$$\left(\frac{\partial x}{\partial \mathbf{A}}\right)_{ij} = \frac{\partial x}{\partial A_{ij}} . \tag{A.19}$$

对于函数 $f(\boldsymbol{x})$, 假定其对向量的元素可导, 则 $f(\boldsymbol{x})$ 关于 $\boldsymbol{x}$ 的一阶导数是一个向量, 其第 i 个分量为

$$\left(\nabla f\left(\boldsymbol{x}\right)\right)_i = \frac{\partial f(\boldsymbol{x})}{\partial x_i} , \tag{A.20}$$

$f(\boldsymbol{x})$ 关于 $\boldsymbol{x}$ 的二阶导数是称为海森矩阵(Hessian matrix)的一个方阵, 其第 i 行第 j 列上的元素为

$$\left(\nabla^2 f\left(\boldsymbol{x}\right)\right)_{ij} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} . \tag{A.21}$$

向量和矩阵的导数满足乘法法则(product rule)

$\boldsymbol{a}$ 相对于 $\boldsymbol{x}$ 为常向量.

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a} , \tag{A.22}$$

$$\frac{\partial \mathbf{A}\mathbf{B}}{\partial \boldsymbol{x}} = \frac{\partial \mathbf{A}}{\partial \boldsymbol{x}}\mathbf{B} + \mathbf{A}\frac{\partial \mathbf{B}}{\partial \boldsymbol{x}} . \tag{A.23}$$

由 $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ 和 式(A.23), 逆矩阵的导数可表示为

$$\frac{\partial \mathbf{A}^{-1}}{\partial x} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1} . \tag{A.24}$$

若求导的标量是矩阵 $\mathbf{A}$ 的元素, 则有

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{B})}{\partial A_{ij}} = B_{ji} , \tag{A.25}$$

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B}^{\mathrm{T}} . \tag{A.26}$$

进而有

$$\frac{\partial \operatorname{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B} , \tag{A.27}$$

$$\frac{\partial \operatorname{tr}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{I} , \tag{A.28}$$

$$\frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}})}{\partial \mathbf{A}} = \mathbf{A}(\mathbf{B} + \mathbf{B}^{\mathrm{T}}) . \tag{A.29}$$

由式(A.15)和(A.29)有

$$\frac{\partial \|\mathbf{A}\|_F^2}{\partial \mathbf{A}} = \frac{\partial \operatorname{tr}(\mathbf{A}\mathbf{A}^{\mathrm{T}})}{\partial \mathbf{A}} = 2\mathbf{A} \ . \tag{A.30}$$

链式法则(chain rule)是计算复杂导数时的重要工具. 简单地说, 若函数 f 是 g 和 h 的复合, 即 $f(x) = g\left(h\left(x\right)\right)$, 则有

$$\frac{\partial f(x)}{\partial x} = \frac{\partial g\left(h\left(x\right)\right)}{\partial h(x)} \cdot \frac{\partial h(x)}{\partial x} \ . \tag{A.31}$$

例如在计算下式时, 将 $\mathbf{A}\boldsymbol{x} - \boldsymbol{b}$ 看作一个整体可简化计算:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{x}}(\mathbf{A}\boldsymbol{x} - \boldsymbol{b})^{\mathrm{T}}\mathbf{W}(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) &= \frac{\partial(\mathbf{A}\boldsymbol{x} - \boldsymbol{b})}{\partial \boldsymbol{x}} \cdot 2\mathbf{W}(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) \\ &= 2\mathbf{A}\mathbf{W}(\mathbf{A}\boldsymbol{x} - \boldsymbol{b}) \ . \end{aligned} \tag{A.32}$$

A.3 奇异值分解

任意实矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$ 都可分解为

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} \ , \tag{A.33}$$

其中, $\mathbf{U} \in \mathbb{R}^{m\times m}$ 是满足 $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$ 的 m 阶酉矩阵(unitary matrix); $\mathbf{V} \in \mathbb{R}^{n\times n}$ 是满足 $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$ 的 n 阶酉矩阵; $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ 是 $m \times n$ 的矩阵, 其中 $(\mathbf{\Sigma})_{ii} = \sigma_i$ 且其他位置的元素均为 0, σ_i 为非负实数且满足 $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant 0$.

常将奇异值按降序排列以确保 $\mathbf{\Sigma}$ 的唯一性.

式(A.33)中的分解称为奇异值分解(Singular Value Decomposition, 简称 SVD), 其中 $\mathbf{U}$ 的列向量 $\boldsymbol{u}_i \in \mathbb{R}^m$ 称为 $\mathbf{A}$ 的左奇异向量(left-singular vector), $\mathbf{V}$ 的列向量 $\boldsymbol{v}_i \in \mathbb{R}^n$ 称为 $\mathbf{A}$ 的右奇异向量(right-singular vector), σ_i 称为奇异值(singular value). 矩阵 $\mathbf{A}$ 的秩(rank)就等于非零奇异值的个数.

当 $\mathbf{A}$ 为对称正定矩阵时, 奇异值分解与特征值分解结果相同.

奇异值分解有广泛的用途, 例如对于低秩矩阵近似(low-rank matrix approximation)问题, 给定一个秩为 r 的矩阵 $\mathbf{A}$, 欲求其最优 k 秩近似矩阵 $\widetilde{\mathbf{A}}$, $k \leqslant r$, 该问题可形式化为

$$\begin{aligned} \min_{\widetilde{\mathbf{A}}\in\mathbb{R}^{m\times n}} \quad & \|\mathbf{A} - \widetilde{\mathbf{A}}\|_F \\ \text{s.t.} \quad & \operatorname{rank}(\widetilde{\mathbf{A}}) = k \ . \end{aligned} \tag{A.34}$$

奇异值分解提供了上述问题的解析解: 对矩阵 $\mathbf{A}$ 进行奇异值分解后, 将矩阵 $\mathbf{\Sigma}$ 中的 $r-k$ 个最小的奇异值置零获得矩阵 $\mathbf{\Sigma}_k$, 即仅保留最大的 k 个奇异值, 则

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^{\mathrm{T}} \tag{A.35}$$

就是式(A.34)的最优解, 其中 $\mathbf{U}_k$ 和 $\mathbf{V}_k$ 分别是式(A.33)中的前 k 列组成的矩阵. 这个结果称为 Eckart-Young-Mirsky 定理.

B 优化

B.1 拉格朗日乘子法

拉格朗日乘子法(Lagrange multipliers)是一种寻找多元函数在一组约束下的极值的方法. 通过引入拉格朗日乘子, 可将有 d 个变量与 k 个约束条件的最优化问题转化为具有 $d+k$ 个变量的无约束优化问题求解.

先考虑一个等式约束的优化问题. 假定 $\boldsymbol{x}$ 为 d 维向量, 欲寻找 $\boldsymbol{x}$ 的某个取值 $\boldsymbol{x}^*$, 使目标函数 $f(\boldsymbol{x})$ 最小且同时满足 $g(\boldsymbol{x})=0$ 的约束. 从几何角度看, 该问题的目标是在由方程 $g(\boldsymbol{x})=0$ 确定的 $d-1$ 维曲面上寻找能使目标函数 $f(\boldsymbol{x})$ 最小化的点. 此时不难得到如下结论:

函数等值线与约束曲面相切.

可通过反证法证明: 若梯度 $\nabla f(\boldsymbol{x}^*)$ 与约束曲面不正交, 则仍可在约束曲面上移动该点使函数值进一步下降.

- 对于约束曲面上的任意点 $\boldsymbol{x}$, 该点的梯度 $\nabla g(\boldsymbol{x})$ 正交于约束曲面;
- 在最优点 $\boldsymbol{x}^*$, 目标函数在该点的梯度 $\nabla f(\boldsymbol{x}^*)$ 正交于约束曲面.

