第一章：介绍

（2）典型例子

授课教师：曹语

课程主页：`https://yucaoyc.github.io/math3806`

目录

最小二乘

- 最小二乘至少可以追溯到 Gauss 在 1820s 的论文。
- 今天我们先不把它当“公式推导”，而把它当成一个非常实用的任务：

  给一堆有噪声的数据，找一条“尽量贴近整体趋势”的曲线。

- 假设有数据点 (x_i, b_i), $i = 0, 1, 2, \cdots, \ell$，希望找到一个函数 S。对每个数据点，预测误差记为 $\delta_i = S(x_i) - b_i$。问题是怎样衡量“整体误差最小”?
- 问题：把 δ_i 看成一个向量后，若用欧式距离衡量“误差总大小”，应该选哪一个?

  (A) $\sum_{i=1}^{\ell} \delta_i$　　(B) $\sum_{i=1}^{\ell} |\delta_i|$　　(C) $\sum_{i=1}^{\ell} (\delta_i)^2$

- 如果使用欧式距离来描述误差，我们得到：

$$\min_{S} \sum_{i=0}^{\ell} \left(S(x_i) - b_i\right)^2$$

- 函数 S 太抽象，我们先从最常见、最好算的函数族入手：多项式。

- 考虑 $S(x) = c_0 + c_1 x + \cdots + c_q x^q$

$$\min_{c_0, c_1, \cdots, c_q} \sum_{i=0}^{\ell} \left[\sum_{j=0}^{q} c_j \left(x_i\right)^j - b_i\right]^2 = f_0(c_0, c_1, \cdots, c_q)$$

- 此处参数 $c_0, c_1, \cdots, c_q$ 是我们要优化的变量；它们所有可能的取值，构成了**备选解空间**。

热身：线性拟合 ($q = 1$)

如果我们把 S 限制为线性函数 $S(x) = c_0 + c_1 x$，则问题变成了

$$\min_{c_0, c_1} \sum_{i=0}^{\ell} (c_0 + c_1 x_i - b_i)^2$$

- 目标函数是关于 c_0, c_1 的开口向上的抛物面
- 极值点一定满足偏导数为 0

$$\begin{cases} \frac{\partial f_0}{\partial c_0} = 2\sum(S(x_i) - b_i) = 0 \\ \frac{\partial f_0}{\partial c_1} = 2\sum(S(x_i) - b_i)x_i = 0 \end{cases}$$

解出这个方程组，就能得到 (c_0, c_1)。

具体而言，我们对 c_0 和 c_1 分别求导:

- 关于 c_0 的导数为 $\frac{\partial f_0}{\partial c_0} = 2\sum_{i=0}^{\ell}(c_0 + c_1 x_i - b_i) = 0$
- 关于 c_1 的导数为 $\frac{\partial f_0}{\partial c_1} = 2\sum_{i=0}^{\ell}(c_0 + c_1 x_i - b_i)\, x_i = 0$

我们就得到了最优的线性拟合函数:

$$\begin{bmatrix} \sum_{i=0}^{\ell} 1 & \sum_{i=0}^{\ell} x_i \\ \sum_{i=0}^{\ell} x_i & \sum_{i=0}^{\ell} x_i^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{\ell} b_i \\ \sum_{i=0}^{\ell} b_i x_i \end{bmatrix}$$

一般形式

为找到最小化误差的参数，我们对每个参数求偏导并令其为零（找驻点）:

$$\frac{\partial f_0}{\partial c_k} = 2\sum_{i=0}^{\ell}\left(\sum_{j=0}^{q} c_j\,(x_i)^j \;-\; b_i\right)\; x_i^k = 0 \tag{1}$$

这一步本质是在做一件熟悉的事：“多元函数求极值 $\Rightarrow$ 导数等于 0”。

引入记号 $\boldsymbol{A}_{i,j} = x_i^j$，则方程可改写成

$$\sum_{i=0}^{\ell}\sum_{j=0}^{q} \boldsymbol{A}_{i,j}\boldsymbol{A}_{i,k}c_j = \sum_{i=0}^{\ell} \boldsymbol{A}_{i,k}b_i$$

