线性优化与最小二乘

刘士新

(sxliu@mail.neu.edu.cn)

信息科学与工程学院　机器学习与智能决策研究所

2019.11

7　最小二乘

7.1 最小二乘问题及求解

最小二乘问题形式如下：

$$\min \ \|\boldsymbol{Ax}-\boldsymbol{b}\|^2 \tag{7.1}$$

其中，$\boldsymbol{A}$ 是一个高的 $m\times n\ (m>n)$ 的矩阵，因此，$\boldsymbol{Ax}=\boldsymbol{b}$ 是超定的。对于大多数 $\boldsymbol{b}$，没有 $\boldsymbol{x}$ 满足 $\boldsymbol{Ax}=\boldsymbol{b}$。最小二乘问题寻找 $\boldsymbol{x}$ 使得 $\|\boldsymbol{Ax}-\boldsymbol{b}\|^2$ 最小。

最小二乘问题的列解释：记 $\boldsymbol{A}_j$ 为矩阵 $\boldsymbol{A}$ 的第 j 列，则

$$\|\boldsymbol{Ax}-\boldsymbol{b}\|^2=\|(x_1\boldsymbol{A}_1+x_2\boldsymbol{A}_2+\cdots+x_n\boldsymbol{A}_n)-\boldsymbol{b}\|^2。$$

最小二乘问题可以理解为寻找矩阵 $\boldsymbol{A}$ 的列向量的一个最接近 $\boldsymbol{b}$ 的线性组合。

最小二乘问题的行解释：记 $\boldsymbol{a}_i^T$ 为矩阵 $\boldsymbol{A}$ 的第 i 行，残差 $r_i=\boldsymbol{a}_i^T\boldsymbol{x}-b_i$，则

$$\|\boldsymbol{Ax}-\boldsymbol{b}\|^2=(\boldsymbol{a}_1^T\boldsymbol{x}-b_1)^2+\cdots+(\boldsymbol{a}_m^T\boldsymbol{x}-b_m)^2。$$

最小二乘问题可以理解为求解残差平方和最小的优化问题。

假设矩阵 $\boldsymbol{A}$ 的列向量是互相独立的，则 Gram 矩阵 $\boldsymbol{A}^T\boldsymbol{A}$ 是可逆的。定义

$$f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j - b_i\right)^2,$$

则 $f(\boldsymbol{x})$ 的最优解 $\hat{\boldsymbol{x}}$ 必须满足

$$\frac{\partial f}{\partial x_k}(\hat{\boldsymbol{x}}) = \nabla f(\hat{\boldsymbol{x}})_k = 0,\ k = 1,\ \cdots,\ n,$$

其中，

$$\begin{aligned}
\nabla f(\boldsymbol{x})_k &= \frac{\partial f}{\partial x_k}(\boldsymbol{x}) \\
&= \sum_{i=1}^{m} 2\left(\sum_{j=1}^{n} A_{ij}x_j - b_i\right)(A_{ik}) \\
&= \sum_{i=1}^{m} 2(\boldsymbol{A}^T)_{ki}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_i \\
&= \left(2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right)_k。
\end{aligned}$$

最优解 $\hat{\boldsymbol{x}}$ 满足，

$$\nabla f(\hat{\boldsymbol{x}}) = 2\boldsymbol{A}^T(\boldsymbol{A}\hat{\boldsymbol{x}} - \boldsymbol{b}) = \boldsymbol{0} \quad \rightarrow \quad \boldsymbol{A}^T\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{b},$$

则，

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} = \boldsymbol{A}^{\dagger}\boldsymbol{b}。$$

矩阵 $\boldsymbol{A} \in R^{m\times n}$ 的列向量是互相独立的。因此，$\boldsymbol{A}$ 可以分解为 $\boldsymbol{A} = \boldsymbol{QR}$，其中，$\boldsymbol{Q} \in R^{m\times n}$为正交矩阵（$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$），$\boldsymbol{R} \in R^{n\times n}$ 为上三角矩阵。于是，

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{QR})^T(\boldsymbol{QR}) = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{QR} = \boldsymbol{R}^T\boldsymbol{R}$$

$$\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T = (\boldsymbol{R}^T\boldsymbol{R})^{-1}(\boldsymbol{QR})^T = \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}\boldsymbol{R}^T\boldsymbol{Q}^T = \boldsymbol{R}^{-1}\boldsymbol{Q}^T。$$

采用矩阵 $\boldsymbol{QR}$ 分解的最小二乘问题求解算法如下：

步骤 1：计算矩阵 $\boldsymbol{A}$ 的 $\boldsymbol{QR}$ 分解，$\boldsymbol{A} = \boldsymbol{QR}$。

步骤 2：计算 $\boldsymbol{Q}^T\boldsymbol{b}$。

步骤 3：求解 $\boldsymbol{R}\hat{\boldsymbol{x}} = \boldsymbol{Q}^T\boldsymbol{b}$，获得 $\hat{\boldsymbol{x}}$。

