核函数 K 要满足 x 到 x_i 的距离越远函数值越小(这样，如果 x 附近样本点越多，函数值就会越大)， $\frac{1}{nh^d}$ 为了保证 $\int_{-\infty}^{+\infty} p(x) = 1$ (引理: $f(x)$ 与 $cf(cx)$ 的面积相同，其中 c 为常数)， $c_{k,d}, k$ 在下文提及。下图是高斯核估计， \times 代表一组数据点 $\{-3, -2, 0, 2, 2.5, 3, 4\}$ ：

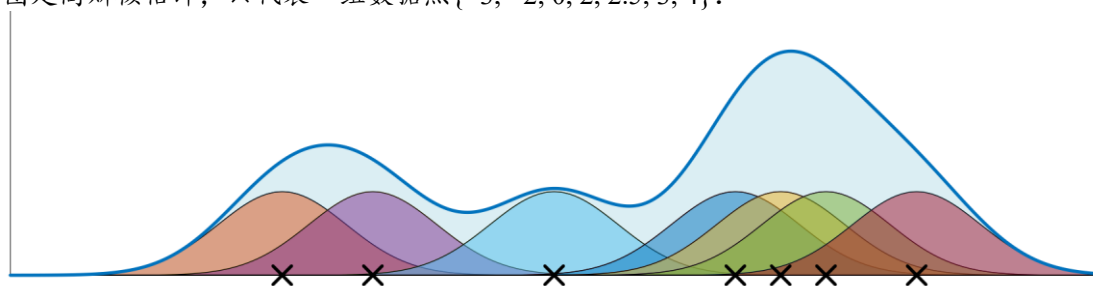


图 6. 用 7 个高斯核函数构造得到的概率密度估计曲线

常见的核函数是径向对称核 $K(x) = c_{k,d} k(\|x\|^2)$ ，式中 $c_{k,d}$ 是归一化常数(核函数与水平面构成图形的面积为 1)， $k(x)$ 为核的剖面(profile)函数，是 $\|x\|$ 的减函数且对点 x 关于原点径向对称。如：

Epanechnikov 核: $k(x) = \begin{cases} 1-x & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$, $K(x) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2) (1-\|x\|^2) & \|x\| \leq 1 \\ 0 & \text{else} \end{cases}$ ，式中 c_d

是 d 维单位球的体积。Epanechnikov 剖面函数 $k(x)$ 在 $x=1$ 处不可导。

高斯核: $k(x) = e^{-\frac{1}{2}x^2}$, $K(x) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}\|x\|^2}$, 这里的归一化系数用到了 n 重积分里的结论 $\int_{\mathbb{R}^n} \exp(-x^T x) dx = \pi^{\frac{n}{2}} \Rightarrow \int_{\mathbb{R}^d} \exp(-\frac{1}{2}x^T x) dx = 2\pi^{\frac{d}{2}}$ 。

EM 算法 Expectation Maximization(期望最大化): 每个样本包含可观测数据 x_i 、可观测属性 w_i 、不可观测属性(隐变量) z_i (常为类别标签值), 比如两个班成绩已知, 但不知道每个成绩对应的班级是什么。

初始: 随机指定参数 θ 的值。

E 步: 依据当前的 θ 值, 计算 x 在 w_i 的条件下的条件概率, 相当于是 θ, w_i 的条件下 x_i 的 z_i 的值。

M 步: 依据 z_i 的值更新 θ 值。

参考: BV1RT411G7jJ

均值漂移 Mean Shift: 找到概率密度函数的极大值点。

令 $g(x) = -k'(x) = -\nabla_x k\left(\left\|\frac{x-x_i}{h}\right\|^2\right) = -k'\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \frac{2}{h^2}(x-x_i)$, 因为剖面函数是减函数,

所以 $g(x) = -k'(x) > 0$, 将 $g(x)$ 代入 $f_{h,K}(x)$ 得 $\nabla f_{h,K}(x) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)(x_i - x) = \dots =$

$$\frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right) \right] \left[\frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \right], \text{ 令 } m_{h,G}(x) = \frac{\sum_{i=1}^n x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \text{ (均值漂}$$

移向量是使用核函数 G 进行加权之后的 x_i 均值与 x 之间的差异), 那么迭代公式: $x_{t+1} = x_t + m_t$, 式中 m_t 是第 t 次迭代时计算出来的均值漂移向量。

令 $f_{h,G}(x) = \frac{c_{g,d}}{nh^2} \sum_{i=1}^n g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)$, 则 $\nabla f_{h,K}(x) = \frac{2c_{k,d}}{h^2 c_{g,d}} f_{h,G}(x) m_{h,G}(x)$, 相当于 $m_{h,G}(x) =$

$\frac{1}{2} h^2 \frac{c_{g,d}}{c_{k,d}} \frac{\nabla f_{h,K}(x)}{f_{h,G}(x)}$, 也就是用核函数 G 计算出来的均值漂移向量正比于核函数 K 的梯度值归一化后的值, 归一化系数依据 x 处用 G 计算出来的密度估计值计算。

3-8. 随机算法

线性同余法: $x_{i+1} = (ax_i + b) \bmod m$, m 控制随机整数的范围, 可取 $a = 7^5, b = 0, m = 2^{31} - 1$ 。

Box-Muller 算法: 设 $u_1, u_2 \sim U[0,1]$ 独立, 则 $z_1 = \sqrt{-2\ln u_1} \cos(2\pi u_2), z_2 = \sqrt{-2\ln u_1} \sin(2\pi u_2)$ 相互独立且服从 $N(0,1)$ 。通过变换 $x = \sigma z + \mu$ 能从 $N(0,1)$ 变成 $N(\mu, \sigma^2)$ 。通过 $z = [z_1 \dots z_n]^T$ 能从 $N(0,1)$ 变成 $N(0, I)$ 。通过 $x = Az + \mu$ 能从 $N(0, I)$ 变成 $N(\mu, \Sigma)$, 其中 A 满足 $\Sigma = AA^T$ 。

遗传算法: 求解 $\arg \min_x f(x)$, 假设优化变量是正整数, 用 8 位编码(如 0000 0010 = 2)。

初始化: 随机生成 n 个可行解 $x_1^{(0)}, \dots, x_n^{(0)}$ 。

for (k=1; k<max_iter; k++)

