5.7 最小二乘问题

SMaLL

[1] 中国石油大学（华东）
SMaLL 课题组
small.sem.upc.edu.cn
liangxijunsd@163.com

2023

最小二乘问题

1. 最小二乘法
2. 高斯-牛顿（Gauss-Newton）法
3. 增量梯度方法

最小二乘法

$$\begin{array}{ll} \text{minimize} & f(x) = \frac{1}{2}\|g(x)\|^2 = \frac{1}{2}\sum_{i=1}^{m}\|g_i(x)\|^2 \\ \text{subject to} & x \in \Re^n \end{array}$$

其中 g 是一个连续可微函数，具有分量函数 $g_1, \ldots, g_m$ ，其中 $g_i : \Re^n \to \Re^{r_i}$.

示例

假设模型：

$$z = h(\theta, x).$$

目标. 估计 n 个参数 $\theta \in \mathbb{R}^n$.

- h 是已知的能够刻画这个模型的函数
- $\theta \in \Re^n$ 是未知参数的向量
- $x \in \Re^p$ 是模型的输入
- $z \in \Re^r$ 是模型的输出

示例

- 数据. m 个输入-输出对 $(x_1, z_1), \ldots, (x_m, z_m)$
- 模型. 最小化误差平方和

$$\frac{1}{2}\sum_{i=1}^{m}\|z_i - h(\theta, x_i)\|^2$$

- E.g., 用三次多项式近似拟合数据:

$$h(\theta, x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$$

其中 $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)$ 是未知系数的向量.

最小二乘问题

1. 最小二乘法
2. 高斯-牛顿（Gauss-Newton）法
3. 增量梯度方法

Sec. 2 高斯-牛顿法

注意下面的迭代格式中, x^k 表示迭代点，对应最小二乘拟合问题中的参数，并不是样本点。

- 给定迭代点 x^k, 高斯-牛顿迭代法的基本思想是用如下函数的线性化函数近似 g:

$$\tilde{g}(x, x^k) = g(x^k) + \nabla g(x^k)^T (x - x^k)$$

- 最小化线性化函数 $\tilde{g}$ 的范数:

$$\begin{aligned} x^{k+1} &= \arg\min_{x \in \Re^n} \frac{1}{2} \|\tilde{g}(x, x^k)\|^2 \\ &= \arg\min_{x \in \Re^n} \frac{1}{2} \{\|g(x^k)\|^2 + 2(x - x^k)^T \nabla g(x^k) g(x^k) \\ &\quad + (x - x^k)^T \nabla g(x^k) \nabla g(x^k)^T (x - x^k)\}. \end{aligned}$$

高斯-牛顿法

假设 $\nabla g\left(x^{k}\right) \nabla g\left(x^{k}\right)^{T} \in \mathbb{R}^{n \times n}$ 是可逆的:

$$x^{k+1}=x^{k}-\left(\nabla g\left(x^{k}\right) \nabla g\left(x^{k}\right)^{T}\right)^{-1} \nabla g\left(x^{k}\right) g\left(x^{k}\right) \tag{1}$$

- 如果 g 本身是线性函数 $\Rightarrow \|g(x)\|^{2}=\left\|\tilde{g}\left(x, x^{k}\right)\right\|^{2}$, 则该方法会在一次迭代后收敛。
- 方向
$$-\left(\nabla g\left(x^{k}\right) \nabla g\left(x^{k}\right)^{T}\right)^{-1} \nabla g\left(x^{k}\right) g\left(x^{k}\right)$$
是下降方向因为 $\nabla g\left(x^{k}\right) g\left(x^{k}\right)=\nabla\left(0.5\|g(x)\|^{2}\right)|_{x=x^{k}}$ 且 $\left(\nabla g\left(x^{k}\right) \nabla g\left(x^{k}\right)^{T}\right)^{-1}$ 是正定矩阵.