以上算法的复杂性为 $O(2mn^2)$。

7.2 最小二乘数据拟合

7.2.1 特征工程

假设原始特征为 $\boldsymbol{x} \in R^n$ 向量。特征工程是选择基函数 $f_i: R^n \to R$，$i = 1$，$\cdots$，p，然后再基于基函数 $f_i(\boldsymbol{x})$，$i = 1$，$\cdots$，p，进行数据拟合，其中，p 可以$< n$、$= n$、$> n$。

例如，按如下方法选择基函数：

$$f_1(\boldsymbol{x}) = 1，\ f_i(\boldsymbol{x}) = x_{i-1}，\ i = 2, \cdots, n+1$$

则线性回归模型形式如下：

$$\begin{aligned} \hat{y} &= \theta_1 f_1(\boldsymbol{x}) + \theta_2 f_2(\boldsymbol{x}) + \cdots + \theta_{n+1} f_{n+1}(\boldsymbol{x}) \\ &= \theta_1 + \theta_2 x_1 + \cdots + \theta_{n+1} x_n \end{aligned}$$

常用的基函数形式：

（1）标准化：

$$f_1(\boldsymbol{x}) = 1，\ f_i(\boldsymbol{x}) = (x_{i-1} - \mu_{i-1})/\sigma_{i-1}，\ i = 2，\ \cdots，\ n+1$$

其中，μ_i 是均值，σ_i 是标准差。

（2）对数变换：如果 $x_i \geq 0$ 且波动范围很大，则

$$\tilde{x}_i = \log(1 + x_i)。$$

（3）截尾处理：如果数据绝对值很大且认为超出合理范围，则设定上下限阈值 a_i、b_i，

$$\tilde{x}_i = \begin{cases} a_i, & x_i < a_i \\ x_i, & a_i \le x_i \le b_i \\ b_i, & x_i > b_i \end{cases}$$

（4）分类（标签）特征：

x_i	$f_1(x_i)$	$f_2(x_i)$
-1	1	0
0	0	0
1	0	1

（5）分段线性函数：

包含 k 个结点 $a_1 < a_2 < \cdots < a_k$ 的任意连续分段线性函数均可由以下 $p = k + 2$ 个基函数描述：

$$f_1(x) = 1，\ f_2(x) = x，\ f_{i+2}(x) = (x - a_i)_+，\ i = 1，2，\cdots，k，$$

其中，$(u)_+ = \max\{u，\ 0\}$。如图 7.1 所示。基于分段线性函数的回归效果如图 7.2 所示。

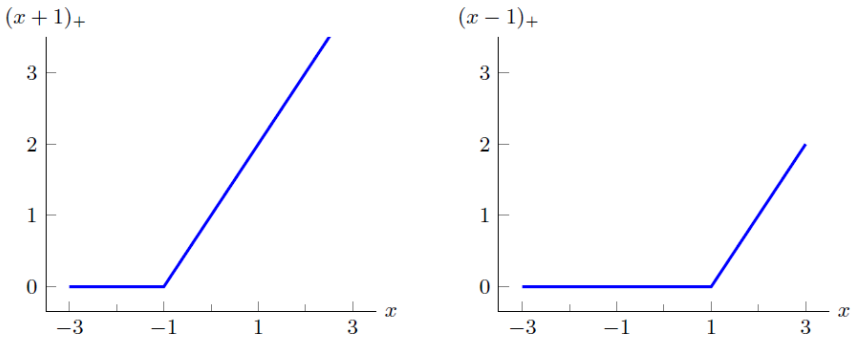

图 7.1 分段线性函数 $(x+1)_+ = \max\{x+1，\ 0\}$ 和 $(x-1)_+ = \max\{x-1，\ 0\}$

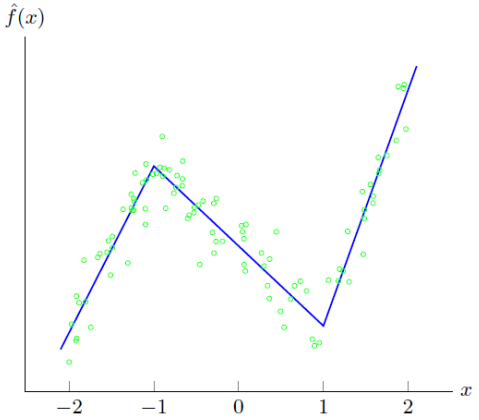

图 7.2 分段线性函数回归效果

（6）广义加性模型：

$$f_i(x_i) = x_i,$$
$$f_{n+i}(x_i) = \min\{x_i + a, 0\},$$
$$f_{2n+i}(x_i) = \max\{x_i - b, 0\},$$

其中，a、b 为参数，则基函数可以定义为：

$$\begin{aligned}\psi_i(x_i) &= \theta_i f_i(x_i) + \theta_{n+i} f_{n+i}(x_i) + \theta_{2n+i} f_{2n+i}(x_i) \\ &= \theta_i x_i + \theta_{n+i} \min\{x_i + a, 0\} + \theta_{2n+i} \max\{x_i - b, 0\}\end{aligned}$$