评估: 计算 $f(x_i^{(k-1)})$ 并按解的优劣(适应度函数)分配抽样概率。

选择: 按照抽样概率选出 $x_1^{(k)}, \dots, x_n^{(k)}$, 选择一部分进行交叉, 进行随机变异。

交叉 Crossover: 对两个编码互换若干个二进制位。

变异 Mutation: 随机选择编码的若干二进制位进行变异(1→0, 0→1)。

蒙特卡洛算法: 用几何模拟概率。

$$\textcircled{1} E[f(\mathbf{x})] = \int_{\mathbb{R}^n} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \Rightarrow E[f(\mathbf{x})] = \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}_i)$$

$$\textcircled{2} \int_D f(\mathbf{x})d\mathbf{x} \stackrel{p(\mathbf{x})=\frac{1}{s(D)}, \mathbf{x} \in D}{\Rightarrow} \int_D f(\mathbf{x})d\mathbf{x} = \int_D \frac{1}{p(\mathbf{x})} p(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int_D s(D) \frac{1}{s(D)} f(\mathbf{x})d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^N s(D)f(\mathbf{x}_i)$$

比如计算 $\iint_D f(x,y)dxdy$, $D: a \leq x \leq b, c \leq y \leq d$, 区域 D 的测度是 $s(D) = (b-a)(d-c)$, 那么

$$\iint_D f(x,y)dxdy = \frac{1}{N} \sum_{i=1}^n (b-a)(d-c)f(x_i, y_i) = \frac{(b-a)(d-c)}{N} \sum_{i=1}^n f(x_i, y_i)。$$

3-9.采样方法

拒绝采样 Rejection Sampling: 先生成容易采样的概率分布 $q(\mathbf{x})$ (提议分布), 再拒绝一部分样本, 使剩下的样本服从难以直接采样的目标概率分布 $p(\mathbf{x})$ 。显然任意点处 $c \cdot q(\mathbf{x}) \geq p(\mathbf{x})$, 其中 c 是手动设置的参数, 也就是 $q(\mathbf{x})$ 在乘上系数之后要能覆盖住 $p(\mathbf{x})$ 。

对 $q(\mathbf{x})$ 中的 \mathbf{x} 计算其概率值 $\alpha(\mathbf{x}) = \frac{p(\mathbf{x})}{c \cdot q(\mathbf{x})}$, 然后以 $\alpha(\mathbf{x})$ 的概率接受这个样本(这可以通过生成随机数 $z \sim U(0,1)$, 如果 $z \leq \alpha(\mathbf{x})$ 就接受 \mathbf{x})。

提议分布通常选用均匀分布、正态分布这样简单的分布。 c 要尽量小以防止算法效率太低(大部分候选样本被拒绝)。

重要性采样 Importance Sampling: 如果只计算数学期望而不是需要服从概率分布 $p(\mathbf{x})$, 此时可用重要性采样。同样构造 $q(\mathbf{x})$ 并采样, 然后计算随机变量函数 $f(\mathbf{x})$ 对概率分布 $p(\mathbf{x})$ 的期望: $E_p[f(\mathbf{x})] = \int_{\mathbb{R}^n} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^n} f(\mathbf{x})w(\mathbf{x})q(\mathbf{x})d\mathbf{x} = E_q[f(\mathbf{x})w(\mathbf{x})]$, 其中 $w(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}$ 称为权重。也就是从提议分布中采样出样本, 然后计算此分布下 $f(\mathbf{x})w(\mathbf{x})$ 的数学期望, 该值等价于 $E_p[f(\mathbf{x})]$ 。

3-10.信息论

①基本思想: 越不可能发生的事情发生了, 其包含的信息量就越大。

②信息熵 Information Entropy: 表述随机变量不确定度的度量。熵越大, 意味着发生的可能性越大。

$$H(X) = E_p[-\ln p(x)] = - \sum_{x \in X} p(x) \log p(x)$$

所有样本等几率出现的情况下, 熵达到最大值(所有可能的事件等概率时不确定性最高), 对于样本等几率分布而言, 样本数越大, 熵值越大(可能的事件越多, 不确定性越高)。也就是说, 样本点越多、越均匀分布, 熵值越大、不确定性越强、信息量越小。

为什么公式是对数? 由 $H(x,y) = H(x) + H(y)$, $P(x,y) = P(x)P(y)$, 那么 $H(x)$ 一定与 $P(x)$ 的对数有关。

为什么有一个负号: 负号是为了确保信息一定是正数或者是 0, 总不能为负数吧。

为什么底数为 2 或 e: **①信息量满足低概率事件对应于高的信息量。** **②熵是服从某一特定概率分布事件的理论最小平均编码长度**, 如果用 0,1 编码, 底数就是 2。也可以说, 熵是当分配的概率真正匹配数据生成过程时的信息量的期望。注: 在通信领域通常以 2 为底(以比特为单位), 在机器学习里通常以 e 为底(以奈特为单位)。下文均取以 e 为底。

推广到连续性随机变量, **微分熵 Differential Entropy:** $H(p) = - \int_{-\infty}^{+\infty} p(x) \ln p(x) dx$ 。

值域: 离散型 $0 \leq H(p) \leq \ln n$, 连续性 $0 \leq H(p) \leq \ln(\sqrt{2\pi}\sigma) + \frac{1}{2}$ 。当取某一值的概率为 1、取其他值的概率为 0 时, 熵有极小值。当**均匀分布**(离散型)/**正态分布**(连续型)时, 熵有极大值(另外, 正态分布的熵只与方差有关而与均值无关)。

$$\text{联合熵 Joint Entropy: } H(X, Y) = \begin{cases} -\sum_x \sum_y p(x, y) \ln p(x, y) \\ -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \ln p(x, y) dx dy \end{cases}$$

联合熵是熵对多维概率分布的推广。

多维正态分布 $N(\mu, \Sigma)$ 的联合熵 $H(x) = \frac{n}{2} \ln(2\pi) + \frac{1}{2} \ln|\Sigma| + \frac{n}{2}$ 只与协方差矩阵有关而与均值向量无关。

③交叉熵 Cross Entropy:

$$H(P, Q) = E_p[-\ln q(x)] = \begin{cases} -\sum_x p(x) \ln q(x) \\ -\int_{-\infty}^{+\infty} p(x) \ln q(x) dx \end{cases}$$

交叉熵可认为表示使用基于 Q 的编码对来自 P 的编码所需要的字节数, 也就是衡量 P, Q 两个概率分布的差异(但不是距离, 因为 $H(p, q) \neq H(q, p)$)。

④KL 散度(相对熵 Relative Entropy):

$$D_{KL}(p \| q) = \begin{cases} \sum_x p(x) \ln \frac{p(x)}{q(x)} \\ \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx \end{cases}$$

KL 散度可认为表示使用基于 Q 的编码对来自 P 的编码所需要的额外字节数。当处处满足 $p(x) = q(x)$ 时 KL 散度取得最小值 0, 因此 $H(P) \leq H(P, Q)$ 。