于是方程可以更紧凑地写成矩阵形式（后续计算更方便）：

$$\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^{\mathsf{T}}\boldsymbol{b}$$

其中

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{00} & \boldsymbol{A}_{01} & \cdots & \boldsymbol{A}_{0q} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{A}_{\ell 0} & \boldsymbol{A}_{\ell 1} & \cdots & \boldsymbol{A}_{\ell q} \end{bmatrix} = \begin{bmatrix} 1 & x_0^1 & \cdots & x_0^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_\ell^1 & \cdots & x_\ell^q \end{bmatrix} \qquad \boldsymbol{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_q \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_0 \\ \vdots \\ b_\ell \end{bmatrix}$$

矩阵 $\boldsymbol{A}$ 称为范德蒙德矩阵 (Vandermonde Matrix)。

目标函数可以更简洁地写为

$$
\begin{aligned}
f_0(c_0, c_1, \cdots, c_q) &= \sum_{i=0}^{\ell} \left[\sum_{j=0}^{q} c_j \left(x_i\right)^j - b_i \right]^2 \\
&= \sum_{i=0}^{\ell} \left(\sum_{j=0}^{q} \boldsymbol{A}_{i,j} c_j - b_i \right)^2 \\
&= (\boldsymbol{A}\boldsymbol{c} - \boldsymbol{b})^\top (\boldsymbol{A}\boldsymbol{c} - \boldsymbol{b}) \\
&= \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{b}\|_2^2
\end{aligned}
$$

一般的最小二乘问题

$$\min_{\boldsymbol{c}} \ f_0(\boldsymbol{c}) = \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{b}\|_2^2$$

- 其中 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ 是数据矩阵，$\boldsymbol{c} \in \mathbb{R}^n$ 是待求参数，$\boldsymbol{b} \in \mathbb{R}^m$ 是观测数据。
- 当 $\boldsymbol{A}$ 列满秩（参数之间不“重复”）时，解可写成

$$\boldsymbol{c} = \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{b}$$

讨论：多项式怎么选？参数越多越好吗？

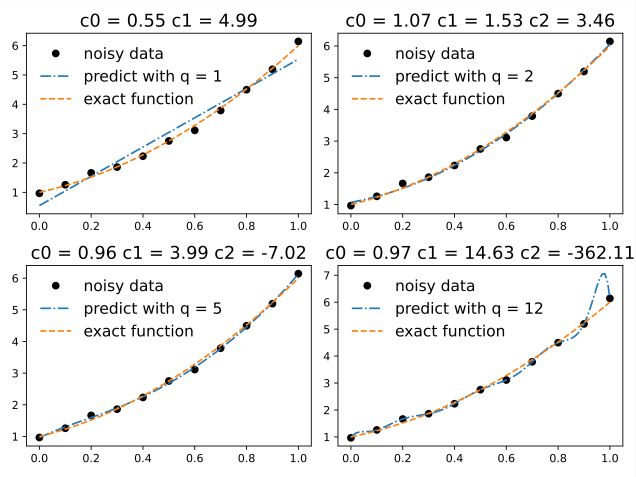
问题： 最小二乘中通常希望 $m \geq n$（数据量不少于参数个数）。直觉上，如果只有 2 个数据点 $(m = 2)$，却用二次函数拟合 $(n = 3)$，会发生什么？

提示：

- 线性代数视角：方程数少于未知数，$\boldsymbol{A}$ 是“矮矩阵”，解不唯一（有无穷多条抛物线穿过这两点）。
- 机器学习视角：即使所有点都在曲线上，这也没意义——我们是在“死记硬背”数据，而不是“学习”规律 **(过拟合/Overfitting)**。