此时，回归模型的形式为：

$$\hat{y} = \psi_1(x_1) + \cdots + \psi_n(x_n)$$

模型有 $3n$ 个参数。在选用此类基函数时，通常先对 x_i 进行标准化处理，并取 $a = b = 1$。

（7）交叉项或多项式特征：$x_i x_j$，x_i^2，x_i^3，…

（8）其他预测模型的预测值：将其他模型的预测值作为当前模型的输入特征使用。

（9）与聚类中心的距离：对数据进行聚类分析，获得 k 个聚类，聚类中

心分别为 z_1，z_2，…，z_k。定义特征 $f(x)$ 为

$$f(x) = e^{-\|x-z_i\|/\sigma^2}$$

其中，σ 为参数。

（10）随机特征：构造原始特征随机线性组合的非线性函数作为新特征。如果想构造 K 个随机特征，首先生成随机矩阵 $\boldsymbol{R} \in R^{K\times n}$，然后将 $(\boldsymbol{Rx})_+$ 或 $|\boldsymbol{Rx}|$ 作为 K 个随机特征。也可以使用 $(\cdot)_+$ 和 $|\cdot|$ 以外的其他非线性函数。虽然随机特征违反直觉，但在某些应用中非常有效。

（11）定制化特征：根据专业知识和经验构建新特征。

7.2.2 最小二乘数据拟合算例

例 7.1 销售量预测（数据源：Advertising.csv）。用最小二乘数据拟合预测销售量 y 和电视、广播、报纸三种广告投入之间关系。线性回归模型形式如下：

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

表 7.1 广告数据集（Advertising.csv）

NO.	$\boldsymbol{x}=(x_1, x_2, x_3)$			y
	TV	Radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
…	…	…	…	…
200	232.1	8.6	8.7	13.4

$$\boldsymbol{A}=\begin{bmatrix}1 & 230.1 & 37.8 & 69.2\\ 1 & 44.5 & 39.3 & 45.1\\ \vdots & \vdots & \vdots & \vdots\\ 1 & 232.1 & 8.6 & 8.7\end{bmatrix}$$

$$\boldsymbol{b}=[22.1 \quad 10.4 \quad \cdots \quad 13.4]^T$$

求解最小二乘问题可得系数为：$\theta_0=2.93889$，$\theta_1=0.0457646$，$\theta_2=0.18853$，$\theta_3=-0.00103749$。

例 7.2 多项式数据拟合。选取基函数 $f_i(x) = x^{i-1}$，$i = 1$，$\cdots$，p，则数据拟合模型为

$$\hat{y} = \theta_1 + \theta_2 x^1 + \theta_3 x^2 + \cdots + \theta_p x^{p-1}$$

$$\boldsymbol{A} = \begin{bmatrix} 1 & x^{(1)} & \left(x^{(1)}\right)^2 & \cdots & \left(x^{(1)}\right)^{p-1} \\ 1 & x^{(2)} & \left(x^{(2)}\right)^2 & \cdots & \left(x^{(2)}\right)^{p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x^{(N)} & \left(x^{(N)}\right)^2 & \cdots & \left(x^{(N)}\right)^{p-1} \end{bmatrix}$$

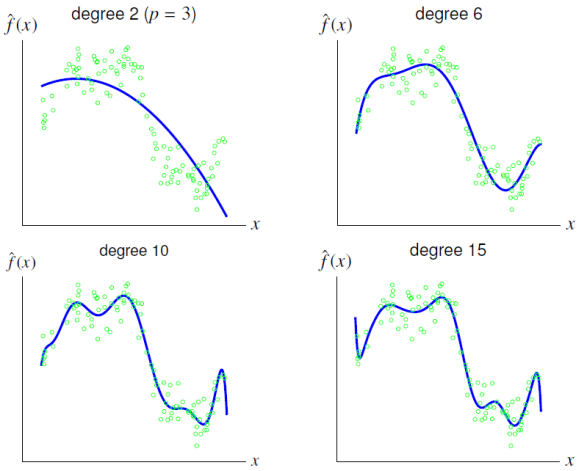

图 7.3 $N = 100$ 个数据点的拟合效果

例 7.3 商品房价格预测（数据源 house_sales_data.txt）。数据集包含了 774 条萨克拉门托地区的商品房销售数据。原始的数据特征包括：

x_1 为房屋面积；

x_2 为居室数量；

x_3 为房屋的分契式公寓属性，如果属于分契式公寓则 $x_3 = 1$，否则 $x_3 = 0$；

x_4 为房屋地址属性，通过邮政编码变换得到。

模型 1：

$$\hat{y} = \theta_1 + \theta_2 x_1 + \theta_3 x_2$$

求解模型后得到$\theta_1 = 54.40$、$\theta_2 = 148.73$、$\theta_3 = -18.85$。预测结果如表 7.2 所示。

表 7.2 模型 1 拟合结果（面积单位：1000 平方英尺，价格单位：$1000）

House	x_1（面积）	x_2（居室数）	y（实际价格）	$\hat{y}$（预测价格）
1	0.846	1	115.00	161.37
2	1.324	2	234.50	213.61
3	1.150	3	198.00	168.88
4	3.037	4	528.00	430.67
5	3.984	5	572.50	552.66