KL 散度非负, 不满足对称性 ($D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$)。

多维正态分布: $D_{KL}(p_1 \| p_2) = \frac{1}{2} \left(\ln \frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) \right)$, 式中两

个 d 维正态分布满足: $p_i = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)\right)}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}}$ ($i=1, 2$)。

当第一个正态分布各个分量独立 (Σ 是对角矩阵), 第二个正态分布是 $N(\mathbf{0}, \mathbf{I})$, 有

$$D_{KL}\left(N\left((\mu_1, \dots, \mu_d)^T, \text{diag}(\sigma_1^2, \dots, \sigma_d^2)\right) \| N(\mathbf{0}, \mathbf{I})\right) = \frac{1}{2} \sum_{i=1}^d (\sigma_i^2 + \mu_i^2 - \ln \sigma_i^2 - 1)。$$

$$d=2 \text{ 时 } D_{KL}(p_1 \| p_2) = \frac{1}{2} \left(\ln \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2} - 1 \right)。$$

KL 散度与交叉熵的关系:

$$D_{KL}(P \| Q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx = \int_{-\infty}^{+\infty} p(x) \ln p(x) dx - \int_{-\infty}^{+\infty} p(x) \ln q(x) dx = -H(P) + H(P, Q)$$

通常以 $p(x)$ 为目标去拟合出一个 $q(x)$ 来近似 $p(x)$, 因为 $H(p)$ 一般不变, 因此在优化时可以直接使用交叉熵 $H(P, Q)$ 来作为优化目标。

KL 散度的应用:

随机近邻嵌入 SNE(Stochastic Neighbor Embedding): 将高维向量组 \mathbf{x}_i 投影到低维 \mathbf{y}_i ,

将 $\mathbf{x}_i, \mathbf{x}_j$ 的距离用概率来刻画, $p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)}$, 式中 σ_i 是以 \mathbf{x}_i 为中心的正态

分布的标准差, $p_{i|i} = 0$ 。投影到低维后 $q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)}$, 式中标准差统一设置为 $\frac{1}{\sqrt{2}}$ 。

所有样本点成为 \mathbf{x}_i 的邻居的概率记为 p_i (离散型概率分布), 投影到低维后记为 q_i , 目标就是让 p_i, q_i 尽可能接近。定义损失函数 $L(\mathbf{y}_i) = \sum_{i=1}^n D_{KL}(p_i | q_i) = \sum_{i=1}^n \sum_{j=1}^n p_{j|i} \ln \frac{p_{j|i}}{q_{j|i}}$, 式中 n 为样本数。

注: 通过控制标准差来调整 \mathbf{x}_i 与其邻近点之间的距离关系, 较大的 σ 意味着高斯分布更加宽广, 其曲线更加平缓, 这使得在 \mathbf{x}_i 的周围范围内的点都有较高的概率成为邻近点。这和 RBF 有异曲同工之妙, 较小的 σ 在 RBF 里就是将数据之间变得更加不“亲近”, 也就是更能分割, 在 SNE 里就是成为临近点的概率下降, 只有更临近的点才更可能成为临近点。

变分推断 Variational Inference(变分贝叶斯): 对于贝叶斯推断 $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} =$

$\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) d\mathbf{z}} = \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{\int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) d\mathbf{z}}$, 式中 \mathbf{x}, \mathbf{z} 分别是可见变量和隐变量。当 \mathbf{z} 高维时, 分母不便计

算。考虑构造 $p(\mathbf{z}|\mathbf{x}) = q(\mathbf{z})$, 优化目标是 $\min_q D_{KL}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x})) = \min \int_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x})} d\mathbf{z} + \ln p(\mathbf{x})$ 。

因为 $\ln p(\mathbf{x})$ 是定值, 也就是 $\min_q \int_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x})} d\mathbf{z}$, 记证据下界 $L(q(\mathbf{z})) = -\int_{\mathbf{z}} q(\mathbf{z}) \ln \frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x})} d\mathbf{z} = -E_{q(\mathbf{z})}[\ln p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}) \| p(\mathbf{z}))$ 。目标就是 $\max_q L(q(\mathbf{z}))$, 可以限定 $q(\mathbf{z})$ 是正态分布等。

⑤JS 散度 Jensen-Shannon:

$D_{JS}(p \| q) = \frac{1}{2} D_{KL}(p \| m) + \frac{1}{2} D_{KL}(q \| m)$, 其中 $m(x) = \frac{p(x) + q(x)}{2}$ 。

JS 散度非负且满足对称性。同样是衡量两个概率分布之间的差异。

比如 GAN 网络的目标是最小化生成样本的概率分布与真实样本的概率分布的 JS 散度。

⑥互信息 Mutual Information:

$$I(X, Y) = \begin{cases} \sum_x \sum_y p(x, y) \frac{\ln p(x, y)}{p(x)p(y)} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dx dy \end{cases}$$

互信息非负且满足对称性。衡量的是两个随机变量的依赖程度(联合概率 $p(x, y)$ 与边缘概率之积 $p(x)p(y)$ 的差异程度)。

互信息与熵的关系是 $H(X, Y) = H(X) + H(Y) - I(X, Y)$, $I(X, Y) \leq H(X)$, $I(X, Y) \leq H(Y)$, 类似于集合里 $A \cup B = A + B - A \cap B$ 。

互信息可用于**特征选择**, X 是某个特征, Y 是类别标签值, 互信息越大, 特征越有用。

⑦条件熵 Conditional Entropy:

$$\begin{aligned} H(Y|X) &= \sum_{x \in X} p(x) H(Y|X=x) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \ln p(y|x) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \ln p(y|x) \\ &= \begin{cases} -\sum_{x \in X} \sum_{y \in Y} p(x, y) \ln \frac{p(x, y)}{p(x)} \\ -\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(x, y) \ln \frac{p(x, y)}{p(x)} dx dy \end{cases} \end{aligned}$$

条件熵非负。当且仅当 X, Y 相互独立时 $H(Y|X) = H(Y)$ 。

条件熵与熵的关系 是 $H(Y|X) = H(X, Y) - H(X)$ ，类似于集合里 $B - A = A \cup B - A$ 。

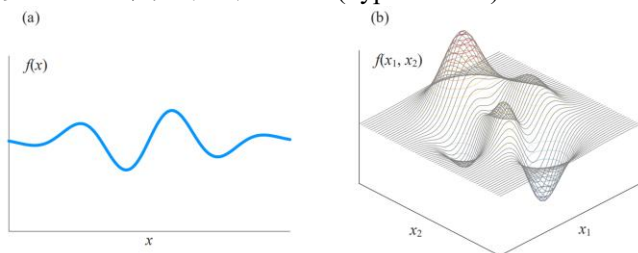
四、优化理论

4-0. 约定写法

(1) 最小化优化问题可以写为： $\arg \min_x f(x)$ 。

$\arg \min$ 即 argument of the minima(最小值的参数)。 x 是自变量(多变量一般写成列向量)。 $f(x)$ 为目标函数(objective function), 可以是函数解析式, 也可以是无法用解析式表达的模型。