由此可知, 在最优点 $\boldsymbol{x}^*$, 如附图B.1 所示, 梯度 $\nabla g(\boldsymbol{x})$ 和 $\nabla f(\boldsymbol{x})$ 的方向必相同或相反, 即存在 $\lambda \neq 0$ 使得

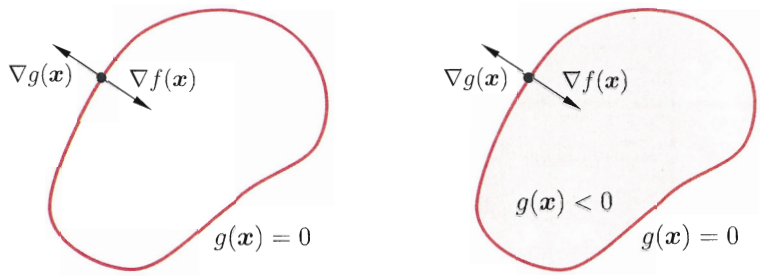
$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0 \,, \tag{B.1}$$

对等式约束, λ 可能为正也可能为负.

λ 称为拉格朗日乘子. 定义拉格朗日函数

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}) \,, \tag{B.2}$$

不难发现, 将其对 $\boldsymbol{x}$ 的偏导数 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda)$ 置零即得式(B.1), 同时, 将其对 λ 的偏导数 $\nabla_{\lambda} L(\boldsymbol{x}, \lambda)$ 置零即得约束条件 $g(\boldsymbol{x})=0$. 于是, 原约束优化问题可转化为对拉格朗日函数 $L(\boldsymbol{x}, \lambda)$ 的无约束优化问题.



(a) 等式约束 (b) 不等式约束

附图B. 1 拉格朗日乘子法的几何含义: 在 (a) 等式约束 $g(\boldsymbol{x}) = 0$ 或 (b) 不等式约束 $g(\boldsymbol{x}) \leqslant 0$ 下, 最小化目标函数 $f(\boldsymbol{x})$. 红色曲线表示 $g(\boldsymbol{x}) = 0$ 构成的曲面, 而其围成的阴影区域表示 $g(\boldsymbol{x}) < 0$.

现在考虑不等式约束 $g(\boldsymbol{x}) \leqslant 0$, 如附图B. 1 所示, 此时最优点 $\boldsymbol{x}^*$ 或在 $g(\boldsymbol{x}) < 0$ 的区域中, 或在边界 $g(\boldsymbol{x}) = 0$ 上. 对于 $g(\boldsymbol{x}) < 0$ 的情形, 约束 $g(\boldsymbol{x}) \leqslant 0$ 不起作用, 可直接通过条件 $\nabla f(\boldsymbol{x}) = 0$ 来获得最优点; 这等价于将 λ 置零然后对 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda)$ 置零得到最优点. $g(\boldsymbol{x}) = 0$ 的情形类似于上面等式约束的分析, 但需注意的是, 此时 $\nabla f(\boldsymbol{x}^*)$ 的方向必与 $\nabla g(\boldsymbol{x}^*)$ 相反, 即存在常数 $\lambda > 0$ 使得 $\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0$. 整合这两种情形, 必满足 $\lambda g(\boldsymbol{x}) = 0$. 因此, 在约束 $g(\boldsymbol{x}) \leqslant 0$ 下最小化 $f(\boldsymbol{x})$, 可转化为在如下约束下最小化式(B.2) 的拉格朗日函数:

$$\begin{cases} g(\boldsymbol{x}) \leqslant 0; \\ \lambda \geqslant 0\ ; \\ \mu_j g_j(\boldsymbol{x}) = 0\ . \end{cases} \tag{B.3}$$

式(B.3)称为 Karush-Kuhn-Tucker (简称KKT)条件.

上述做法可推广到多个约束. 考虑具有 m 个等式约束和 n 个不等式约束, 且可行域 $\mathbb{D} \subset \mathbb{R}^d$ 非空的优化问题

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \quad (i = 1, \ldots, m)\ , \\ & g_j(\boldsymbol{x}) \leqslant 0 \quad (j = 1, \ldots, n)\ . \end{aligned} \tag{B.4}$$

引入拉格朗日乘子 $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^{\mathrm{T}}$ 和 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}}$, 相应的拉格

朗日函数为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \ , \tag{B.5}$$

由不等式约束引入的 KKT 条件($j = 1, 2, \ldots, n$)为

$$\begin{cases} g_j(\boldsymbol{x}) \leqslant 0; \\ \mu_j \geqslant 0 \ ; \\ \mu_j g_j(\boldsymbol{x}) = 0 \ . \end{cases} \tag{B.6}$$

一个优化问题可以从两个角度来考察, 即“主问题”(primal problem)和“对偶问题”(dual problem). 对主问题(B.4), 基于式(B.5), 其拉格朗日“对偶函数”(dual function) $\Gamma : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ 定义为

在推导对偶问题时, 常通过将拉格朗日乘子 $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ 对 $\boldsymbol{x}$ 求导并令导数为 0, 来获得对偶函数的表达形式.

$$\begin{aligned} \Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x} \in \mathbb{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= \inf_{\boldsymbol{x} \in \mathbb{D}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \right) \ . \end{aligned} \tag{B.7}$$

$\boldsymbol{\mu} \succeq 0$ 表示 $\boldsymbol{\mu}$ 的分量均为非负.

若 $\tilde{\boldsymbol{x}} \in \mathbb{D}$ 为主问题(B.4)可行域中的点, 则对任意 $\boldsymbol{\mu} \succeq 0$ 和 $\boldsymbol{\lambda}$ 都有

$$\sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x}) \leqslant 0 \ , \tag{B.8}$$

进而有

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathbb{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ \leqslant \ L(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \ \leqslant \ f(\tilde{\boldsymbol{x}}) \ . \tag{B.9}$$

若主问题(B.4)的最优值为 p^*, 则对任意 $\boldsymbol{\mu} \succeq 0$ 和 $\boldsymbol{\lambda}$ 都有

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leqslant p^* \ , \tag{B.10}$$

即对偶函数给出了主问题最优值的下界. 显然，这个下界取决于 $\boldsymbol{\mu}$ 和 $\boldsymbol{\lambda}$ 的值. 于是, 一个很自然的问题是: 基于对偶函数能获得的最好下界是什么? 这就引出了优化问题

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad \Gamma(\boldsymbol{\lambda},\boldsymbol{\mu}) \quad \text{s.t.} \quad \boldsymbol{\mu} \succeq 0 \,. \tag{B.11}$$

式(B.11)就是主问题(B.4)的对偶问题, 其中 $\boldsymbol{\lambda}$ 和 $\boldsymbol{\mu}$ 称为“对偶变量”(dual variable). 无论主问题(B.4)的凸性如何, 对偶问题(B.11)始终是凸优化问题.

考虑式(B.11)的最优值 d^*, 显然有 $d^* \leqslant p^*$, 这称为“弱对偶性”(weak duality)成立; 若 $d^* = p^*$, 则称为“强对偶性”(strong duality)成立, 此时由对偶问题能获得主问题的最优下界. 对于一般的优化问题, 强对偶性通常不成立. 但是, 若主问题为凸优化问题, 如式(B.4)中 $f(\boldsymbol{x})$ 和$g_j(\boldsymbol{x})$ 均为凸函数, $h_i(\boldsymbol{x})$ 为仿射函数, 且其可行域中至少有一点使不等式约束严格成立, 则此时强对偶性成立. 值得注意的是, 在强对偶性成立时, 将拉格朗日函数分别对原变量和对偶变量求导, 再并令导数等于零, 即可得到原变量与对偶变量的数值关系. 于是, 对偶问题解决了, 主问题也就解决了.

这称为 Slater 条件.