模型 2：

$$\hat{y} = \sum_{i=1}^{8} \theta_i f_i(\boldsymbol{x}) \qquad (7.2)$$

其中，基函数 $f_1(\boldsymbol{x}) = 1$，$f_2(\boldsymbol{x}) = x_1$，$f_3(\boldsymbol{x}) = \max\{x_1 - 1.5，0\}$。即，$f_2(\boldsymbol{x})$ 为房屋面积，$f_3(\boldsymbol{x})$ 是房屋面积超过1.5（1500 平方英尺）部分的值。基函数

$f_1(\boldsymbol{x})$、$f_2(\boldsymbol{x})$ 和 $f_3(\boldsymbol{x})$ 为商品房价格预测模型贡献了一个基于房屋面积的分段函数，

$$\theta_1 f_1(\boldsymbol{x}) + \theta_2 f_2(\boldsymbol{x}) + \theta_3 f_3(\boldsymbol{x}) = \begin{cases} \theta_1 + \theta_2 x_1 & x_1 \leq 1.5 \\ \theta_1 - 1.5\theta_3 + (\theta_2 + \theta_3) x_1 & x_1 > 1.5 \end{cases}$$

选取基函数 $f_4(\boldsymbol{x}) = x_2$，$f_5(\boldsymbol{x}) = x_3$，分别直接取值居室数量和分契式公寓属性。基函数 $f_6(\boldsymbol{x})$、$f_7(\boldsymbol{x})$ 和 $f_8(\boldsymbol{x})$ 从邮政编码构造，如表 7.3 所示。

求解模型 (7.2) 得到 $\theta_1 = 115.62$、$\theta_2 = 175.41$、$\theta_3 = -42.75$、$\theta_4 = -17.88$、$\theta_5 = -19.05$、$\theta_6 = -100.91$、$\theta_3 = -108.79$、$\theta_1 = -24.77$。

表 7.3 针对邮政编码数据 x_4 定义的基函数 $f_6(\boldsymbol{x})$、$f_7(\boldsymbol{x})$ 和 $f_8(\boldsymbol{x})$

ZIP codes	x_4	$f_6(\boldsymbol{x})$	$f_7(\boldsymbol{x})$	$f_8(\boldsymbol{x})$
95811, 95814, 95816, 95817, 95818, 95819	1	0	0	0
95608, 95610, 95621, 95626, 95628, 95655, 95660, 95662, 95670, 95673, 95683, 95691, 95742, 95815, 95821, 95825, 95827, 95833, 95834, 95835	2	1	0	0
95624, 95632, 95690, 95693, 95757, 95758, 95820, 95822, 95823, 95824, 95826, 95828, 95829, 95831, 95832	3	0	1	0
95603, 95614, 95630, 95635, 95648, 95650, 95661, 95663, 95677, 95678, 95682, 95722, 95746, 95747, 95762, 95765	4	0	0	1

7.3 最小二乘分类

对于 2 分类 ($y = \pm 1$) 问题，当 $y = 1$ 时，回归函数值应该接近1，当 $y = -1$ 时，回归函数值应该接近-1。可以使用符号函数 $\hat{y}(\boldsymbol{x}) = \text{sign}(\tilde{y}(\boldsymbol{x}) - \alpha)$ 对 y 进行预测，即：

$$\hat{y}(\boldsymbol{x}) = \begin{cases} +1, & \tilde{y}(\boldsymbol{x}) \geq \alpha \\ -1, & \tilde{y}(\boldsymbol{x}) < \alpha \end{cases}$$

其中，α 为决策阈值。

对于 $K > 2$ 的多分类问题，可以采用 1 对多的分类方法，取分类预测值为：

$$\hat{y}(\boldsymbol{x}) = \underset{\ell \in \{1,\ldots,K\}}{\text{argmax}}\, \tilde{y}_\ell(\boldsymbol{x})$$

选择 $\tilde{y}_\ell(\boldsymbol{x})$ 最大值对应的 ℓ 作为分类标签。

例 7.4 图像分类问题：基于手写数字数据库 MNIST，应用最小二乘模型对数字 0 和其他数字进行分类。MNIST 数据库来自美国国家标准与技术研究所，包含 70000 张 28×28 的手写数字灰度图像。图像数据已经被转化为