测试例子

真解为 $S(x) = 3x^2 + 2x + 1$；我们基于真值对数据做了扰动



练习

考虑二维医疗数据 $\{(x_i, y_i)\}$，可把它理解为两个特征。已知部分样本的标签：

$$z_i \in \{\text{不健康} = 0, \text{健康} = 1\}.$$

目标是学习一个预测函数 $\phi : (x, y) \mapsto z$。

请尝试把它写成“数据拟合 + 优化”的形式（可按以下步骤思考）：

- 先假设一个参数化模型（如线性模型）；
- 再选一个误差函数来衡量预测与标签差异；
- 最后写出“最小化总误差”的优化问题。

目录

例子

最小二乘里，备选解在 $\mathbb{R}^n$ 中几乎“任意可取”；但现实决策常常带有约束（预算、资源、时间等）。 $\Longrightarrow$ 我们将以**线性规划问题**为例介绍

<u>问题</u>： 假如你有 10 元本金，准备买可乐和冰棍去山上卖。 2 元一听的可乐可赚 1 元；3 元一根的冰棍可赚 3 元。你会怎么进货，才能让利润最高?

例子

最小二乘里，备选解在 $\mathbb{R}^n$ 中几乎“任意可取”；但现实决策常常带有约束（预算、资源、时间等）。 $\Longrightarrow$ 我们将以**线性规划问题**为例介绍

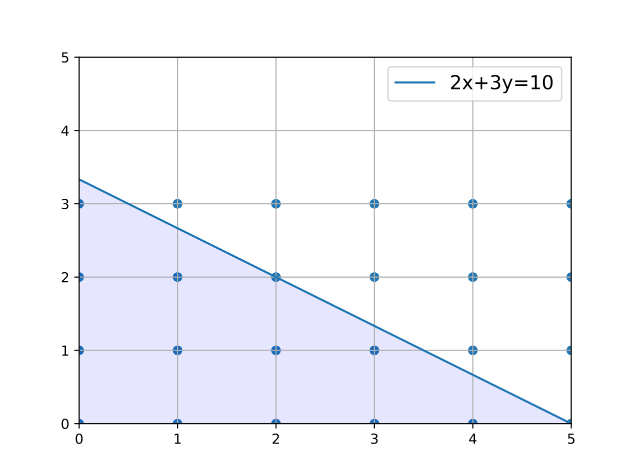
问题： 假如你有 10 元本金，准备买可乐和冰棍去山上卖。 2 元一听的可乐可赚 1 元；3 元一根的冰棍可赚 3 元。你会怎么进货，才能让利润最高？

倘若你购买 x 听可乐，与 y 根冰棍，则

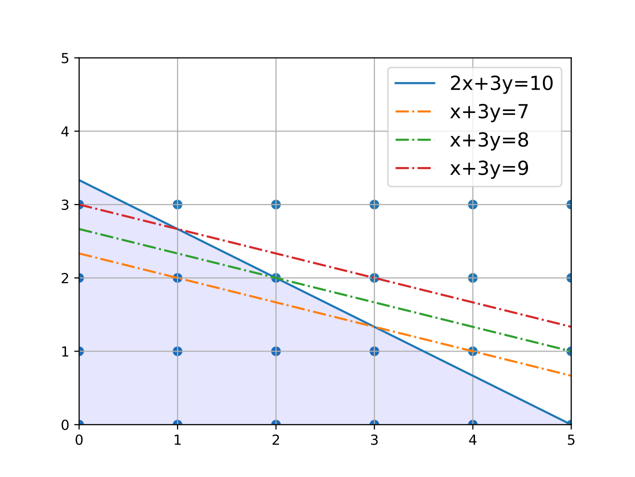
$$\begin{aligned} \text{优化目标：} \quad & \max \ x + 3y; \\ \text{限制条件：} \quad & 2x + 3y \leq 10 \\ & x, y \text{ 是非负整数.} \end{aligned}$$

限制条件

由于 x, y 都是整数，我们仅需考虑格点。该问题的可行解是什么?