B.2 二次规划

二次规划(Quadratic Programming, 简称 QP)是一类典型的优化问题, 包括凸二次优化和非凸二次优化. 在此类问题中, 目标函数是变量的二次函数, 而约束条件是变量的线性不等式.

假定变量个数为 d, 约束条件的个数为 m, 则标准的二次规划问题形如

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b} \,, \end{aligned} \tag{B.12}$$

非标准二次规划问题中可以包含等式约束. 注意到等式约束能用两个不等式约束来代替; 不等式约束可通过增加松弛变量的方式转化为等式约束.

其中 $\boldsymbol{x}$ 为 d 维向量, $\mathbf{Q} \in \mathbb{R}^{d\times d}$ 为实对称矩阵, $\mathbf{A} \in \mathbb{R}^{m\times d}$ 为实矩阵, $\boldsymbol{b} \in \mathbb{R}^m$ 和 $\boldsymbol{c} \in \mathbb{R}^d$ 为实向量, $\mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b}$ 的每一行对应一个约束.

若 $\mathbf{Q}$ 为半正定矩阵, 则式(B.12)目标函数是凸函数, 相应的二次规划是凸二次优化问题; 此时若约束条件 $\mathbf{A}\boldsymbol{x} \leqslant \boldsymbol{b}$ 定义的可行域不为空, 且目标函数在此可行域有下界, 则该问题将有全局最小值. 若 $\mathbf{Q}$ 为正定矩阵, 则该问题有唯一的全局最小值. 若 $\mathbf{Q}$ 为非正定矩阵, 则式(B.12)是有多个平稳点和局部极小点的 NP 难问题.

常用的二次规划解法有椭球法(ellipsoid method)、内点法(interior point)、增广拉格朗日法(augmented Lagrangian)、梯度投影法(gradient projection) 等. 若 $\mathbf{Q}$ 为正定矩阵, 则相应的二次规划问题可由椭球法在多项式时间内求解.

B.3 半正定规划

半正定规划(Semi-Definite Programming, 简称SDP)是一类凸优化问题, 其中的变量可组织成半正定对称矩阵形式, 且优化问题的目标函数和约束都是这些变量的线性函数.

给定 $d \times d$ 的对称矩阵 $\mathbf{X}$、$\mathbf{C}$,

$$\mathbf{C} \cdot \mathbf{X} = \sum_{i=1}^{d} \sum_{j=1}^{d} C_{ij} X_{ij} \ , \tag{B.13}$$

若 $\mathbf{A}_i \ (i = 1, 2, \ldots, m)$ 也是 $d \times d$ 的对称矩阵, $b_i \ (i = 1, 2, \ldots, m)$ 为 m 个实数, 则半正定规划问题形如

$$\begin{aligned} \min_{\mathbf{X}} \quad & \mathbf{C} \cdot \mathbf{X} \\ \text{s.t.} \quad & \mathbf{A}_i \cdot \mathbf{X} = b_i \ , \ i = 1, 2, \ldots, m \\ & \mathbf{X} \succeq 0 \ . \end{aligned} \tag{B.14}$$

$\mathbf{X} \succeq 0$ 表示 $\mathbf{X}$ 半正定.

半正定规划与线性规划都拥有线性的目标函数和约束, 但半正定规划中的约束 $\mathbf{X} \succeq 0$ 是一个非线性、非光滑约束条件. 在优化理论中, 半正定规划具有一定的一般性, 能将几种标准的优化问题(如线性规划、二次规划)统一起来.

常见的用于求解线性规划的内点法经过少许改造即可求解半正定规划问题, 但半正定规划的计算复杂度较高, 难以直接用于大规模问题.

B.4 梯度下降法

梯度下降法(gradient descent)是一种常用的一阶(first-order)优化方法, 是求解无约束优化问题最简单、最经典的方法之一.

一阶方法仅使用目标函数的一阶导数, 不利用其高阶导数.

考虑无约束优化问题 $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, 其中 $f(\boldsymbol{x})$ 为连续可微函数. 若能构造一个序列 $\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots$ 满足

$$f(\boldsymbol{x}^{t+1}) < f(\boldsymbol{x}^t), \ \ t = 0, 1, 2, \ldots \tag{B.15}$$

则不断执行该过程即可收敛到局部极小点. 欲满足式(B.15), 根据泰勒展式有

$$f(\boldsymbol{x} + \Delta \boldsymbol{x}) \simeq f(\boldsymbol{x}) + \Delta \boldsymbol{x}^{\mathrm{T}} \nabla f(\boldsymbol{x}) \ , \tag{B.16}$$

于是, 欲满足 $f(\boldsymbol{x}+\Delta\boldsymbol{x})<f(\boldsymbol{x})$, 可选择

$$\Delta\boldsymbol{x}=-\gamma\nabla f(\boldsymbol{x})\ , \tag{B.17}$$

每步的步长 γ_t 可不同.

其中步长 γ 是一个小常数. 这就是梯度下降法.

L-Lipschitz条件是指对于任意 $\boldsymbol{x}$, 存在常数 L 使得 $\|\nabla f(\boldsymbol{x})\|\leqslant L$ 成立.

若目标函数 $f(\boldsymbol{x})$ 满足一些条件, 则通过选取合适的步长, 就能确保通过梯度下降收敛到局部极小点. 例如若 $f(\boldsymbol{x})$ 满足 L-Lipschitz 条件, 则将步长设置为 $1/(2L)$ 即可确保收敛到局部极小点. 当目标函数为凸函数时, 局部极小点就对应着函数的全局最小点, 此时梯度下降法可确保收敛到全局最优解.

当目标函数 $f(\boldsymbol{x})$ 二阶连续可微时, 可将式(B.16)替换为更精确的二阶泰勒展式, 这样就得到了牛顿法(Newton's method). 牛顿法是典型的二阶方法, 其迭代轮数远小于梯度下降法. 但牛顿法使用了二阶导数 $\nabla^2 f(\boldsymbol{x})$, 其每轮迭代中涉及到海森矩阵(A.21)的求逆, 计算复杂度相当高, 尤其在高维问题中几乎不可行. 若能以较低的计算代价寻找海森矩阵的近似逆矩阵, 则可显著降低计算开销, 这就是拟牛顿法(quasi-Newton method).

B.5 坐标下降法

求解极大值问题时亦称"坐标上升法" (coordinate ascent).

坐标下降法(coordinate descent)是一种非梯度优化方法, 它在每步迭代中沿一个坐标方向进行搜索, 通过循环使用不同的坐标方向来达到目标函数的局部极小值.

不妨假设目标是求解函数 $f(\boldsymbol{x})$ 的极小值, 其中 $\boldsymbol{x}=(x_1,x_2,\ldots,x_d)^{\mathrm{T}}\in\mathbb{R}^d$ 是一个 d 维向量. 从初始点 $\boldsymbol{x}^0$ 开始, 坐标下降法通过迭代地构造序列 $\boldsymbol{x}^0,\boldsymbol{x}^1,\boldsymbol{x}^2,\ldots$ 来求解该问题, $\boldsymbol{x}^{t+1}$ 的第 i 个分量 x_i^{t+1} 构造为

$$x_i^{t+1}=\underset{y\in\mathbb{R}}{\arg\min}\ f(x_1^{t+1},\ldots,x_{i-1}^{t+1},y,x_{i+1}^{t},\ldots,x_d^{t})\ . \tag{B.18}$$

通过执行此操作, 显然有

$$f(\boldsymbol{x}^0)\geqslant f(\boldsymbol{x}^1)\geqslant f(\boldsymbol{x}^2)\geqslant\ldots \tag{B.19}$$

与梯度下降法类似, 通过迭代执行该过程, 序列 $\boldsymbol{x}^0,\boldsymbol{x}^1,\boldsymbol{x}^2,\ldots$ 能收敛到所期望的局部极小点或驻点(stationary point).

