第四章：应用

(1) 逼近与拟合

曹语

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

背景和目标

在 1d 数据拟合问题中，我们已经使用过如下的范数逼近问题：

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

其中 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ 来源于问题的数据；$\boldsymbol{x} \in \mathbb{R}^n$ 是优化变量。

目标： 由于这类问题的广泛应用，将更加系统地看该问题的多种变化。

回顾测试： 如果我们有一维数据 $\left\{(z_i, y_i)\right\}_{i=1}^m$，我们希望拟合 $y = a_1 z + a_0$，该 $\boldsymbol{A}$ 和 $\boldsymbol{b}$ 应该是什么？

我们后续会假设 $\boldsymbol{A}$ 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

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

目录

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 $\boldsymbol{b}$

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

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

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**

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

如果我们进一步知道误差是正态分布（标准差为 σ），则该问题等价于

$$\min_{\boldsymbol{x}} \ \big|y_i - \boldsymbol{a}_i^\top \boldsymbol{x}\big|^2 = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

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?

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

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

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

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

该问题还是一般的范数逼近问题：

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

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

L^2-逼近

$$
\begin{aligned}
\text{minimize} \quad & \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \\
& = \boldsymbol{r}_1^2 + \boldsymbol{r}_2^2 + \cdots + \boldsymbol{r}_m^2 \\
& = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{b}
\end{aligned}
$$

（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）

$$\text{minimize} \quad \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_\infty = \max\{|\boldsymbol{r}_1|, \cdots, |\boldsymbol{r}_m|\}$$

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

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

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

$$(|\boldsymbol{r}_1|^p + \cdots + |\boldsymbol{r}_m|^p)^{1/p}, \qquad \text{或者等价于} \qquad |\boldsymbol{r}_1|^p + \cdots + |\boldsymbol{r}_m|^p.$$

该形式可以更一般变成

$$\begin{array}{ll} \text{minimize} & \phi(\boldsymbol{r}_1) + \cdots + \phi(\boldsymbol{r}_m) \\ \text{subject to} & \boldsymbol{r} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \end{array}$$

$\phi : \mathbb{R} \to \mathbb{R}$ 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 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（$a > 0$ ）

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

- 对数障碍罚函数（$a > 0$)

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

数值实验： 见代码部分

提示：

- 对于两个罚函数，ϕ_1 和 $\phi_2 = c\phi_1$，两者的效果是没有区别的。
- 惩罚函数的形状对于优化结果具有很大的影响。

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?

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**，

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

【数值实验结果见代码部分】

- 如果我们要求对于野值的鲁棒性高，选择 L^1（或者 Huber 罚函数）
- 一般而言，L^2 是一个好的选择
- 如果我们希望残差值一致相对比较小，可以选择 L^p（其中 p 比较大）

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

- 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

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

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

应用场景：比如若 $\boldsymbol{A}$ 接近奇异矩阵时，求解原问题会比较不稳定；使用了正则化后会变稳定。

例子：$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1+\epsilon & 1 \end{bmatrix}$，其中 $\epsilon \approx 0$。当 $\epsilon = 0$ 时候，该矩阵是奇异矩阵

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

$$\boldsymbol{x}^* = \begin{bmatrix} \frac{b_2 - b_1}{\epsilon} \\ b_1 + \frac{b_1 - b_2}{\epsilon} \end{bmatrix}$$

如果数据 $\boldsymbol{b}$ 有一些测量误差，而且 ϵ 很小，比如 $\approx 10^{-3}$，则结果很不稳定。

如果加了正则项，则最优解为

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

更加具体的，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

其他正则化形式（不要求）

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

其中矩阵

$$\boldsymbol{D} = n^2 \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(n-2)\times n},$$

应用场景：求物理问题

目录

总结

主要需要了解的内容：

- 对于 L^∞ 和 L^1 逼近，能知道它对应什么线性规划问题。
- 知道罚函数的设计对于最优解的影响，知道 L^1 范数作为罚函数可以降低野值的影响（相比于 L^2）。
- 知道正则化的含义，以及正则化可能可以增加稳定性。

阅读作业 & 参考资料：

- 课本第 6.1 - 6.3 章