$28 \times 28 = 784$ 维的向量形式存储，例如，[0.0 0.0 … 0.0 0.380 0.376 0.301 0.462 … 0.239 … 0.0 0.0]，标签以 10 维向量形式存储，例如，[0.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0]。图 7.4 显示了部分数据集示例。数据库的下载网址为 http://yann.lecun.com/exdb/mnist/。

MNIST 数据库将数据分成 2 组，一组是包含 60000 张图像的训练集，另一组是包含 10000 张图像的测试集。将 $\boldsymbol{x}$ 设置为 494 维的向量，第 1 维是常量 1，其余 493 维是至少在 600 张图片中像素值不为 0 的像素。如果图像为数字 0，则取 $y = 1$，否则取 $y = -1$。

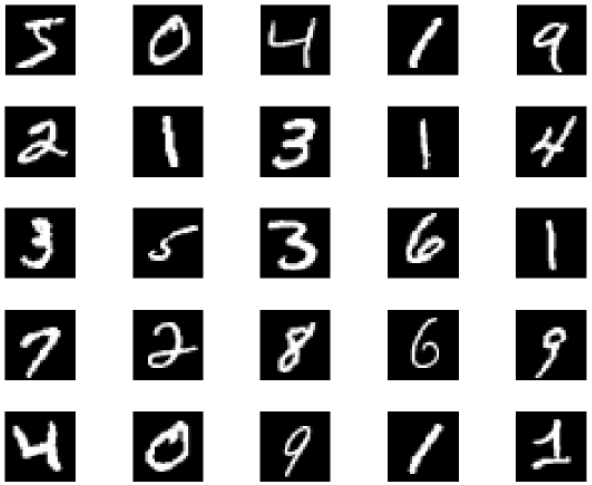

图 7.4 来自 MNIST 数据集的 25 张手写数字图像

应用最小二乘模型进行分类，实验结果如表 7.4 ~ 7.5 所示。

表 7.4 训练集结果（错误率 1.6%）

	$\hat{y}=+1$	$\hat{y}=-1$	Total
$y=+1$	5158	765	5923
$y=-1$	167	53910	54077
All	5325	54675	60000

表 7.5 测试集结果（错误率 1.6%）

	$\hat{y}=+1$	$\hat{y}=-1$	Total
$y=+1$	864	116	980
$y=-1$	42	8978	9020
All	906	9094	10000

按照如下方法构造 5000 个随机特征：生成随机矩阵 $\boldsymbol{R} \in R^{5000\times 494}$，$\boldsymbol{R}_{ij}$ 随机地取值 ± 1，取 $\max\{0，(\boldsymbol{Rx})_j\}$，$j = 1，\cdots，5000$，作为新特征，与原来的 494 个特征一起，构成 5494 个特征。实验结果表明：添加 5000 个随机特征以后，最小二乘分类精度显著提高。

作业 3：用你熟悉的计算机语言完成例题 7.4 的实验，给出训练集分类错误率和测试集分类错误率。**作业提交时间：**2019 年 11 月 23 日。

8　多目标最小二乘

8.1 多目标最小二乘问题及求解

多目标最小二乘问题是寻找 $\boldsymbol{x} \in R^n$，使得以下 k 个目标函数总体最小：

$$\min J_1 = \|\boldsymbol{A}_1\boldsymbol{x} - \boldsymbol{b}_1\|^2，\ \cdots，\ J_k = \|\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{b}_k\|^2$$

其中，$\boldsymbol{A}_i$ 是 $m_i \times n$ 矩阵，$\boldsymbol{b}_i$ 是 m_i 维向量，$i = 1，\ \cdots，\ k$。

求解多目标最小二乘问题的典型方法是将 k 个目标转换为如下加权目标函数：

$$J = \lambda_1 J_1 + \cdots + \lambda_k J_k = \lambda_1\|\boldsymbol{A}_1\boldsymbol{x} - \boldsymbol{b}_1\|^2 + \cdots + \lambda_k\|\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{b}_k\|^2 \qquad (8.1)$$

其中，$\lambda_i > 0$，$i = 1，\ \cdots，\ k$，为目标 J_i 的权重系数。

可以把加权目标函数 (8.1) 写成如下标准最小二乘问题形式进行求解，

$$J = \left\| \begin{bmatrix} \sqrt{\lambda_1}(\boldsymbol{A}_1\boldsymbol{x} - \boldsymbol{b}_1) \\ \vdots \\ \sqrt{\lambda_k}(\boldsymbol{A}_k\boldsymbol{x} - \boldsymbol{b}_k) \end{bmatrix} \right\|^2，$$

即，

$$J = \left\| \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{A}_1 \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{A}_k \end{bmatrix} \boldsymbol{x} - \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{b}_1 \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{b}_k \end{bmatrix} \right\|^2 = \left\| \widetilde{\boldsymbol{A}}\boldsymbol{x} - \widetilde{\boldsymbol{b}} \right\|^2,$$