坐标下降法不需计算目标函数的梯度, 在每步迭代中仅需求解一维搜索问题, 对于某些复杂问题计算较为简便. 但若目标函数不光滑, 则坐标下降法有可能陷入非驻点(non-stationary point).

C 概率分布

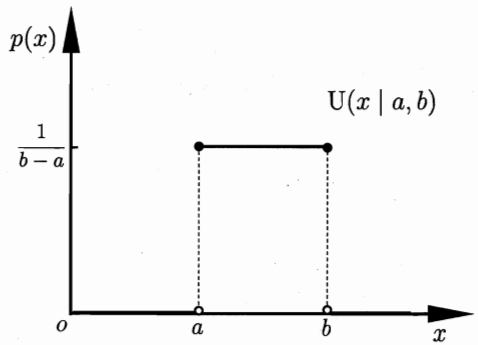
C.1 常见概率分布

本节简要介绍几种常见概率分布. 对于每种分布, 我们将给出概率密度函数以及期望 $\mathbb{E}[\cdot]$、方差 $\mathrm{var}[\cdot]$ 和协方差 $\mathrm{cov}[\cdot,\cdot]$ 等几个主要的统计量.

C.1.1 均匀分布

这里仅介绍连续均匀分布.

均匀分布(uniform distribution)是关于定义在区间 $[a,b]$ $(a<b)$ 上连续变量的简单概率分布, 其概率密度函数如附图C.1 所示.



附图C. 1 均匀分布的概率密度函数

$$p(x\mid a,b)=\mathrm{U}(x\mid a,b)=\frac{1}{b-a}\ ; \tag{C.1}$$

$$\mathbb{E}[x]=\frac{a+b}{2}\ ; \tag{C.2}$$

$$\mathrm{var}[x]=\frac{(b-a)^2}{12}\ . \tag{C.3}$$

不难发现, 若变量 x 服从均匀分布 $\mathrm{U}(x\mid 0,1)$ 且 $a<b$, 则 $a+(b-a)x$ 服从均匀分布 $\mathrm{U}(x\mid a,b)$.

C.1.2 伯努利分布

以瑞士数学家雅各布·伯努利 (Jacob Bernoulli, 1654–1705)的名字命名.

伯努利分布(Bernoulli distribution)是关于布尔变量 $x\in\{0,1\}$ 的概率分布, 其连续参数 $\mu\in[0,1]$ 表示变量 $x=1$ 的概率.

$$P(x\mid\mu)=\mathrm{Bern}(x\mid\mu)=\mu^x(1-\mu)^{1-x}\ ; \tag{C.4}$$

$$\mathbb{E}[x] = \mu \; ; \tag{C.5}$$

$$\mathrm{var}[x] = \mu(1-\mu) \; . \tag{C.6}$$

C.1.3 二项分布

二项分布(binomial distribution)用以描述 N 次独立的伯努利实验中有 m 次成功(即 $x=1$)的概率, 其中每次伯努利实验成功的概率为 $\mu \in [0,1]$.

$$P(m \mid N, \mu) = \mathrm{Bin}(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \; ; \tag{C.7}$$

$$\mathbb{E}[x] = N\mu \; ; \tag{C.8}$$

$$\mathrm{var}[x] = N\mu(1-\mu) \; . \tag{C.9}$$

对于参数 μ, 二项分布的共轭先验分布是贝塔分布. 共轭分布参见 C.2.

当 $N=1$ 时, 二项分布退化为伯努利分布.

C.1.4 多项分布

若将伯努利分布由单变量扩展为 d 维向量 $\boldsymbol{x}$, 其中 $x_i \in \{0,1\}$ 且 $\sum_{i=1}^d x_i = 1$, 并假设 x_i 取 1 的概率为 $\mu_i \in [0,1]$, $\sum_{i=1}^d \mu_i = 1$, 则将得到离散概率分布

$$P(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^d \mu_i^{x_i} \; ; \tag{C.10}$$

$$\mathbb{E}[x_i] = \mu_i \; ; \tag{C.11}$$

$$\mathrm{var}[x_i] = \mu_i(1-\mu_i) \; ; \tag{C.12}$$

$$\mathrm{cov}[x_j, x_i] = \mathbb{I}[j=i] \, \mu_i \; . \tag{C.13}$$

对于参数 $\boldsymbol{\mu}$, 多项分布的共轭先验分布是狄利克雷分布. 共轭分布参见 C.2.

在此基础上扩展二项分布则得到多项分布(multinomial distribution), 它描述了在 N 次独立实验中有 m_i 次 $x_i = 1$ 的概率.

$$\begin{aligned} P(m_1, m_2, \ldots, m_d \mid N, \boldsymbol{\mu}) &= \mathrm{Mult}(m_1, m_2, \ldots, m_d \mid N, \boldsymbol{\mu}) \\ &= \frac{N!}{m_1! \, m_2! \, \ldots \, m_d!} \prod_{i=1}^d \mu_i^{m_i} \; ; \end{aligned} \tag{C.14}$$

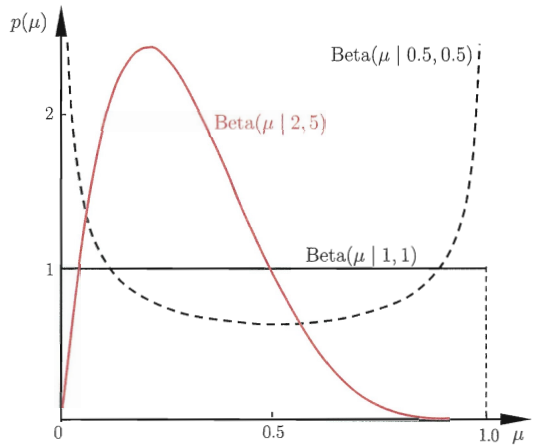
$$\mathbb{E}[m_i] = N\mu_i \; ; \tag{C.15}$$

$$\mathrm{var}[m_i] = N\mu_i(1-\mu_i)\ ; \tag{C.16}$$

$$\mathrm{cov}[m_j, m_i] = -N\mu_j\mu_i\ . \tag{C.17}$$

C.1.5 贝塔分布

贝塔分布(Beta distribution)是关于连续变量 $\mu \in [0,1]$ 的概率分布, 它由两个参数 $a>0$ 和 $b>0$ 确定, 其概率密度函数如附图C.2 所示.



附图C. 2 贝塔分布的概率密度函数

$$p(\mu \mid a,b) = \mathrm{Beta}(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

$$= \frac{1}{B(a,b)}\mu^{a-1}(1-\mu)^{b-1}\ ; \tag{C.18}$$

$$\mathbb{E}[\mu] = \frac{a}{a+b}\ ; \tag{C.19}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}\ , \tag{C.20}$$

其中 $\Gamma(a)$ 为 Gamma 函数

$$\Gamma(a) = \int_0^{+\infty} t^{a-1}e^{-t}\mathrm{dt}\ , \tag{C.21}$$

$B(a,b)$ 为 Beta 函数

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\ . \tag{C.22}$$

当 $a=b=1$ 时, 贝塔分布退化为均匀分布.

C.1.6 狄利克雷分布

以德国数学家狄利克雷(1805—1859)的名字命名.

狄利克雷分布(Dirichlet distribution) 是关于一组 d 个连续变量 $\mu_i \in [0,1]$ 的概率分布, $\sum_{i=1}^{d}\mu_i = 1$. 令 $\boldsymbol{\mu} = (\mu_1;\mu_2;\ldots;\mu_d)$, 参数 $\boldsymbol{\alpha} = (\alpha_1;\alpha_2;\ldots;\alpha_d)$, $\alpha_i > 0$, $\hat{\alpha} = \sum_{i=1}^{d}\alpha_i$.

$$p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_i)}\prod_{i=1}^{d}\mu_i^{\alpha_i-1}\ ; \tag{C.23}$$

$$\mathbb{E}[\mu_i] = \frac{\alpha_i}{\hat{\alpha}}\ ; \tag{C.24}$$

$$\mathrm{var}[\mu_i] = \frac{\alpha_i(\hat{\alpha}-\alpha_i)}{\hat{\alpha}^2(\hat{\alpha}+1)}\ ; \tag{C.25}$$

$$\mathrm{cov}[\mu_j,\mu_i] = \frac{\alpha_j\alpha_i}{\hat{\alpha}^2(\hat{\alpha}+1)}\ . \tag{C.26}$$

当 $d=2$ 时, 狄利克雷分布退化为贝塔分布.

