第四章：应用

(1) 拟合问题

授课教师：曹语

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

为什么拟合不止一种?

在数据拟合中，我们常见的目标都可以写成

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \tag{1}$$

其中 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ 和 $\boldsymbol{b} \in \mathbb{R}^m$ 由数据给出，$\boldsymbol{x} \in \mathbb{R}^n$ 是待求参数。

但是不同问题的侧重点不同:

- 是更关心整体误差，还是最大误差?
- 如果数据中有少量野值，是否希望模型更鲁棒?
- 如果参数对数据很敏感，是否要加正则化?

学习目标

学完本节后，希望能够：

- 把简单的数据拟合写成 $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$，并理解残差；
- 知道 L^2、L^1、L^∞ 拟合分别强调什么；
- 知道 L^1、L^∞ 如何改写成线性规划；
- 理解罚函数 ϕ 的形状会怎样影响最优解；
- 知道为什么 L^1/Huber 更鲁棒，以及为什么正则化可能提高稳定性。

回顾练习：如何写成矩阵形式?

如果我们有一维数据 $\{(z_i, y_i)\}_{i=1}^m$，希望拟合直线

$$y = a_1 z + a_0,$$

如何把它写成

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|?$$

回顾练习：如何写成矩阵形式?

如果我们有一维数据 $\{(z_i, y_i)\}_{i=1}^m$，希望拟合直线

$$y = a_1 z + a_0,$$

如何把它写成

$$\min_{\boldsymbol{x}} \ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|?$$

答案是

$$\boldsymbol{x} = \begin{bmatrix} a_1 \\ a_0 \end{bmatrix}, \qquad \boldsymbol{A} = \begin{bmatrix} z_1 & 1 \\ z_2 & 1 \\ \vdots & \vdots \\ z_m & 1 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

基本记号

我们后续默认 $\boldsymbol{A}$ 的列线性独立。

讨论： 如果 $\boldsymbol{A}$ 的列线性相关，会发生什么？

我们定义残差

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$$

后面各种模型的主要差别，就在于我们如何度量残差。

目录

（1）线代角度

$$\boldsymbol{A}\boldsymbol{x} = x_1\bar{\boldsymbol{a}}_1 + \cdots + x_n\bar{\boldsymbol{a}}_n$$

其中，$\bar{\boldsymbol{a}}_1, \cdots, \bar{\boldsymbol{a}}_n$ 是矩阵 $\boldsymbol{A}$ 的列。

所以范数逼近问题是用一些既有向量的线性组合，来逼近某个向量 $\boldsymbol{b}$

（2）统计角度

假设观测值是 $\boldsymbol{a}_i \in \mathbb{R}^n$，待估计值是 $\boldsymbol{x} \in \mathbb{R}^n$，并且假设误差模型为线性

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{x} + v_i$$

这里把观测向量记为 $\boldsymbol{y}$；它与前面抽象模型中的 $\boldsymbol{b}$ 扮演同样的角色。

如果我们假设误差 v_i 独立同分布（IID），且其分布为 p，则该观测数据的概率密度为 $p_{\boldsymbol{x}}(\boldsymbol{y}) = \prod_{i=1}^m p\big(y_i - \boldsymbol{a}_i^\top \boldsymbol{x}\big)$。我们希望找到 $\boldsymbol{x}$ 最大化概率密度，即**最大似然**

$$\max_{\boldsymbol{x}} \; p_{\boldsymbol{x}}(\boldsymbol{y}) \;\longleftrightarrow\; \max_{\boldsymbol{x}} \; \log p_{\boldsymbol{x}}(\boldsymbol{y}) \;\longleftrightarrow\; \max_{\boldsymbol{x}} \sum_{i=1}^m \log p\big(y_i - \boldsymbol{a}_i^\top \boldsymbol{x}\big).$$

统计角度：高斯误差对应最小二乘

如果我们进一步知道误差是正态分布（标准差为 σ），则最大似然等价于最小化残差平方和：

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left| y_i - \boldsymbol{a}_i^\top \boldsymbol{x} \right|^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

其中，

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

提问： 若误差分布为无偏的 Laplace 分布，即 $p(z) = \frac{1}{2b} e^{-|z|/b}$，则上述问题会变成什么？

（3）几何角度

我们引入符号 $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x}$，因此 $\boldsymbol{u}$ 的所有范围是

$$\boldsymbol{u} \in \text{Range}(\boldsymbol{A}) \qquad \text{（矩阵}\boldsymbol{A}\text{的值域）}$$

公式(1)的问题等价于

$$\begin{array}{ll} \text{minimize} & \|\boldsymbol{u} - \boldsymbol{b}\| \\ \text{subject to} & \boldsymbol{u} \in \text{Range}(\boldsymbol{A}) \end{array}$$

它具有什么几何含义?