SMaLL Statistical Machine Learning Lyceum

高斯-牛顿法

为了确保矩阵 $\nabla g\left(x^k\right) \nabla g\left(x^k\right)^T$ 是奇异矩阵（或接近奇异）时该方法也有效，迭代公式修正为:

$$x^{k+1} = x^k - \alpha^k \left(\nabla g\left(x^k\right) \nabla g\left(x^k\right)^T + \Delta^k\right)^{-1} \nabla g\left(x^k\right) g\left(x^k\right)$$

其中 Δ^k 是一个对角矩阵，使得:

$$\nabla g\left(x^k\right) \nabla g\left(x^k\right)^T + \Delta^k \text{ 为正定矩阵.}$$

- 高斯-牛顿法所使用的方向与梯度相关并且符合梯度下降法的收敛结果。

与牛顿法的关系

- 假定每个 g_i 是一个标量函数,

$$\nabla^2\left(0.5\|g(x)\|^2\right) = \nabla g\left(x^k\right)\nabla g\left(x^k\right)^T + \sum_{i=1}^{m}\nabla^2 g_i\left(x^k\right) g_i\left(x^k\right)$$

- 在高斯-牛顿法中 "近似的 Hessian 矩阵为":

$$\nabla g\left(x^k\right)\nabla g\left(x^k\right)^T.$$

- 高斯-牛顿法的迭代公式:

$$x^{k+1} = x^k - \left(\nabla g\left(x^k\right)\nabla g\left(x^k\right)^T\right)^{-1}\nabla g\left(x^k\right) g\left(x^k\right) \tag{2}$$

- 优点: 它简化了计算.
- 缺点: 收敛速度较慢.

与牛顿法的关系

- 如果被忽略项 $\sum_{i=1}^{m} \nabla^2 g_i\left(x^k\right) g_i\left(x^k\right)$ 在解附近很小 $\rightarrow$ 良好的收敛速度
- E.g.1 当 g 接近线性时, 或者当分量 $g_i(x)$ 在解附近很小时.
- E.g.2 $g(x)=0,\ m=n.$
  $\rightarrow$ 被忽略的项在解处接近零。
- 假定 $\nabla g\left(x^k\right)$ 是可逆的,

$$\left(\nabla g\left(x^k\right) \nabla g\left(x^k\right)^T\right)^{-1} \nabla g\left(x^k\right) g\left(x^k\right)=\left(\nabla g\left(x^k\right)^T\right)^{-1} g\left(x^k\right)$$

- 高斯-牛顿法的迭代公式化简为:

$$x^{k+1}=x^k-\left(\nabla g\left(x^k\right)^T\right)^{-1} g\left(x^k\right) \tag{3}$$

最小二乘问题

1. 最小二乘法
2. 高斯-牛顿（Gauss-Newton）法
3. 增量梯度方法

Sec. 3 增量梯度方法

- 每个分量 g_i: 一个数据块.
- $x^k \to x^{k+1}$: 数据块的循环
- 初始: $\psi_0 = x^k$
- 进行 m 步循环:

$$\psi_i = \psi_{i-1} - \alpha^k \nabla g_i(\psi_{i-1})\, g_i(\psi_{i-1}), \quad i = 1, \ldots, m$$

其中 $\alpha^k > 0$ 为步长, 方向为第 i 个数据块的梯度:

$$\nabla \left(\frac{1}{2} \|g_i(x)\|^2 \right) \Big|_{x=\psi_{i-1}} = \nabla g_i(\psi_{i-1})\, g_i(\psi_{i-1})$$

SMaLL Statistical Machine Learning Lyceum

增量梯度方法

该方法可以写作如下形式

$$x^{k+1} = x^k - \alpha^k \sum_{i=1}^{m} \nabla g_i\left(\psi_{i-1}\right) g_i\left(\psi_{i-1}\right) \tag{4}$$

不同点

- 增量梯度方法的方向:

$$-\sum_{i=1}^{m} \nabla g_i\left(\psi_{i-1}\right) g_i\left(\psi_{i-1}\right)$$

- 梯度法的方向:

$$-\nabla f(x^k) = -\sum_{i=1}^{m} \nabla g_i\left(x^k\right) g_i\left(x^k\right)$$

增量梯度方法

优势：

- 适合数据流场景;
- 每一轮参数的更新速度都大大加快.

练习题

1. 编程. 生成 m 个输入-输出对 $(x_1, z_1), \ldots, (x_m, z_m)$, 其中 $x_i \in \mathbb{R}$, 考虑三次多项式近似拟合数据的最小化误差平方和模型:

$$\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} \|z_i - h(\theta, x_i)\|^2 \tag{5}$$

其中 h 为三次多项式 $h(\theta, x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$, 编程实现高斯-牛顿法, 求解上述最小化误差平方和模型, 计算拟合误差。