C.1.7 高斯分布

高斯分布(Gaussian distribution)亦称正态分布(normal distribution), 是应用最为广泛的连续概率分布.

σ 为标准差.

对于单变量 $x \in (-\infty,\infty)$, 高斯分布的参数为均值 $\mu \in (-\infty,\infty)$ 和方差 $\sigma^2 > 0$. 附图C.3 给出了在几组不同参数下高斯分布的概率密度函数.

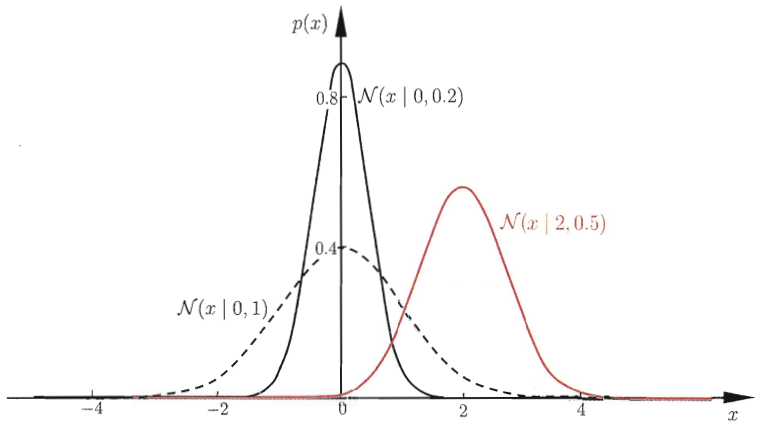
$$p(x \mid \mu,\sigma^2) = \mathcal{N}(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\ ; \tag{C.27}$$

$$\mathbb{E}[x] = \mu\ ; \tag{C.28}$$

$$\mathrm{var}[x] = \sigma^2\ . \tag{C.29}$$

对于 d 维向量 $\boldsymbol{x}$, 多元高斯分布的参数为 d 维均值向量 $\boldsymbol{\mu}$ 和 $d\times d$ 的对称正定协方差矩阵 $\boldsymbol{\Sigma}$.

$$\begin{aligned} p(\boldsymbol{x} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^d\det(\boldsymbol{\Sigma})}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\ ; \end{aligned} \tag{C.30}$$



附图C. 3 高斯分布的概率密度函数

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}\ ; \tag{C.31}$$

$$\mathrm{cov}[\boldsymbol{x}] = \boldsymbol{\Sigma}\ . \tag{C.32}$$

C.2 共轭分布

假设变量 x 服从分布 $P(x \mid \Theta)$, 其中 Θ 为参数, $X = \{x_1, x_2, \ldots, x_m\}$ 为变量 x 的观测样本, 假设参数 Θ 服从先验分布 $\Pi(\Theta)$. 若由先验分布 $\Pi(\Theta)$ 和抽样分布 $P(X \mid \Theta)$ 决定的后验分布 $F(\Theta \mid X)$ 与 $\Pi(\Theta)$ 是同种类型的分布, 则称先验分布 $\Pi(\Theta)$ 为分布 $P(x \mid \Theta)$ 或 $P(X \mid \Theta)$ 的共轭分布(conjugate distribution).

例如, 假设 $x \sim \mathrm{Bern}(x \mid \mu)$, $X = \{x_1, x_2, \ldots, x_m\}$ 为观测样本, $\bar{x}$ 为观测样本的均值, $\mu \sim \mathrm{Beta}(\mu \mid a, b)$, 其中 a, b 为已知参数, 则 μ 的后验分布

$$\begin{aligned} F(\mu \mid X) &\propto \mathrm{Beta}(\mu \mid a, b) P(X \mid \mu) \\ &= \frac{\mu^{a-1}(1-\mu)^{b-1}}{B(a,b)} \mu^{m\bar{x}} (1-\mu)^{m-m\bar{x}} \\ &= \frac{1}{B(a+m\bar{x}, b+m-m\bar{x})} \mu^{a+m\bar{x}-1} (1-\mu)^{b+m-m\bar{x}-1} \\ &= \mathrm{Beta}(\mu \mid a', b')\ , \end{aligned} \tag{C.33}$$

亦为贝塔分布, 其中 $a' = a + m\bar{x}$, $b' = b + m - m\bar{x}$, 这意味着贝塔分布与伯努利分布共轭. 类似可知, 多项分布的共轭分布是狄利克雷分布, 而高斯分布的共轭分布仍是高斯分布.

这里仅考虑高斯分布方差已知、均值服从先验的情形.

先验分布反映了某种先验信息, 后验分布既反映了先验分布提供的信息、又反映了样本提供的信息. 当先验分布与抽样分布共轭时, 后验分布与先验分布属于同种类型, 这意味着先验信息与样本提供的信息具有某种同一性. 于是, 若使用后验分布作为进一步抽样的先验分布, 则新的后验分布仍将属于同种类型. 因此, 共轭分布在不少情形下会使问题得以简化. 例如在式(C.33)的例子中, 对服从伯努利分布的事件 X 使用贝塔先验分布, 则贝塔分布的参数值 a 和 b 可视为对伯努利分布的真实情况(事件发生和不发生)的预估. 随着“证据”(样本)的不断到来, 贝塔分布的参数值从 a, b 变化为 $a+m\bar{x}, b+m-m\bar{x}$, 且 $a/(a+b)$ 将随着 m 的增大趋近于伯努利分布的真实参数值 $\bar{x}$. 显然, 使用共轭先验之后, 只需调整 a 和 b 这两个预估值即可方便地进行模型更新.

C.3 KL散度

KL散度(Kullback-Leibler divergence), 亦称相对熵(relative entropy)或信息散度(information divergence), 可用于度量两个概率分布之间的差异. 给定两个概率分布 P 和 Q, 二者之间的KL散度定义为

这里假设两个分布均为连续型概率分布; 对于离散型概率分布, 只需将定义中的积分替换为对所有离散值遍历求和.

$$\mathrm{KL}(P\|Q)=\int_{-\infty}^{\infty}p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x\ , \tag{C.34}$$

其中 $p(x)$ 和 $q(x)$ 分别为 P 和 Q 的概率密度函数.

KL散度满足非负性, 即

$$\mathrm{KL}(P\|Q)\geqslant 0\ , \tag{C.35}$$

当且仅当 $P=Q$ 时 $\mathrm{KL}(P\|Q)=0$. 但是, KL散度不满足对称性, 即

$$\mathrm{KL}(P\|Q)\neq\mathrm{KL}(Q\|P)\ , \tag{C.36}$$

度量应满足四个基本性质, 参见9.3节.

因此, KL散度不是一个度量(metric).

若将KL散度的定义(C.34)展开, 可得

$$\begin{aligned}\mathrm{KL}(P\|Q)&=\int_{-\infty}^{\infty}p(x)\log p(x)\mathrm{d}x-\int_{-\infty}^{\infty}p(x)\log q(x)\mathrm{d}x\\&=-H(P)+H(P,Q)\ ,\end{aligned} \tag{C.37}$$

其中 $H(P)$ 为熵(entropy), $H(P,Q)$ 为 P 和 Q 的交叉熵(cross entropy). 在信

息论中, 熵 $H(P)$ 表示对来自 P 的随机变量进行编码所需的最小字节数, 而交叉熵 $H(P,Q)$ 则表示使用基于 Q 的编码对来自 P 的变量进行编码所需的字节数. 因此, KL散度可认为是使用基于 Q 的编码对来自 P 的变量进行编码所需的“额外”字节数; 显然, 额外字节数必然非负, 当且仅当 $P=Q$ 时额外字节数为零.