答案



最优策略是购买 3 根冰棍，即使没有用完所有的钱；最高可挣 9 元。

实数空间的线性规划：资源分配的通用模型

不少工程问题（化工生产、物流调度、金融投资）都可以看作“在有限资源约束下，追求最大效益”。很多时候变量可以是连续实数（如生产 1000.5 吨原料）。

$$
\begin{array}{lll}
\text{minimize} & \boldsymbol{c}^\top \boldsymbol{x} & \text{(成本最小化，或 -利润最大化)} \\
\text{subject to} & \boldsymbol{G}\boldsymbol{x} \preceq \boldsymbol{h} & \text{(资源上限约束)} \\
 & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} & \text{(平衡/守恒约束)}
\end{array}
$$

- $\boldsymbol{x}$ 是决策变量；$\boldsymbol{G}\boldsymbol{x} \preceq \boldsymbol{h}$ 表示每一个分量都有不等式约束。
- 线性规划 (Linear Programming, LP) 是运筹学核心工具。

虽然没有直接解析解（像最小二乘那样），但现代求解器（如单纯形法/内点法）可以飞速解出线性规划。

练习： 某工厂生产两种产品 A 和 B。

- 每生产 1 单位产品 A 需要 2 小时加工时间和 1 单位原材料；
- 每生产 1 单位产品 B 需要 1 小时加工时间和 3 单位原材料；
- 工厂每天最多有 100 小时加工时间；
- 每天最多有 90 单位原材料；
- 每生产 1 单位产品 A 可获利润 40 元；
- 每生产 1 单位产品 B 可获利润 30 元。

请给出一个线性规划模型，来最大化工厂的利润。

目录

拓展：Chebyshev 逼近问题

最小二乘使用的是 L^2 误差；另一个很常见的目标是“控制最坏误差”：

$$\min_{c}\left(\max_{i=1,2,\cdots,N}|a_i c - b_i|\right)$$

其中为简化，我们仅考虑 $a_i, c, b_i \in \mathbb{R}$。

该问题可改写为一个线性规划：

$$\begin{aligned}
\text{minimize} \quad & t \\
\text{subject to} \quad & a_i c - t \le b_i, \qquad i = 1, 2, \cdots, N \\
& -a_i c - t \le -b_i, \qquad i = 1, 2, \cdots, N.
\end{aligned}$$

优化变量为 (t, c)。

信息： 同一拟合任务，若更换误差的衡量标准，优化模型也会随之改变。

前一页的证明

$$
\begin{aligned}
&\min_{c}\left(\max_{i=1,2,\cdots,N}|a_i c - b_i|\right)\\
=&\min_{c}\min_{t\geq|a_i c-b_i|,\ 1\leq i\leq N} t\\
=&\min_{t,c} t \qquad \text{subject to } t \geq |a_i c - b_i|, 1 \leq i \leq N
\end{aligned}
$$

不同向量的范数

目标：学习向量的 L^p 范数（含欧式距离与最大范数），并理解其建模含义。

对于一个向量 $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top$

- 内积 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_i x_i y_i$（记号可用尖括号或圆括号）
- $\|\boldsymbol{x}\|_2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2} = (\sum_i x_i^2)^{1/2}$ 为范数（模）

性质：

- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$
- Cauchy-Schwarz 不等式：$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$
- 三角不等式：$\|\boldsymbol{x} + \boldsymbol{y}\|_2 \leq \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2$

L^p 范数

把范数中下标 2 改为任何的 $p \in [1, \infty)$：

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$$ p-范数

例子：$\boldsymbol{x} = [1, 3]$。请计算 $\|\boldsymbol{x}\|_5$，$\|\boldsymbol{x}\|_{10}$，$\|\boldsymbol{x}\|_{20}$，观察 p 增大时结果如何变化。

L^p 范数

把范数中下标 2 改为任何的 $p \in [1, \infty)$：

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$ p-范数

例子：$\boldsymbol{x} = [1, 3]$。请计算 $\|\boldsymbol{x}\|_5$，$\|\boldsymbol{x}\|_{10}$，$\|\boldsymbol{x}\|_{20}$，观察 p 增大时结果如何变化。