其中，$\widetilde{\boldsymbol{A}} \in R^{m \times n}$，$\widetilde{\boldsymbol{b}} \in R^m$，$m = m_1 + \cdots + m_k$，

$$\widetilde{\boldsymbol{A}} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{A}_1 \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{A}_k \end{bmatrix}, \quad \widetilde{\boldsymbol{b}} = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{b}_1 \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{b}_k \end{bmatrix}。 \tag{8.2}$$

假设 $\widetilde{\boldsymbol{A}}$ 的列是线性独立的，则加权目标函数 (8.1) 存在唯一的最优解 $\hat{\boldsymbol{x}}$，

$$\begin{aligned} \hat{\boldsymbol{x}} &= \left(\widetilde{\boldsymbol{A}}^T\widetilde{\boldsymbol{A}}\right)^{-1}\widetilde{\boldsymbol{A}}^T\widetilde{\boldsymbol{b}} \\ &= \left(\lambda_1 {\boldsymbol{A}_1}^T\boldsymbol{A}_1 + \cdots + \lambda_k {\boldsymbol{A}_k}^T\boldsymbol{A}_k\right)^{-1}\left(\lambda_1 {\boldsymbol{A}_1}^T\boldsymbol{b}_1 + \cdots + \lambda_k {\boldsymbol{A}_k}^T\boldsymbol{b}_k\right)。 \end{aligned} \tag{8.3}$$

此处，$\widetilde{\boldsymbol{A}}$ 的列线性独立是指不存在非零向量 $\boldsymbol{x}$ 使得 $\boldsymbol{A}_i\boldsymbol{x} = \boldsymbol{0}$，$i = 1$，$\cdots$，$k$。即，只要 $\boldsymbol{A}_1$，$\cdots$，$\boldsymbol{A}_k$ 中有一个矩阵的列是线性独立的，则 $\widetilde{\boldsymbol{A}}$ 是线性独立。实际上，即使 $\boldsymbol{A}_1$，$\cdots$，$\boldsymbol{A}_k$ 中没有任何矩阵的列是线性独立的，$m_i < n$，$i = 1$，$\cdots$，k。如果 $m_1 + \cdots + m_k > n$，且 $\widetilde{\boldsymbol{A}}$ 是线性独立的，则式 (8.3) 仍适用。

8.2 多目标最小二乘应用

8.2.1 提克洛夫（Tikhonov）正则化

寻找 $\boldsymbol{x}$ 最小化如下函数，

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|^2+\lambda\|\boldsymbol{x}\|^2 \tag{8.4}$$

对给定的 $\lambda>0$，该问题称作 Tikhonov 正则化反演，根据美国数学家 Andrey Tikhonov 命名。

对于该问题，对应式 (8.2) 中的 $\widetilde{\boldsymbol{A}}$ 为

$$\widetilde{\boldsymbol{A}}=\begin{bmatrix}\boldsymbol{A}\\ \sqrt{\lambda}\boldsymbol{I}\end{bmatrix},$$

其列向量总是线性独立的。因而，$\widetilde{\boldsymbol{A}}$ 对应的 Gram 矩阵 $\widetilde{\boldsymbol{A}}^T\widetilde{\boldsymbol{A}}=\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I}$ 总是可逆的。于是，Tikhonov 正则化的解为

$$\hat{\boldsymbol{x}}=(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{y}。 \tag{8.5}$$

8.2.2 数据拟合中的正则化

考虑如下数据拟合模型，

$$\hat{f}(\boldsymbol{x})=\theta_1 f_1(\boldsymbol{x})+\cdots+\theta_p f_p(\boldsymbol{x}) \tag{8.6}$$

其中，θ_i 为拟合参数，可以解释为预测值对 $f_i(\boldsymbol{x})$ 的依赖程度。如果 θ_i 很大，则预测值对 $f_i(\boldsymbol{x})$ 的变化或波动非常敏感。我们希望预测模型中除了常数基函数，例如，$f_1(\boldsymbol{x}) = 1$，以外，对应的拟合参数 θ_i 尽可能小一些，降低预测值对 $f_i(\boldsymbol{x})$ 变化的敏感程度。因此，在最小化预测误差 $\|\boldsymbol{y} - \boldsymbol{A\theta}\|^2$ 的基础上，希望 $\left\|\boldsymbol{\theta}_{2:p}\right\|^2$ 尽可能小，此处，假设$f_1(\boldsymbol{x}) = 1$。于是，多目标最小二乘函数为：

$$\|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\left\|\boldsymbol{\theta}_{2:p}\right\|^2 \tag{8.7}$$

其中，λ 称作正则化参数。

令 $f_1(\boldsymbol{x}) = v$ 为常数，$\boldsymbol{\beta} = \boldsymbol{\theta}_{2:p}$，则回归模型可描述为 $\hat{y} = \boldsymbol{X\beta} + v\boldsymbol{1}$，加权目标函数 (8.7) 可表示为

$$\|\boldsymbol{y} - \boldsymbol{X\beta} - v\boldsymbol{1}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