后　记

写作本书的主因, 是 2016 年准备在南京大学开设“机器学习”课. 十五年前笔者曾主张开设此课, 但那时国内对机器学习闻之不多, 不少人听到这个名字的第一反应是“学习什么机器?”学校估计学生兴趣不大, 于是笔者开设了“数据挖掘”这门名字听上去就觉得很有用的课. 被评为省优秀研究生课程后, 又给本科生单开了一门“数据挖掘导论”. 这两门课很受欢迎, 选修学生很多, 包括不少外来蹭听生. 虽然课上有一多半其实在讲机器学习, 但笔者仍一直希望专开一门机器学习课, 因笔者以为机器学习迟早会变成计算机学科的基础内容.

图灵奖得主 E. W. Dijkstra 曾说“计算机科学并不仅是关于计算机, 就像天文学并不仅是关于望远镜”. 正如天文学早期的研究关注如何制造望远镜, 计算机科学早期研究是在关注如何令计算机运转. 到了今天, 建造强大的天文望远镜虽仍重要, 但天文学更要紧的是“用”望远镜来开展研究. 类似地, 计算机科学发展至今, 也该到了从关注“造”计算机转入更关注“用”计算机来认识和改造世界的阶段, 其中最重要的无疑是用计算机对数据进行分析, 因为这是计算的主要目的, 而这就离不开机器学习. 十多年前在国内某次重要论坛上笔者刚抛出此观点就被专家迎头指斥, 但今日来看, 甚至很多计算机学科外人士都已对机器学习的重大价值津津乐道, 现在才开设机器学习基础课似乎已有点嫌晚了.

1995 年在南大图书馆偶然翻看了《机器学习: 通往人工智能的途径》, 这算是笔者接触机器学习的开始. 那时机器学习在国内问津者寥, 甚至连科研人员申请基金项目也无合适代码方向可报. 周边无专家可求教, 又因国内科研经费匮乏而几无国际交流, 加之学校尚无互联网和电子文献库, 能看到的最新文献仅是两年前出版且页数不全的某 IEEE 汇刊……可谓举步维艰, 经历的困惑和陷阱不可胜数. 笔者切身体会到, 入门阶段接触的书籍是何等重要, 对自学者尤甚. 一本好书能让人少走许多弯路, 材料不佳则后续要花费数倍精力方能纠偏. 中文书当然要国人自己来写. 虽已不需靠“写书出名”, 且深知写教科书极耗时间精力, 但踌躇后笔者仍决定动手写这本书, 唯望为初学者略尽绵薄之力.

有人说“一千个人眼中就有一千个哈姆雷特”, 一个学科何尝不是如此. 之所以不欲使用市面上流行的教科书(主要是英文的), 除了觉得对大多数中国学生来说中文教科书更便于学习, 另一个原因则是希望从笔者自己的视角来展现机器学习.

2013 年中开始规划提纲, 由此进入了焦躁的两年. 该写哪些内容、先写什么后写什么、从哪个角度写、写到什么程度, 总有千丝万缕需考虑. 及至写作进行, 更是战战兢兢, 深恐不慎误人子弟. 写书难, 写教科书更难. 两年下来, 甘苦自知. 子曰:“取乎其上, 得乎其中; 取乎其中, 得乎其下”, 且以顶级的态度, 出一本勉强入得方家法眼之书.

本书贯穿以西瓜为例, 一则因为瓜果中笔者尤喜西瓜, 二则因为西瓜在笔者所生活的区域有个有趣的蕴义. 朋友小聚、请客吃饭, 菜已全而主未知, 或馔未齐而人待走, 都挺尴尬. 于是聪明人发明了“潜规则”: 席终上西瓜. 无论整盘抑或小碟, 宾主见瓜至, 则心领神会准备起身, 皆大欢喜. 久而久之, 无论菜肴价格贵贱、场所雅鄙, 宴必有西瓜. 若将宴席比作(未来)应用系统, 菜肴比作所涉技术, 则机器学习好似那必有的西瓜, 它可能不是最“高大上”的, 但却是离不了的、没用上总觉得不甘心的.

本书写作过程从材料搜集, 到习题设计, 再到阅读校勘, 都得到了笔者的很多学生、同事和学术界朋友的支持和帮助, 在此谨列出他们的姓名以致谢意(姓氏拼音序): 陈松灿, 戴望州, 高阳, 高尉, 黄圣君, 黎铭, 李楠, 李武军, 李宇峰, 钱超, 王魏, 王威廉, 吴建鑫, 徐淼, 俞扬, 詹德川, 张利军, 张敏灵, 朱军. 书稿在 LAMDA 组学生 2015 年暑期讨论班上试讲, 高斌斌、郭翔宇、李绍园、钱鸿、沈芷玉、叶翰嘉、张腾等同学又帮助发现了许多笔误. 特别感谢李楠把笔者简陋的手绘图转变为精致的插图, 俞扬帮助调整排版格式和索引, 刘冲把笔者对封面设计的想法具体表现出来.

中国计算机学会终身成就奖得主、中国科学院院士陆汝钤先生是我国人工智能事业的开拓者之一, 他在 1988 年和 1996 年出版的《人工智能》(上、下册)曾给予笔者很多启发. 承蒙陆老师厚爱在百忙中为本书作序, 不胜惶恐之至. 陆老师在序言中提出的问题很值得读者在本书之后的进阶学习与研究中深思.

感谢清华大学出版社薛慧老师为本书出版所做的努力. 十二年前笔者入选国家杰出青年科学基金时薛老师即邀著书, 笔者以年纪尚轻、学力未逮婉辞. 十年前“机器学习及其应用”研讨会(MLA)从陆汝钤院士肇始的复旦大学智能信息处理重点实验室移师南京, 参会人数从复旦最初的 20 人, 发展到 2010 年 400 余人, 此后在清华、复旦、西电达 800 余人，今年再回南大竟至 1300 余人, 场面热烈. MLA 倡导“学术至上、其余从简”, 不搞繁文缛节，参会免费. 但即便如此, 仍有很多感兴趣的师生因旅费不菲而难以参加. 于是笔者提议每两年以《机器学习及其应用》为题出版一本报告选集以飨读者. 这个主意得到了薛老师、陆老师以及和笔者一起长期组织 MLA、去年因病去世的王珏老师的大力支持. 此类专业性学术文集销量不大, 出版社多半要贴钱. 笔者曾跟薛老师说, 自著的第一本中文书必交由薛老师在清华出版, 或可稍为出版社找补. 转眼《机器学习及其应用》系列已出到第六本, 薛老师或以为十年前是玩笑话, 某日告之书快完稿时她蓦然惊喜.

最后要感谢笔者的家人, 本书几乎耗尽了两年来笔者所有的节假日和空闲时间. 写作时垂髫犬子常跑来案边, 不是问“爸爸去哪儿?”而是看几眼然后问“爸爸你又写了几页?”为了给他满意的答复, 笔者埋头努力.

周志华

2015 年 11 月于南京渐宽斋

索　引

MACHINE LEARNING

机器学习

周志华 著

清华大学出版社
北 京

内 容 简 介

机器学习是计算机科学的重要分支领域. 本书作为该领域的入门教材，在内容上尽可能涵盖机器学习基础知识的各方面. 全书共 16 章，大致分为 3 个部分：第 1 部分（第 1～3 章）介绍机器学习的基础知识；第 2 部分（第 4～10 章）讨论一些经典而常用的机器学习方法（决策树、神经网络、支持向量机、贝叶斯分类器、集成学习、聚类、降维与度量学习）；第 3 部分（第 11～16 章）为进阶知识，内容涉及特征选择与稀疏学习、计算学习理论、半监督学习、概率图模型、规则学习以及强化学习等. 每章都附有习题并介绍了相关阅读材料，以便有兴趣的读者进一步钻研探索.