答案：对该例子，当 $p \to \infty$ 时，$\|\boldsymbol{x}\|_p \to 3$（逐渐逼近最大分量）

L^p 范数

把范数中下标 2 改为任何的 $p \in [1, \infty)$：

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \quad \text{p-范数}$$

例子：$\boldsymbol{x} = [1, 3]$。请计算 $\|\boldsymbol{x}\|_5$, $\|\boldsymbol{x}\|_{10}$, $\|\boldsymbol{x}\|_{20}$，观察 p 增大时结果如何变化。

答案：对该例子，当 $p \to \infty$ 时，$\|\boldsymbol{x}\|_p \to 3$（逐渐逼近最大分量）

更一般的，当 $p \to \infty$，

$$\lim_{p\to\infty} \|\boldsymbol{x}\|_p \overset{\text{现象/结论}}{=} \max_i |x_i| \overset{\text{定义为}}{=} \|\boldsymbol{x}\|_\infty$$

常用的例子

- L^1 范数（曼哈顿距离）：$\sum |x_i|$。【在稀疏优化中非常关键】
- L^2 范数（欧氏距离）：$\sqrt{\sum x_i^2}$。【能量、物理场、普通的空间距离】
- L^∞ 范数（最大模）：$\max |x_i|$。【最坏情况分析】

不同范数的联系

练习： 证明 $\|\boldsymbol{x}\|_\infty \leqslant \|\boldsymbol{x}\|_p \leqslant n^{1/p}\|\boldsymbol{x}\|_\infty$，其中 n 是向量的维数。

更一般情况的范数定义

$\|\cdot\|$ 是一个函数：输入是一个向量 $\boldsymbol{x}$，输出 $\|\boldsymbol{x}\|$ 是一个非负实数。

定义 (向量的范数)

我们说它是一个范数如果它满足如下条件：

- $\|\boldsymbol{x}\| \geq 0$ ($\|\boldsymbol{x}\| = 0$ 当且仅当 $\boldsymbol{x} = 0$)
- $\|\alpha \boldsymbol{x}\| = |\alpha| \|\boldsymbol{x}\|$ 对于任何一个实数 α
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$

我们使用下标来注明这个范数是什么意义下的。

该量本质是用来描述向量的大小。

不同向量的范数 $\rightarrow$ 不同的优化问题

假设有数据 a_i, b_i，我们希望用线性函数拟合 $b = ca$。

- 若我们使用 L^∞ 范数，我们希望

  $$\min_c \max_{i=1,2,\cdots,N} |a_i c - b_i|,$$

  该问题等价为线性规划；（关注“最坏误差”）
- 若我们使用 L^2 范数，我们希望

  $$\min_c \sum_{i=1,2\cdots,N} |a_i c - b_i|^2$$

  该问题是典型最小二乘。（最常用，计算高效）

小结：同一份数据，不同范数对应不同“偏好”，也会导向不同算法。

目录

总结

- 理解最小二乘拟合的框架，以及能推导出最优解的表达式
- 理解线性规划的问题来源以及其一般形式

阅读作业 & 参考资料：

- 课本第 1 章
- 第 4.3 章（线性规划问题）

参考答案

- 第3页：C
- 第10页：参数量多于数据量将注定可以拟合成功，但其效果不一定好；会出现拟合的函数不唯一
- 设

$$x_1 = \text{每天生产产品}A\text{ 的数量}, \quad x_2 = \text{每天生产产品}B\text{ 的数量},$$

$$\begin{aligned} \text{maximize} \quad & 40x_1 + 30x_2 \\ \text{subject to} \quad & \begin{cases} 2x_1 + x_2 \leq 100 \\ x_1 + 3x_2 \leq 90 \\ x_1 \geq 0,\ x_2 \geq 0 \end{cases} \end{aligned}$$