当只有一/两个优化变量 x 时, 目标函数类似如下 a, b 图。当优化变量增多时, $f(x)$ 在多维空间内形成一个超曲面(hypersurface)。



最大化优化问题取负转为最小化优化问题。

(2) 约束条件:

变量的取值范围叫做定义域(domain)、搜索空间(search space)、选择集(choice set)。

范围内的每一个点为一个潜在解(candidate solution) 或可行解(feasible solution)。

优化变量取值范围的条件被称作约束条件(constraints)。根据约束条件的有无, 优化问题分为无约束优化问题(unconstrained optimization)和受约束优化问题(constrained optimization)。

五类约束条件:

上下界(lower and upper bounds): $l \leq x \leq u$

线性不等式(linear inequalities): $g(x) = Ax - b \leq 0$

线性等式(linear equalities): $h(x) = Ax - b = 0$

非线性不等式(nonlinear inequalities): $c(x) \leq 0$

非线性等式(nonlinear equalities): $c(x) = 0$

因此, 最终的最小化优化问题可以写为:

$$\begin{aligned} & \arg \min_x f(x) \\ & \text{s.t. } l \leq x \leq u \\ & \quad g(x) = Ax - b \leq 0 \\ & \quad h(x) = Ax - b = 0 \\ & \quad c(x) \leq 0 \\ & \quad c(x) = 0 \end{aligned}$$

其中, subject to 代表受限于、约束于、取决于、依赖于, 简写成 s.t.。

约束条件的表示方法:

比如 $\begin{cases} -2 \leq x_1 \leq 1 \\ -1 \leq x_2 \leq 1 \end{cases}$ 可写为 $\begin{bmatrix} -2 \\ -1 \end{bmatrix} \leq x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\begin{cases} x_1 - 0.5x_2 \geq -1 \\ x_1 + 2x_2 \geq 1 \\ x_1 + x_2 \leq 2 \end{cases}$ 可写为 $\begin{bmatrix} -1 & 0.5 \\ -1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \leq \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ 。

对于等式 $Ax - b = 0$ 可以化为两个不等式 $Ax - b \leq 0$ 且 $Ax - b \geq 0$ 。

对于没有显式约束的变量 x_i , 可以化为 $x_i = \bar{x}_i - \underline{x}_i$, 其中 $\bar{x}_i, \underline{x}_i \in (-\infty, +\infty)$ 。

常常有这样的二次规划形式:

$$\begin{aligned} & \arg \min_x \left(\frac{1}{2} x^T Q x + c^T x \right) \\ & \text{s.t. } Ax \leq b \end{aligned}$$

其中 $x \in \mathbb{R}^n$, Q 是 $n \times n$ 的二次项系数矩阵, $c \in \mathbb{R}^n$ 是一次项系数向量。 $A_{m \times n}$ 是约束系

数矩阵, $b \in \mathbb{R}^m$ 是不等式约束的常数向量。

4-1. 凸 convex

1. 凸集 Convex Set

对 n 维空间的点集 C , 对 C 中任意两点 x, y 与 $0 \leq \theta \leq 1$ 均有 $\theta x + (1-\theta)y \in C$ 。

直观上就是任意两点连线, 线段上的所有点都属于集合 C 。

由 $Ax \leq b$ 约束的可行域是凸集, 因为 $A(\theta x + (1-\theta)y) = \theta Ax + (1-\theta)Ay \leq \theta b + (1-\theta)b \leq b$ 。

因此实际问题(见 4-0 的五种约束条件)的可行域通常是凸集。

2. 凸函数

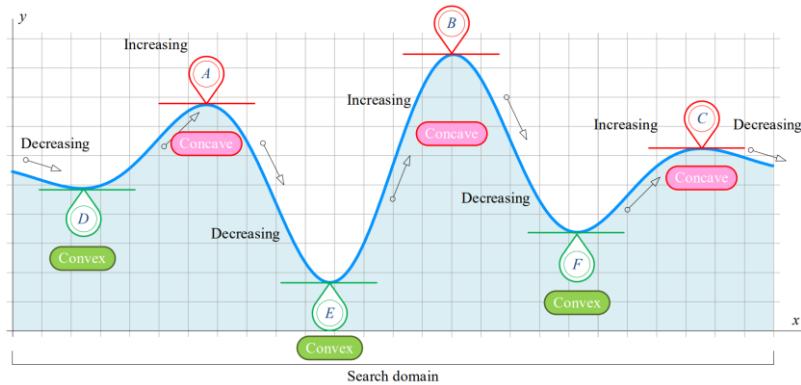
设 $\nabla^2 f(x)$ 是 $f(x)$ 在 x 处的二阶导数矩阵或者 Hessian 矩阵, 则

(1) f 是凸集 S 上的凸函数 $\Leftrightarrow \forall x \in S, \nabla^2 f(x)$ 半正定

(2) $\forall x \in S, \nabla^2 f(x)$ 正定 $\Rightarrow f$ 是凸集 S 上的严格凸函数

(3) 凸函数的线性组合还是凸函数

如下图, D、E、F 三点区域函数都是局部为凸(convex), A、B、C 所在的局域函数都是局部为凹(concave)。凸优化(convex optimization)正是研究定义于凸集中的凸函数最小化的问题。凸, 即下凸, $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ 或 $f''(x) \geq 0$ 都是凸函数的充要条件。



(1) 典型的凸函数

① log-sum-exp 函数 $f(\vec{x}) = \log(\sum_{i=1}^n e^{x_i})$

② 线性函数、指数函数、负熵、范数等

(2) 凸优化问题

凸优化问题(Convex optimization problem)要求目标函数为凸函数, 而且定义域为凸集, 即要求目标函数 $f_0(x)$ 和约束函数 $f_i(x) (i=1, \dots, m)$ 均为凸函数。

对凸优化问题, 局部最优就是全局最优。

4-2. 多元函数的极大/小值问题

1. 无约束的优化问题

驻点条件: $\nabla f(x) = 0$, 如果 x^* 是驻点, 然后:

考虑 Hessian 矩阵 (对梯度向量的每个分量再次求梯度, 一般是对称矩阵):

$$H(x) = \nabla^2 f(x) = \nabla(\nabla f(x)) = \begin{bmatrix} \nabla\left(\frac{\partial f}{\partial x_1}\right) \\ \nabla\left(\frac{\partial f}{\partial x_2}\right) \\ \vdots \\ \nabla\left(\frac{\partial f}{\partial x_n}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{bmatrix}$$