本书可作为高等院校计算机、自动化及相关专业的本科生或研究生教材，也可供对机器学习感兴趣的研究人员和工程技术人员阅读参考.

图书在版编目(CIP)数据

机器学习/周志华著. --北京：清华大学出版社，2016
ISBN 978-7-302-42328-7

Ⅰ. ①机…　Ⅱ. ①周…　Ⅲ. ①机器学习　Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2015)第 287090 号

责任编辑：薛　慧
封面设计：何凤霞
责任校对：刘玉霞
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编：**100084
社 总 机：010-62770175　　**邮　　购：**010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：北京亿浓世纪彩色印刷有限公司
经　　销：全国新华书店
开　　本：210mm×235mm　　**印　张：**27.75　　**字　数：**626 千字
版　　次：2016 年 1 月第 1 版　　**印　次：**2016 年 1 月第 1 次印刷
印　　数：1～5000
定　　价：88.00 元

产品编号：064027-01

前　言

这是一本面向中文读者的机器学习教科书, 为了使尽可能多的读者通过本书对机器学习有所了解, 作者试图尽可能少地使用数学知识. 然而, 少量的概率、统计、代数、优化、逻辑知识似乎不可避免. 因此, 本书更适合大学三年级以上的理工科本科生和研究生, 以及具有类似背景的对机器学习感兴趣的人士. 为方便读者, 本书附录给出了一些相关数学基础知识简介.

全书共 16 章, 大体上可分为 3 个部分：第 1 部分包括第 1~3 章, 介绍机器学习基础知识; 第 2 部分包括第 4~10 章, 介绍一些经典而常用的机器学习方法; 第 3 部分包括第 11~16 章, 介绍一些进阶知识. 前 3 章之外的后续各章均相对独立, 读者可根据自己的兴趣和时间情况选择使用. 根据课时情况, 一个学期的本科生课程可考虑讲授前 9 章或前 10 章; 研究生课程则不妨使用全书.

书中除第 1 章外, 每章都给出了十道习题. 有的习题是帮助读者巩固本章学习, 有的是为了引导读者扩展相关知识. 一学期的一般课程可使用这些习题, 再辅以两到三个针对具体数据集的大作业. 带星号的习题则有相当难度, 有些并无现成答案, 谨供富有进取心的读者启发思考.

本书在内容上尽可能涵盖机器学习基础知识的各方面, 但作为机器学习入门读物且因授课时间的考虑, 很多重要、前沿的材料未能覆盖, 即便覆盖到的部分也仅是管中窥豹, 更多的内容留待读者在进阶课程中学习. 为便于有兴趣的读者进一步钻研探索, 本书每章均介绍了一些阅读材料, 谨供读者参考.

笔者以为, 对学科相关的重要人物和事件有一定了解, 将会增进读者对该学科的认识. 本书在每章最后都写了一个与该章内容相关的小故事, 希望有助于读者增广见闻, 并且在紧张的学习过程中稍微放松调剂一下.

书中不可避免地涉及大量外国人名, 若全部译为中文, 则读者在日后进一步阅读文献时或许会对不少人名产生陌生感, 不利于进一步学习. 因此, 本书仅对一般读者耳熟能详的名字如“图灵”等加以直接使用, 对故事中的一些主要人物给出了译名, 其他则保持外文名.

机器学习发展极迅速, 目前已成为一个广袤的学科, 罕有人士能对其众多分支领域均有精深理解. 笔者自认才疏学浅, 仅略知皮毛, 更兼时间和精力所限, 书中错谬之处在所难免, 若蒙读者诸君不吝告知, 将不胜感激.

周志华

2015 年 6 月

序　言

在人工智能界有一种说法, 认为机器学习是人工智能领域中最能够体现智能的一个分支. 从历史来看, 机器学习似乎也是人工智能中发展最快的分支之一. 在二十世纪八十年代的时候, 符号学习可能还是机器学习的主流, 而自从二十世纪九十年代以来, 就一直是统计机器学习的天下了. 不知道是否可以这样认为: 从主流为符号机器学习发展到主流为统计机器学习, 反映了机器学习从纯粹的理论研究和模型研究发展到以解决现实生活中实际问题为目的的应用研究, 这是科学研究的一种进步. 有关机器学习的专著国内出版的不是很多. 前两年有李航教授的《统计学习方法》出版, 以简要的方式介绍了一批重要和常用的机器学习方法. 此次周志华教授的鸿篇巨著《机器学习》则全面而详细地介绍了机器学习的各个分支, 既可作为教材, 又可作为自学用书和科研参考书.

翻阅书稿的过程引起了一些自己的思考, 平时由于和机器学习界的朋友接触多了, 经常获得一些道听途说的信息以及专家们对机器学习现状及其发展前途的评论. 在此过程中, 难免会产生一些自己的疑问. 我借此机会把它写下来放在这里, 算是一种“外行求教机器学习”.

问题一: 在人工智能发展早期, 机器学习的技术内涵几乎全部是符号学习. 可是从二十世纪九十年代开始, 统计机器学习犹如一匹黑马横空出世, 迅速压倒并取代了符号学习的地位. 人们可能会问: 在满目的统计学习期刊和会议文章面前, 符号学习是否被彻底忽略了? 它还能成为机器学习的研究对象吗? 它是否将继续在统计学习的阴影里生活并苟延残喘? 对这个问题有三种可能的答案: 一是告诉符号学习:“你就是该退出历史舞台, 认命吧!”二是告诉统计学习:“你的一言堂应该关门了!”单纯的统计学习已经走到了尽头, 再想往前走就要把统计学习和符号学习结合起来. 三是事物发展总会有“三十年河东, 三十年河西”的现象, 符号学习还有“翻身”的日子. 第一种观点我没有听人明说过, 但是我想恐怕有可能已经被许多人默认了. 第二种观点我曾听王珏教授多次说过. 他并不认为统计学习会衰退, 而只是认为机器学习已经到了一个转折点, 从今往后, 统计学习应该和知识的利用相结合, 这是一种“螺旋式上升, 进入更高级的形式”, 否则, 统计学习可能会停留于现状而止步不前. 王珏教授还认为：进入转折点的标志就是 Koller 等的《概率图模型》一书的出版. 至于第三种观点, 恰好我收到老朋友, 美国人工智能资深学者、俄亥俄大学 Chandrasekaran 教授的来信, 他正好谈起符号智能被统计智能“打压”的现象, 并且正好表达了河东河西的观点. 我请求他允许我把这段话引进正在撰写的序言中, 他爽快地同意了, 仅仅修改了几处私人通信的口吻. 全文如下:“最近几年, 人工智能在很大程度上集中于统计学和大数据. 我同意由于计算能力的大幅提高, 这些技术曾经取得过某些令人印象深刻的成果. 但是我们完全有理由相信, 虽然这些技术还会继续改进、提高, 总有一天这个领域(指 AI)会对它们说再见, 并转向更加基本的认知科学研究. 尽管钟摆的摆回去还需要一段时间, 我

相信定有必要把统计技术和对认知结构的深刻理解结合起来.”看来, Chandrasekaran 教授也并不认为若干年以后 AI 真会回到河西, 他的意见和王珏教授的意见基本一致, 但不仅限于机器学习, 而是涉及整个人工智能领域. 只是王珏教授强调知识, 而 Chandrasekaran 教授强调更加基本的“认知”.