（3）几何角度

我们引入符号 $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x}$，因此 $\boldsymbol{u}$ 的所有范围是

$$\boldsymbol{u} \in \text{Range}(\boldsymbol{A}) \qquad （矩阵\boldsymbol{A}的值域）$$

公式(1)的问题等价于

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{u} - \boldsymbol{b}\| \\ &\text{subject to} && \boldsymbol{u} \in \text{Range}(\boldsymbol{A}) \end{aligned}$$

即 $\boldsymbol{b}$ 到 $\text{Range}(\boldsymbol{A})$ 的最小距离。

目录

加权问题

比如之前的线性数据拟合问题，如果不同的数据的可信度 w_i 有所不同，则我们或许希望优化

$$\sum_{i=1}^{m} |w_i(a_1 z_i + a_0 - y_i)|^2$$

这里默认 $w_i > 0$；w_i 越大，表示该数据点在拟合中被看得越重要。

该问题还是一般的范数逼近问题：

$$\text{minimize} \quad \|\boldsymbol{W}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2$$

其中 $\boldsymbol{W} = \text{diag}\left(\begin{bmatrix} w_1 & w_2 & \cdots & w_m \end{bmatrix}\right)$。

L^2-逼近

$$
\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \\
& = r_1^2 + r_2^2 + \cdots + r_m^2 \\
& = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{b}
\end{aligned}
$$

（无约束的）最优性条件告诉我们，

$$
\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{A}^\top \boldsymbol{b} = 0 \qquad \Longrightarrow \qquad \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\top \boldsymbol{b}
$$

若 $\boldsymbol{A}^\top \boldsymbol{A}$ 可逆，则

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

L^∞-逼近（Chebyshev 逼近问题）

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_\infty = \max\{|r_1|,\cdots,|r_m|\}$$

回顾练习： 为何该函数是凸函数?

该情况等价于

$$\begin{array}{ll}\text{minimize} & t \\ \text{subject to} & -t\mathbf{1} \preceq \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b} \preceq t\mathbf{1},\end{array}$$

变量为 $\boldsymbol{x}\in\mathbb{R}^n$ 和 $t\in\mathbb{R}$

L^1-逼近

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1 = |r_1| + \cdots + |r_m|$$

该形式被称为**鲁棒估计器**（原因在后面会解释）

该问题等价于

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^\top \boldsymbol{t} \\ \text{subject to} & -\boldsymbol{t} \preceq \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \preceq \boldsymbol{t} \end{array}$$

变量为 $\boldsymbol{x} \in \mathbb{R}^n$ 和 $\boldsymbol{t} \in \mathbb{R}^m$

练习：请说明为什么？

一般的形式

对于 $1 \leqslant p < \infty$, L^p – 范数逼近问题的目标函数为

$$(|r_1|^p + \cdots + |r_m|^p)^{1/p}, \qquad \text{或者等价于} \qquad |r_1|^p + \cdots + |r_m|^p.$$

该形式可以更一般变成

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & \boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \end{array}$$

$\phi : \mathbb{R} \to \mathbb{R}$ 称为**（残差）罚函数**

$\phi(u)$ **刻画了我们不喜欢残差值** u **的程度**

例子：L^p 与 deadzone

- L^p 范数逼近 $\phi(u) = |u|^p$（$p \geqslant 1$）
  - (L^2) 二次罚函数 $\phi(u) = u^2$
  - (L^1) 绝对值罚函数 $\phi(u) = |u|$

- 带有死区的线性罚函数（$a > 0$ ）

$$\phi(u) = \begin{cases} 0 & |u| \leqslant a \\ |u| - a & |u| > a \end{cases}$$

- deadzone 适合“允许小误差存在，但不希望大误差出现”的场景。

例子：对数障碍罚函数

- 对数障碍罚函数（$a > 0$）

$$\phi(u) = \begin{cases} -a^2 \log\left(1 - (u/a)^2\right) & |u| < a \\ \infty & |u| \geqslant a \end{cases}$$

- 它表达的是：误差不能超过阈值 a。

参与式观察：从罚函数形状猜结果

课堂观察： 如果只看 $\phi(u)$ 的图像，你会猜哪种模型更偏好哪类残差？

- u^2：对大残差惩罚很重；
- $|u|$：对大残差的增长较慢，因此通常更鲁棒；
- $|u|^p$（p 较大）：倾向于让所有残差都比较小；
- deadzone：小残差可以被“忽略”；
- log barrier：残差不能超过给定阈值。

数值实验： 见配套代码。

对于野值或大误差的灵敏性

有时由于设备的误差或者实验测量的问题，具有一些误差很大的测量值：$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$，其中误差 $\boldsymbol{\epsilon}$ 的部分分量很大。如何降低这些大的分量的影响?

对于野值或大误差的灵敏性

有时由于设备的误差或者实验测量的问题，具有一些误差很大的测量值：$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$，其中误差 $\boldsymbol{\epsilon}$ 的部分分量很大。如何降低这些大的分量的影响?