如果 $H(x^*)$ 是正定的, 则 x^* 是极小值点;

如果 $H(x^*)$ 是负定的, 则 x^* 是极大值点;

如果 $H(x^*)$ 是不定的, 则 x^* 是鞍点, 即在某个方向上有极小值, 在另一方向上有极大值。

如, 求 $f(x, y) = 3x^2 + 2y^3 - 2xy$ 的驻点。

解: $\nabla f = \begin{bmatrix} 6x - 2y \\ 6y^2 - 2x \end{bmatrix} = \vec{0} \Rightarrow \begin{cases} x^* = 0 \\ y^* = 0 \end{cases} \text{ OR } \begin{cases} x^* = \frac{1}{27} \\ y^* = \frac{1}{9} \end{cases}$ 。Hessian 矩阵为 $H(x, y) = \begin{bmatrix} 6 & -2 \\ -2 & 12y \end{bmatrix}$, 在

$x^* = y^* = 0$ 处是不定的, 因此为鞍点; 在 $x^* = \frac{1}{27}, y^* = \frac{1}{9}$ 处是正定的, 因此 $(\frac{1}{27}, \frac{1}{9})$ 是一个极小值点, 极小值为 -0.0014 。

2. 有约束的优化问题

有约束的情况下, ①可能取不到原来的全局最优解。②最优解可能在约束边界上。

一元函数极值点判定:

寻找极值时注意三类点: ①驻点(一阶导数为 0 的点)。②不可导点。③搜索区域边界点。

二元函数极值点判定:

$f(x_1, x_2)$ 在 (a, b) 邻域内连续且一阶偏导及二阶偏导连续, 满足 $f_1(a, b) = 0, f_2(a, b) = 0$, 令

$A = f_{11}(a, b), B = f_{12}(a, b), C = f_{22}(a, b), \Delta = AC - B^2$:

① $\Delta > 0$: 存在极值。 $A > 0$ 极小值, $A < 0$ 极大值。

② $\Delta < 0$: 不存在极值。

③ $\Delta = 0$: 可能存在也可能不存在, 需要进一步讨论。

其实 Δ 就是黑塞矩阵 $\begin{bmatrix} A & B \\ B & C \end{bmatrix}$ 的 2 阶顺序主子式 $\begin{vmatrix} A & B \\ B & C \end{vmatrix}$, $|A|$ 就是 1 阶顺序主子式。

可视化具体代码可见 [Visualize-ML-Code/Book3_Elements-of-Mathematics-main/Book3_Ch19_Python_Codes](#)

(1) 拉格朗日乘数法-只含等式约束的优化问题

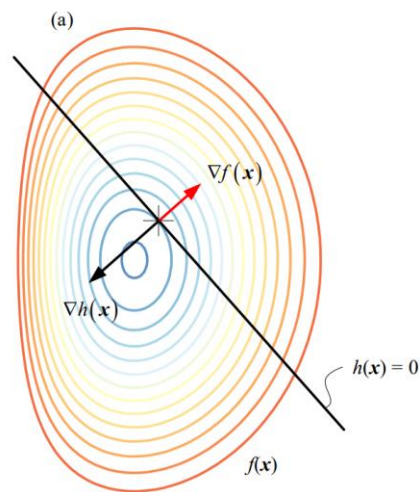
设优化目标 $\min_x f(x)$, 约束条件 $g_i(x) = 0$ 。

构造拉格朗日函数 $L(x, \lambda) = f(x) + \sum_{i=1}^n \lambda_i g_i(x)$, 有

$$\begin{cases} \nabla_x L(x, \lambda) = \nabla_x f(x) + \sum_{i=1}^n \lambda_i \nabla_x g_i(x) = \mathbf{0} \\ \nabla_\lambda L(x, \lambda) = g_i(x) = 0 \end{cases}, \text{ 这是最优解的}$$

必要条件, 求得的不一定是驻点(可能是鞍点)。

偏导中的 $g_i(x) = 0$ 很好理解, 对另一个, 我们简化为只有一个约束条件的情况 $\nabla_x f(x) + \lambda \nabla_x g(x) = \mathbf{0}$, 如图(图中的 $h(x)$ 是这里的 $g(x)$, 等高线代表数值(暖色对应的数值更大)): 如果 $h(x), f(x)$ 只是简单的相交, 此时交点不是极值。只有在相切且非驻点处, 才能取到极值。



(2) 拉格朗日对偶性-原问题转化为对偶问题

对于优化问题:

$$\begin{aligned} & \min_x f(x) \\ & \text{s.t. } g(x) \leq 0 \\ & \quad h(x) = 0 \end{aligned}$$

构造广义拉格朗日乘子函数 $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^q \nu_i h_i(\mathbf{x})$, 其中 $\lambda_i \geq 0$ 。记原问题的最优解是 $p^* = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}, \lambda_i \geq 0} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, 其对偶问题的最优值为 $d^* = \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}, \lambda_i \geq 0} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ 。二者区别是：先最小化拉格朗日乘子还是先最大化 \mathbf{x} 。设原/对偶问题的最优解是 $\mathbf{x}_1, \boldsymbol{\lambda}_1, \boldsymbol{\nu}_1$ 和 $\mathbf{x}_2, \boldsymbol{\lambda}_2, \boldsymbol{\nu}_2$, 则 $p^* = L(\mathbf{x}_1, \boldsymbol{\lambda}_1, \boldsymbol{\nu}_1) \geq L(\mathbf{x}_1, \boldsymbol{\lambda}_2, \boldsymbol{\nu}_2) \geq L(\mathbf{x}_2, \boldsymbol{\lambda}_2, \boldsymbol{\nu}_2) = d^*$ 。也就是**对偶问题的最优解不大于原问题的最优解**。

Slater 条件：一个凸优化问题如果存在一个 \mathbf{x} 使得所有不等式约束都是**严格满足**的(就是不取等, 也就是区域内部至少有一个可行解(非边界)), 那么 $p^* = d^*$ 。

对偶问题常常更简单, 还能提供问题的 \mathbf{x} 解界(对原始问题的最优性提供了一个界限)。

(3) KKT 条件-含不等式约束的优化问题

先只考虑 $g(\mathbf{x}) \leq 0$ 与 $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda} g(\mathbf{x})$:

如果 \mathbf{x} 在内部取到, 则 $g(\mathbf{x}) \leq 0$ 的约束失效, 问题退化为无约束的最优化问题, 最优解 \mathbf{x}^* 满足 $\nabla f(\mathbf{x}^*) = 0, \boldsymbol{\lambda} = 0$ 。