问题二: 王珏教授认为统计机器学习不会“一路顺风”的判据是: 统计机器学习算法都是基于样本数据独立同分布的假设. 但是自然界现象千变万化, 王珏教授认为“哪有那么多独立同分布?”这就引来了下一个问题: “独立同分布”条件对于机器学习来讲真是必需的吗? 独立同分布的不存在一定是一个不可逾越的障碍吗? 无独立同分布条件下的机器学习也许只是一个难题, 而不是不可解问题. 我有一个“胡思乱想”, 认为前些时候出现的“迁移学习”也许会对这个问题的解决带来一线曙光. 尽管现在的迁移学习还要求迁移双方具备“独立同分布”条件, 但是不同分布之间的迁移学习, 同分布和异分布之间的迁移学习也许迟早会出现?

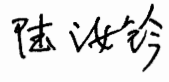
问题三: 近年来出现了一些新的动向, 例如“深度学习”、“无终止学习”等等, 社会上给予了特别关注, 尤其是深度学习. 但它们真的代表了机器学习的新的方向吗? 包括本书作者周志华教授在内的一些学者认为: 深度学习掀起的热潮也许大过它本身真正的贡献, 在理论和技术上并没有太多的创新, 只不过是由于硬件技术的革命, 计算机的速度大大提高了, 使得人们有可能采用原来复杂度很高的算法, 从而得到比过去更精细的结果. 当然这对于推动机器学习应用于实践有很大意义. 但我们不禁要斗胆问一句: 深度学习是否又要取代统计学习了? 事实上, 确有专家已经感受到来自深度学习的压力, 指出统计学习正在被深度学习所打压, 正如我们早就看到的符号学习被统计学习所打压. 不过我觉得这种打压还远没有强大到像统计学习打压符号学习的程度. 这一是因为深度学习的“理论创新”还不明显; 二是因为目前的深度学习主要适合于神经网络, 在各种机器学习方法百花盛开的今天, 它的应用范围还有限, 还不能直接说是连接主义方法的回归; 三是因为统计学习仍然在机器学习中被有效地普遍采用, “得道多助”, 想抛弃它不容易.

问题四: 机器学习研究出现以来,我们看到的主要是从符号方法到统计方法的演变, 用到的数学主要是概率统计. 但是, 数学之大, 就像大海. 难道只有统计方法适合于在机器学习方面应用吗? 当然, 我们也看到了一些其他数学分支在机器学习上的应用的好例子, 例如微分几何在流形学习上的应用, 微分方程在归纳学习上的应用. 但如果和统计方法相比, 它们都只能算是配角. 还有的数学分支如代数可能应用得更广, 但在机器学习中代数一般是作为基础工具来使用, 例如矩阵理论和特征值理论. 又如微分方程求解最终往往归结为代数问题求解. 它们可以算是幕后英雄:“出头露面的是概率和统计, 埋头苦干的是代数和逻辑”. 是否可以想象以数学方法为主角, 以统计方法为配角的机器学习理论呢? 在这方面, 流形学习已经“有点意思”了, 而彭实戈院士的倒排随机微分方程理论之预测金融走势, 也许是用高深数学推动新的机器学习模式的更好例子. 但是从宏观的角度看, 数学理论的介入程度还远远不够. 这里指的主要是深刻的、现代的数学理论, 我们期待着有更多数学家的参与, 开辟机器学习的新模式、新理论、新方向.

问题五: 上一个问题的延续: 符号机器学习时代主要以离散方法处理问题, 统计机器学习时代主要以连续方法处理问题. 这两种方法之间应该没有一条鸿沟. 流形学习中李群、李代数方法的引入给我们以很好的启示. 从微分流形到李群, 再从李群到李代数, 就是一个沟通连续和离散的过程. 然而, 现有的方法在数学上并不完美. 浏览流形学习的文献可知, 许多论文直接把任意数据集看成微分流形, 从而就认定测地线的存在并讨论起降维来了. 这样的例子也许不是个别的, 足可说明数学家介入机器学习研究之必要.

问题六: 大数据时代的出现, 有没有给机器学习带来本质性的影响? 理论上讲, 似乎“大数据”给统计机器学习提供了更多的机遇, 因为海量的数据更加需要统计、抽样的方法. 业界人士估计, 大数据的出现将使人工智能的作用更加突出. 有人把大数据处理分成三个阶段: 收集、分析和预测. 收集和分析的工作相对来说已经做得相当好了, 现在关注的焦点是要有科学的预测, 机器学习技术在这里不可或缺. 这一点大概毋庸置疑. 然而, 同样是使用统计、抽样方法, 同样是收集、分析和预测, 大数据时代使用这类方法和以前使用这类方法有什么本质的不同吗? 量变到质变是辩证法的一个普遍规律. 那么, 从前大数据时代到大数据时代, 数理统计方法有没有发生本质的变化? 反映到它们在机器学习上的应用有无本质变化? 大数据时代正在呼唤什么样的机器学习方法的产生? 哪些机器学习方法又是由于大数据研究的驱动而产生的呢?

以上这些话也许说得远了, 我们还是回到本书上来. 本书的作者周志华教授在机器学习的许多领域都有出色的贡献, 是中国机器学习研究的领军人物之一, 在国际学术界有着很高的声誉. 他在机器学习的一些重要领域, 例如集成学习、半监督学习、多示例和多标记学习等方面都做出了在国际上有重要影响的工作, 其中一些可以认为是中国学者在国际上的代表性贡献. 除了自身的学术研究以外, 他在推动中国的机器学习发展方面也做了许多工作. 例如他和不久前刚过世的王珏教授从 2002 年开始, 组织了系列化的“机器学习及其应用”研讨会. 初在复旦, 后移至南大举行, 越办越兴旺, 从单一的专家报告发展到专家报告、学生论坛和张贴论文三种方式同时举行, 参会者从数十人发展到数百人, 活动搞得有声有色, 如火如荼. 最近更是把研讨会推向全国高校轮流举行. 他和王珏教授紧密合作, 南北呼应, 人称“南周北王”. 王珏教授的离去使我们深感悲伤. 令我们欣慰的是国内不但有周志华教授这样的机器学习领军人物, 而且比周教授更年轻的许多机器学习青年才俊也成长起来了. 中国的机器学习大有希望.



中国科学院数学与系统科学研究院

2015 年 8 月于北京

主要符号表

x	标量
$\boldsymbol{x}$	向量
$\mathbf{x}$	变量集
$\mathbf{A}$	矩阵
$\mathbf{I}$	单位阵
$\mathcal{X}$	样本空间或状态空间
$\mathcal{D}$	概率分布
D	数据样本（数据集）
$\mathcal{H}$	假设空间
H	假设集
$\mathfrak{L}$	学习算法
$(\cdot,\ \cdot,\ \cdot)$	行向量
$(\cdot;\ \cdot;\ \cdot)$	列向量
$(\cdot)^{\mathrm{T}}$	向量或矩阵转置
$\{\cdots\}$	集合
$\lvert\{\cdots\}\rvert$	集合$\{\cdots\}$中元素个数
$\lVert\cdot\rVert_p$	L_p 范数, p 缺省时为 L_2 范数
$P(\cdot),\ P(\cdot \mid \cdot)$	概率质量函数, 条件概率质量函数
$p(\cdot),\ p(\cdot \mid \cdot)$	概率密度函数, 条件概率密度函数
$\mathbb{E}_{\cdot\sim\mathcal{D}}[f(\cdot)]$	函数 $f(\cdot)$ 对 $\cdot$ 在分布 $\mathcal{D}$ 下的数学期望; 意义明确时将省略 $\mathcal{D}$ 和(或) $\cdot$
$\sup(\cdot)$	上确界
$\mathbb{I}(\cdot)$	指示函数, 在 $\cdot$ 为真和假时分别取值为 $1, 0$
$\mathrm{sign}(\cdot)$	符号函数, 在 $\cdot < 0,\ = 0,\ > 0$ 时分别取值为 $-1, 0, 1$

目　录