以上加权目标形式的回归模型称作岭回归。选择不同的正则化参数 λ 会得到不同的回归模型。

例 8.1 岭回归。根据如下模型生成一组噪声数据（数据源：regularized_fit_data.txt）

$$s(t) = c + \sum_{k=1}^{4} \alpha_k \cos(\omega_k t + \phi_k)$$

其中，参数 $c = 1.54$，$\alpha_1 = 0.66$，$\alpha_2 = -0.90$，$\alpha_3 = -0.66$，$\alpha_4 = 0.89$；$\omega_1 = 13.69, \omega_2 = 3.55, \omega_3 = 23.25, \omega_4 = 6.03$；$\phi_1 = 0.21, \phi_2 = 0.02, \phi_3 = -1.87$，$\phi_4 = 1.72$。数据集如图 8.1 所示。

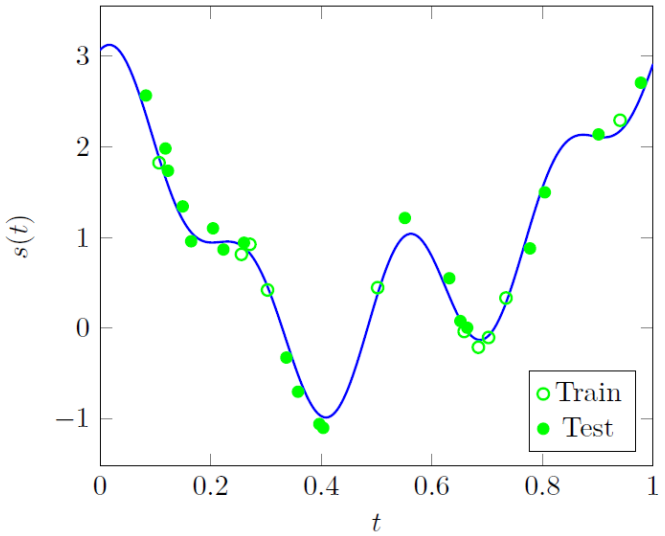

图 8.1 函数 $s(t)$ 和 30 组噪声数据，10 组训练数据，20 组测试数据

应用以下基函数进行数据拟合，

$$f_1(x) = 1，\ f_{k+1}(x) = \cos(\omega_k x + \phi_k)，\ k = 1，\ \cdots，\ 4，$$

结果如图 8.2 所示。结论如下：

（1）最小测试误差 RMS 发生在 $\lambda = 0.08$时；

（2）增加 λ 会压缩系数θ_2，…，θ_5；

（3）点画线显示的是生成数据时的参数；

（4）$\lambda = 0.08$时，最优拟合参数最接近真实值。

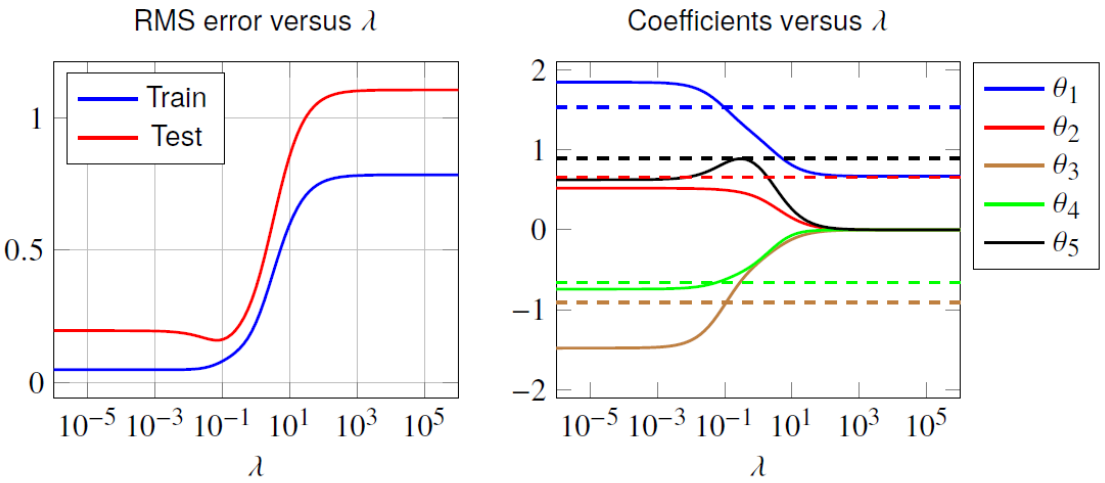

图 8.2 岭回归数据拟合结果

9　约束最小二乘

9.1 约束最小二乘问题及求解

约束最小二乘问题形式如下：

$$\begin{aligned} \min\ & f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 \\ \text{s.t.}\ & \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d} \end{aligned} \tag{9.1}$$

其中，向量 $\boldsymbol{x} \in R^n$、$\boldsymbol{A} \in R^{m\times n}$、$\boldsymbol{b} \in R^m$、$\boldsymbol{C} \in R^{p\times n}$、$\boldsymbol{d} \in R^p$。