如果 \mathbf{x} 在边界取到, 则 $g(\mathbf{x}) \leq 0$ 的约束退化为 $g(\mathbf{x}) = 0$, 最优解满足 $\nabla f(\mathbf{x}^*) + \nabla \boldsymbol{\lambda} g(\mathbf{x}^*) = 0$ 。另外, 因为是最小化 f , 所以可行域 K 的内部的 f 更大, 那么 ∇f 指向 K 内部。因为可行域的内部满足 $g(\mathbf{x}) \leq 0$, 边界满足 $g(\mathbf{x}) = 0$, 说明 ∇g 指向 K 外部。为了让 $\nabla f(\mathbf{x}^*) + \nabla \boldsymbol{\lambda} g(\mathbf{x}^*) = 0$ 成立, 必然有 $\boldsymbol{\lambda} \geq 0$ 。

综上, KKT 条件是:

$$\begin{aligned} \nabla_{\mathbf{x}} f(\mathbf{x}) + \boldsymbol{\lambda} \nabla_{\mathbf{x}} g(\mathbf{x}) &= \mathbf{0} \\ g(\mathbf{x}) &\leq 0 \\ \boldsymbol{\lambda} &\geq 0 \\ \boldsymbol{\lambda} g(\mathbf{x}) &= 0 \end{aligned}$$

KKT 条件是取得极值的必要条件。凸优化问题时 KKT 才是充要条件。

(4) 小结

对于前文提及的有约束的优化问题:

$$\begin{aligned} \arg \min_{\mathbf{x}} f(\mathbf{x}) \\ \text{s.t. } \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \\ g(\mathbf{x}) \leq 0 \\ h(\mathbf{x}) = 0 \end{aligned}$$

根据拉格朗日和 KKT, 令 $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}_h h(\mathbf{x}) + \boldsymbol{\lambda}_g g(\mathbf{x})$, 有:

$$\begin{aligned} \mathbf{l} &\leq \mathbf{x} \leq \mathbf{u} \\ \nabla_{\mathbf{x}, \boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{0} \\ g(\mathbf{x}) \leq 0, \boldsymbol{\lambda}_g > 0, \boldsymbol{\lambda}_g g(\mathbf{x}) &= 0 \\ h(\mathbf{x}) &= 0 \end{aligned}$$

(5) 再谈二次型的优化

①对于这样的二次规划问题:

$$\begin{aligned} \arg \min_{\mathbf{x}} (\mathbf{x}^T Q \mathbf{x}) \quad \text{或} \quad \arg \max_{\mathbf{x}} (\mathbf{x}^T Q \mathbf{x}) \\ \text{s.t. } \mathbf{x}^T \mathbf{x} = 1 \end{aligned}$$

记 $L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T Q \mathbf{x} - \boldsymbol{\lambda} (\mathbf{x}^T \mathbf{x} - 1)$ (负号是为了更便于与特征值分解对应), 得 $2Q\mathbf{x} - 2\boldsymbol{\lambda} \mathbf{x} = \mathbf{0}$, 即 $Q\mathbf{x} = \boldsymbol{\lambda} \mathbf{x}$ 。因此, 最小化问题里取最小的特征值 $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\min}$, 最大化问题中取 $\boldsymbol{\lambda}_{\max}$ 。

②当约束条件变成 $\mathbf{x}^T P \mathbf{x} = 1$ 时, 同理有 $Q\mathbf{x} = \boldsymbol{\lambda} P \mathbf{x}$, 当 P 可逆时, 变为对 $P^{-1}Q$ 进行特征

值分解。

③当优化目标变为 $\arg \max_{\mathbf{x} \neq 0} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{B} \mathbf{x}}$ ，一般 \mathbf{B} 正定，使 $\mathbf{x}^T \mathbf{B} \mathbf{x} > 0$ 。令 $\mathbf{x} = \mathbf{B}^{-\frac{1}{2}} \mathbf{y}$ ，代入优化目标得 $\frac{\mathbf{y}^T \mathbf{B}^{-\frac{1}{2}T} \mathbf{A} \mathbf{B}^{-\frac{1}{2}} \mathbf{y}}{\mathbf{y}^T \mathbf{y}}$ ，注意到 $\frac{\mathbf{x}^T \mathbf{Q} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ 的形式就是瑞利商，它等价于问题①的 $\arg \max_{\mathbf{x}} (\mathbf{x}^T \mathbf{Q} \mathbf{x})$ 。
 $s.t. \quad \mathbf{x}^T \mathbf{x} = 1$

④对于 $\arg \min_{\mathbf{x}} \|\mathbf{Q} \mathbf{x}\|$ 显然等价于 $\arg \min_{\mathbf{x}} (\mathbf{x}^T \mathbf{Q}^T \mathbf{Q} \mathbf{x})$ 与 $\arg \min_{\mathbf{h} \neq 0} \left(\frac{\|\mathbf{Q} \mathbf{h}\|}{\|\mathbf{h}\|} \right)^2$, $\arg \min_{\mathbf{h} \neq 0} \frac{\|\mathbf{Q} \mathbf{h}\|}{\|\mathbf{h}\|}$ ，其中 \mathbf{h}
 $s.t. \quad \|\mathbf{x}\| = 1$ $s.t. \quad \mathbf{x}^T \mathbf{x} = 1$
 $= \frac{\mathbf{h}^T \mathbf{Q}^T \mathbf{Q} \mathbf{h}}{\mathbf{h}^T \mathbf{h}}$

是任意非 $\mathbf{0}$ 向量。

4-3.优化方法

1.一阶优化算法

(1)梯度下降法 Gradient Descent

泰勒展开 $f(\mathbf{x} + \Delta \mathbf{x}) = f(\mathbf{x}) + (\nabla f(\mathbf{x}))^T \Delta \mathbf{x} + o(\|\Delta \mathbf{x}\|)$ ，当 $\Delta \mathbf{x}$ 足够小时，只要 $(\nabla f(\mathbf{x}))^T \Delta \mathbf{x} \leq 0$ ，

就有 $f(\mathbf{x} + \Delta \mathbf{x}) \leq f(\mathbf{x})$ 。而 $(\nabla f(\mathbf{x}))^T \Delta \mathbf{x} = \|\nabla f(\mathbf{x})\| \|\Delta \mathbf{x}\| \cos \theta$ ，当 $\nabla f(\mathbf{x}), \Delta \mathbf{x}$ 的夹角 $\theta = \pi$ 时(也就是在梯度相反的方向上)，函数值下降最快。为了满足 $\Delta \mathbf{x}$ 足够小这一条件，因此设置学习率，所以迭代公式是 $\Delta \mathbf{x} = -\alpha \nabla f(\mathbf{x})$ 。伪代码如下：

```
for (  $\mathbf{x}_0, k=0$  ;  $\|\nabla f(\mathbf{x}_k)\| > eps$  &  $k < N$  ;  $k++$  )
     $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$ 
```