一个可能的选择如下：

$$\phi(u) = \begin{cases} u^2 & |u| \leqslant M; \\ M^2 & |u| > M. \end{cases}$$

该函数是非凸的。

对于野值或大误差的灵敏性

有时由于设备的误差或者实验测量的问题，具有一些误差很大的测量值：$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$，其中误差 $\boldsymbol{\epsilon}$ 的部分分量很大。如何降低这些大的分量的影响?

如果我们要求凸函数，我们可以选择 $\phi(u) = |u|$，或者**Huber 罚函数**，

$$\phi_{\mathsf{hub}}(u) = \begin{cases} u^2 & |u| \leqslant M \\ M(2|u| - M) & |u| > M \end{cases}$$

鲁棒性：课堂结论

- 如果数据里有少量野值，希望减小它们的影响，通常选择 L^1 或 Huber 罚函数；
- L^2 对大残差惩罚更重，因此更容易被少量异常值“拉偏”；
- 如果我们希望所有残差都相对均匀地小，可以考虑 L^p（其中 p 比较大）或 L^∞ 型目标。

目录

正则化

- 我们希望 $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$，以及 $\|\boldsymbol{x}\|$ 同时比较小，则可以考虑如下的正则化形式：

$$\text{minimize } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| + \gamma\|\boldsymbol{x}\|,$$

  其中 $\gamma > 0$ 为问题参数。

- 第一项要求“拟合数据”，第二项要求“参数不要太大”。
- 这里先写成一般形式；后面重点讨论最常见的 L^2 型正则化。

正则化：L^2 型情形

- 若我们考虑 L^2 距离，则一种很常见的形式是

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \delta\|\boldsymbol{x}\|_2^2$$

该问题等价于优化

$$\min \ \boldsymbol{x}^\top\big(\boldsymbol{A}^\top\boldsymbol{A} + \delta\boldsymbol{I}\big)\boldsymbol{x} - 2\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top\boldsymbol{b}$$

因此，最优解为

$$\boldsymbol{x}^* = \underline{\qquad\qquad ? \qquad\qquad}$$

正则化：闭式解与稳定性

答案：$\boldsymbol{x}^* = \left(\boldsymbol{A}^\top \boldsymbol{A} + \delta \boldsymbol{I}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{b}$

场景：若 $\boldsymbol{A}$ 接近奇异矩阵，原问题往往不稳定；加入正则项后会更稳定。

例子：令

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1+\epsilon & 1 \end{bmatrix}, \qquad \epsilon \approx 0.$$

当 $\epsilon = 0$ 时，该矩阵奇异。

对于数据 $\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$，没有加正则化的问题的真解为 $\boldsymbol{x}^* = \begin{bmatrix} \frac{b_2 - b_1}{\epsilon} \\ b_1 + \frac{b_1 - b_2}{\epsilon} \end{bmatrix}$ 若 $\boldsymbol{b}$ 有测量误差，而且 ϵ 很小，比如 $\approx 10^{-3}$，则结果会很不稳定。

正则化：稳定性例子

如果加了正则项，则最优解为

$$\boldsymbol{x}^*_{\text{reg}} = \frac{1}{c}\begin{bmatrix}(\delta-\epsilon)b_1 + (\delta+\epsilon+\delta\epsilon)b_2 \\ b_2(\delta-\epsilon) + b_1(\delta+\epsilon+\epsilon^2)\end{bmatrix}, \qquad c = \delta^2+\epsilon^2+\delta(4+2\epsilon+\epsilon^2)$$

关键点是：与原问题相比，正则化后的系数不再出现 $1/\epsilon$ 型的爆炸项。

正则化：稳定性例子（结论）

考虑 $\epsilon = 10^{-3}$，$\delta = 10^{-2}$。

若数据 b_1 具有误差 10^{-2}，则原问题中 b_1 对于第一个分量造成的误差高达 -10，而正则化中，数据扰动对于最优解的误差仅为

$$\frac{\delta - \epsilon}{\delta^2 + \epsilon^2 + \delta(4 + 2\epsilon + \epsilon^2)} \times 10^{-2} \approx 0.22 \times 10^{-2}$$

因此，增加正则项的优化问题更加稳定。

→ 对于模型训练参数的正则化，在深度学习中也被广泛应用。

目录

课内练习

请回答下面问题：

1. 如果数据里有少量很大的测量误差，且你希望优化结果更鲁棒，你会优先考虑 L^1 还是 L^2？为什么？
2. 请把 L^∞-逼近写成一个线性规划。
3. 为什么在 $\boldsymbol{A}$ 接近奇异时，加入 $\delta\|\boldsymbol{x}\|_2^2$ 会让解更稳定？

总结

主要需要了解的内容：

- 能把简单拟合写成 $\boldsymbol{Ax} - \boldsymbol{b}$，并理解残差。
- 知道 L^∞ 和 L^1 分别对应什么线性规划。
- 知道罚函数会影响最优解，且 L^1/Huber 更鲁棒。
- 知道正则化的含义，以及它可能提高稳定性。

阅读：课本第 6.1–6.3 章