其中 \mathbf{x}_0 的值可以是 $\mathbf{0}$ 或随机值， eps 是很接近 $\mathbf{0}$ 的正数， N 为最大迭代次数。

(2)最速下降法 Steepest Descent

在梯度下降法中步长 $\Delta \mathbf{x} = -\alpha \nabla f(\mathbf{x})$ ，在最速下降法中步长 $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$ 。但我们这次要通过 f 来优化 α ，也就是 $\alpha_k = \arg \min_{\alpha} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ 。因此伪代码如下：

```
for (  $\mathbf{x}_0, k=0, \mathbf{d}_k = -\nabla f(\mathbf{x}_k)$  ;  $\|\nabla f(\mathbf{x}_k)\| > eps$  &  $k < N$  ;  $k++$  )
     $\alpha_k = \arg \min_{\alpha} f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ 
     $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ 
```

(3)梯度下降法的优化版本

解决振荡问题：

引入动量项 $\mathbf{v}_k = -\alpha \nabla f(\mathbf{x}_k) + \mu \mathbf{v}_{k-1}$ ，梯度下降更新公式是 $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_k$ ，展开 \mathbf{v}_k 可得

$\mathbf{v}_k = -\alpha \nabla f(\mathbf{x}_k) - \alpha \mu \nabla f(\mathbf{x}_{k-1}) - \alpha \mu^2 \nabla f(\mathbf{x}_{k-2}) - \dots$ ，说明每次更新梯度的时候记录了之前的负梯度值，依靠惯性保持迭代的前进方向，且负梯度值按系数 μ 指数级衰减。

自适应学习率：

AdaGrad 算法： 设第 k 次迭代的梯度向量是 \mathbf{g}_k ，则 $(\mathbf{x}_{k+1})_i = (\mathbf{x}_k)_i - \alpha \frac{(\mathbf{g}_k)_i}{\sqrt{\sum_{j=1}^k ((\mathbf{g}_j)_i)^2 + \varepsilon}}$ 。

式中 $(\cdot)_i$ 是向量的分量下标, ε 是一个接近0的正数(防止除0)。分母是累积到本次迭代的梯度历史信息, 历史梯度越大会让学习率越小(但这也是问题, 最终分母会越来越大让学习率趋向0), 每一维都有自己的学习率。

RMSProp 算法: 定义 $E[\mathbf{g}^2]_k = \delta E[\mathbf{g}^2]_{k-1} + (1-\delta)\mathbf{g}_k^2$, $(\mathbf{x}_{k+1})_i = (\mathbf{x}_k)_i - \alpha \frac{(\mathbf{g}_k)_i}{\sqrt{(E[\mathbf{g}^2]_k)_i + \varepsilon}}$ 。

式中 \mathbf{g}^2 是对梯度向量的每个分量分别平方, 得到梯度平方累加值的向量 $E[\mathbf{g}^2]$, $0 < \delta < 1$ 是手动设置的衰减系数。该算法避免了随着时间的累积, 学习率趋向0的问题。

AdaDelta 算法: 定义 $\text{RMS}[\mathbf{g}]_k = \sqrt{E[\mathbf{g}^2]_k + \varepsilon}$, $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\text{RMS}[\Delta \mathbf{x}]_{k-1}}{\text{RMS}[\mathbf{g}]_k} \mathbf{g}_k$ 。式中 $\text{RMS}[\Delta \mathbf{x}]$ 与 $\text{RMS}[\mathbf{g}]$ 的计算方法一致。该算法不需要再手动设置学习率。

结合:

Adam 算法: 定义 $\begin{cases} (\mathbf{m}_k)_i = \beta_1(\mathbf{m}_{k-1})_i + (1-\beta_1)(\mathbf{g}_k)_i \\ (\mathbf{v}_k)_i = \beta_2(\mathbf{v}_{k-1})_i + (1-\beta_2)(\mathbf{g}_k)_i^2 \end{cases}$, $(\mathbf{x}_{k+1})_i = (\mathbf{x}_k)_i - \alpha \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k} \frac{(\mathbf{m}_k)_i}{\sqrt{(\mathbf{v}_k)_i + \varepsilon}}$ 。

式中 \mathbf{m}, \mathbf{v} 相当于动量项和学习率, β_1, β_2 是手动设置的参数。

(4) 随机梯度下降法

批量梯度下降: 损失函数 $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}, \text{data}_i)$, 梯度 $\nabla f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \nabla f(\mathbf{x}, \text{data}_i)$ 。

小批量梯度下降: 损失函数 $f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f(\mathbf{x}, \text{data}_i)$, 梯度 $\nabla f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \nabla f(\mathbf{x}, \text{data}_i)$ 。

随机梯度下降: 损失函数 $f(\mathbf{x}) = f(\mathbf{x}, \text{data}_i)$, 梯度 $\nabla f(\mathbf{x}) = \nabla f(\mathbf{x}, \text{data}_i)$ 。

小批量就是从 n 个样本里随机选 m 个, 随机就是只选1个。

2. 二阶优化算法

(1) 牛顿法 Newton Method

泰勒展开 $f(\mathbf{x}) = f(\mathbf{x}_0) + (\nabla f(\mathbf{x}_0))^T (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2)$, 忽略高阶项并对 \mathbf{x} 求梯度得 $\nabla f(\mathbf{x}) \approx \nabla f(\mathbf{x}_0) + \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ 。因此 $\mathbf{x} - \mathbf{x}_0 = \frac{\nabla f(\mathbf{x}_0)}{\nabla^2 f(\mathbf{x}_0)}$, 将梯度向

量记为 $\mathbf{g} = \nabla f(\mathbf{x})$, 黑塞矩阵记为 H , 则 $\mathbf{x} = \mathbf{x}_0 - H^{-1} \mathbf{g}$ 。也就是 $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha H_k^{-1} \mathbf{g}_k$ 。 $-H^{-1} \mathbf{g}$

称为牛顿方向。伪代码如下:

```
for (  $\mathbf{x}_0, k=0$ ;  $\|\mathbf{g}_k\| > \text{eps} \ \& \ k < N$ ;  $k++$  )
```

```
     $\mathbf{d}_k = -H_k^{-1} \mathbf{g}_k$ 
```

```
     $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha \mathbf{d}_k$ 
```

实际常为求解 $H_k \mathbf{d}_k = -\mathbf{g}_k$ 来代替求解 H_k^{-1} 。

(2) 拟牛顿法 Quasi-Newton Method

对 $\nabla f(\mathbf{x}) \approx \nabla f(\mathbf{x}_0) + \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ 赋值 $\mathbf{x} = \mathbf{x}_{k+1}, \mathbf{x}_0 = \mathbf{x}_k$ 可得 $\mathbf{g}_{k+1} - \mathbf{g}_k = H_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_k)$,

令 $\begin{cases} s_k = x_{k+1} - x_k \\ y_k = g_{k+1} - g_k \end{cases}$, 则 $s_k = H_{k+1}^{-1} y_k$, 说明迭代点与其梯度值就能计算 H^{-1} 。

DFP 算法:

for ($x_0, k=0, H_0=I$; $\|g_{k+1}\| > eps \ \& \ k < N$; $k++$)
 $d_k = -H_k g_k$, 用直线搜索得到步长 λ_k , $s_k = \lambda_k d_k, x_{k+1} = x_k + s_k$
 $y_k = g_{k+1} - g_k$, $H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k}$

BFGS 算法:

for ($x_0, k=0, B_0=I$; $\|g_{k+1}\| > eps \ \& \ k < N$; $k++$)
 $d_k = -B_k^{-1} g_k$, 用直线搜索得到步长 λ_k , $s_k = \lambda_k d_k, x_{k+1} = x_k + s_k$
 $y_k = g_{k+1} - g_k$, $B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$

3.分治法

(1)坐标下降法 Coordinate Descent

依次选择 $x_1 \sim x_n$, 求解单个变量的优化问题, 也就是 $\min_x f(x), x = (x_1, \dots, x_n)$ 。

该方法容易失效, 因为单独改变某个变量的值可能均不能使目标函数值下降。比如

$f = |x+y| + 3|y-x|$ 在 $(-2, -2)$ 处。

(2)SMO 算法

每次选择两个分量进行优化。

(3)分阶段优化

AdaBoost, 先训练弱分类器, 然后确定弱分类器的权重系数。

五、图论

1. 谱聚类算法

输入：样本集 $G=(x_1, x_2, \dots, x_n)$ ，相似矩阵的生成方式，降维后的维度 k_1 ，聚类方法，聚类后的维度 k_2

输出：簇划分 $C=(c_1, c_2, \dots, c_{k_2})$

过程：①根据输入的相似矩阵的生成方式构建样本集的相似矩阵 S

②根据 S 构建邻接矩阵 W ，构建度矩阵 A

③计算拉普拉斯矩阵 $L=D-A$

④构建标准化的拉普拉斯矩阵 $D^{-\frac{1}{2}}LD^{\frac{1}{2}}$

⑤计算 $D^{-\frac{1}{2}}LD^{\frac{1}{2}}$ 最小的 k_1 个特征值对应的特征向量 \vec{f}

⑥将各自对应的 \vec{f} 组成的矩阵标准化，最终组成 $n \times k_1$ 维的特征矩阵 F

⑦对 F 中的每一行作为 k_1 维的样本，共 n 个，用输入的聚类算法进行聚类，聚类维数为 k_2

⑧得到簇划分 $C=(c_1, c_2, \dots, c_{k_2})$

【例】以下图为例简单介绍谱聚类实现。

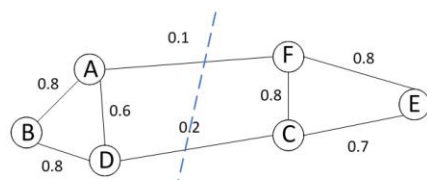


图 2-35 图G的网络结构

$$\text{相似矩阵} \begin{bmatrix} 0 & 0.8 & 0.6 & 0 & 0.1 & 0 \\ 0.8 & 0 & 0.8 & 0 & 0 & 0 \\ 0.6 & 0.8 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0.8 & 0.7 \\ 0.1 & 0 & 0 & 0.8 & 0 & 0.8 \\ 0 & 0 & 0 & 0.7 & 0.8 & 0 \end{bmatrix}, \text{计算度矩阵} D = \begin{bmatrix} 1.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.5 \end{bmatrix}$$

求G的拉普拉斯矩阵为：

$$L = D - A = \begin{bmatrix} 1.5 & -0.8 & -0.6 & 0 & -0.1 & 0 \\ -0.8 & 1.6 & -0.8 & 0 & 0 & 0 \\ -0.6 & -0.8 & 1.6 & -0.2 & 0 & 0 \\ 0 & 0 & -0.2 & 1.7 & -0.8 & -0.7 \\ -0.1 & 0 & 0 & -0.8 & 1.7 & -0.8 \\ 0 & 0 & 0 & -0.7 & -0.8 & 1.5 \end{bmatrix}$$

对其进行标准化：

$$\hat{L} = D^{-\frac{1}{2}}LD^{\frac{1}{2}} = \begin{bmatrix} 1 & -0.52 & -0.39 & 0 & -0.06 & 0 \\ -0.52 & 1 & -0.5 & 0 & 0 & 0 \\ -0.39 & -0.5 & 1 & -0.12 & 0 & 0 \\ 0 & 0 & -0.12 & 1 & -0.47 & -0.44 \\ -0.06 & 0 & 0 & -0.47 & 1 & -0.5 \\ 0 & 0 & 0 & -0.44 & -0.5 & 1 \end{bmatrix}$$

对标准化的矩阵求特征值和特征向量，求得特征向量矩阵：

$$\begin{bmatrix} 0.4026 & -0.3963 & -0.5191 & -0.3751 & -0.3452 & -0.3892 \\ 0.4163 & -0.4434 & -0.0973 & 0.2464 & 0.0301 & 0.7476 \\ 0.4146 & -0.3729 & 0.6023 & 0.1613 & 0.3276 & -0.4399 \\ 0.4142 & 0.3849 & 0.4810 & -0.4317 & -0.4622 & 0.2211 \\ 0.4127 & 0.4193 & -0.2943 & -0.2813 & 0.6972 & 0.0452 \\ 0.3883 & 0.4282 & -0.2009 & 0.7120 & -0.2711 & -0.2122 \end{bmatrix}$$

同时求得对应的特征值为：-0.0008，0.1148，1.3210，1.4643，1.5357，1.5650。

之后可以选择 $K - Means$ 等聚类方法对求得特征向量进行聚类得到簇划分，后续步骤限于篇幅过程暂不展开。

若将得到的结果分为两类，则左边三个点为一类，右边三个点为一类，这也很符合直观的观察结果，左边的三个点之间有较为紧密的联系，右边的三个点间也有较为紧密的联系。