将问题 (9.1) 的约束条件 $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$ 改写成 $\boldsymbol{C}_i^T\boldsymbol{x} = d_i$，$i = 1$，$\cdots$，$p$，其中，$\boldsymbol{C}_i^T$ 为矩阵 $\boldsymbol{C}$ 的第 i 行。构造如下拉格朗日函数，

$$L(\boldsymbol{x},\ \boldsymbol{z}) = f(\boldsymbol{x}) + z_1(\boldsymbol{C}_1^T\boldsymbol{x} - d_1) + \cdots + z_p(\boldsymbol{C}_p^T\boldsymbol{x} - d_p)$$

其中，$\boldsymbol{z} \in R^p$ 为拉格朗日乘子。

由拉格朗日乘子方法可知，如果 $\hat{\boldsymbol{x}}$ 是问题 (9.1) 的最优解，则存在拉格朗日乘子 $\boldsymbol{z}$ 满足如下最优性条件：

$$\frac{\partial L}{\partial x_i}(\hat{\boldsymbol{x}},\ \boldsymbol{z}) = 0,\ i = 1,\ \cdots,\ n, \qquad \frac{\partial L}{\partial z_i}(\hat{\boldsymbol{x}},\ \boldsymbol{z}) = 0,\ i = 1,\ \cdots,\ p \tag{9.2}$$

最优性条件的第 2 组等式可以写成

$$\frac{\partial L}{\partial z_i}\left(\hat{\boldsymbol{x}},\ \boldsymbol{z}\right)=\boldsymbol{C}_i^T\hat{\boldsymbol{x}}-d_i=0,\ i=1,\ \cdots,\ p,$$

表明 $\hat{\boldsymbol{x}}$ 满足 $\boldsymbol{C}\hat{\boldsymbol{x}}=\boldsymbol{d}$ 的约束条件。第 1 组等式可以进一步写成

$$\frac{\partial L}{\partial x_i}\left(\hat{\boldsymbol{x}},\ \boldsymbol{z}\right)=2\sum_{j=1}^{n}(\boldsymbol{A}^T\boldsymbol{A})_{ij}\hat{x}_j-2(\boldsymbol{A}^T\boldsymbol{b})_i+\sum_{j=1}^{p}z_j\left(c_j\right)_i=0,$$

写成矩阵的紧凑形式为

$$2(\boldsymbol{A}^T\boldsymbol{A})\hat{\boldsymbol{x}}-2\boldsymbol{A}^T\boldsymbol{b}+\boldsymbol{C}^T\boldsymbol{z}=\boldsymbol{0},$$

结合可行性条件 $\boldsymbol{C}\hat{\boldsymbol{x}}=\boldsymbol{d}$，可以将最优性条件 (9.2) 写成如下 $n+p$ 个关于变量 $\left(\hat{\boldsymbol{x}},\ \boldsymbol{z}\right)$ 的线性等式

$$\begin{bmatrix}2\boldsymbol{A}^T\boldsymbol{A} & \boldsymbol{C}^T\\ \boldsymbol{C} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{x}}\\ \boldsymbol{z}\end{bmatrix}=\begin{bmatrix}2\boldsymbol{A}^T\boldsymbol{b}\\ \boldsymbol{d}\end{bmatrix}\tag{9.3}$$

式 (9.3) 称作 KKT 方程组，其中，$(n+p)\times(n+p)$ 方阵 $\begin{bmatrix}2\boldsymbol{A}^T\boldsymbol{A} & \boldsymbol{C}^T\\ \boldsymbol{C} & \boldsymbol{0}\end{bmatrix}$ 被称作 KKT 矩阵。该方程组将约束最小二乘问题转化为 $n+p$ 个线性等式的求解问题。

KKT 矩阵可逆的充分必要条件是：矩阵 $\boldsymbol{C}$ 的行相互独立，$\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{C}\end{bmatrix}$ 的列相互独立。当 KKT 矩阵可逆时，KKT 方程组 (9.3) 的解为：

$$\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{z}}\end{bmatrix}=\begin{bmatrix}2\boldsymbol{A}^T\boldsymbol{A} & \boldsymbol{C}^T\\ \boldsymbol{C} & \boldsymbol{0}\end{bmatrix}^{-1}\begin{bmatrix}2\boldsymbol{A}^T\boldsymbol{b}\\ \boldsymbol{d}\end{bmatrix} \tag{9.4}$$

基于以上分析，可以设计基于 KKT 方程组的约束最小二乘问题求解算法如下。

算法 9.1 基于 KKT 方程式的约束最小二乘问题求解算法

步骤 0：满足 KKT 矩阵可逆条件的 $\boldsymbol{A}\in R^{m\times n}$、$\boldsymbol{C}\in R^{p\times n}$，$\boldsymbol{b}\in R^{m}$、$\boldsymbol{d}\in R^{p}$

步骤 1：构造 Gram 矩阵。计算$\boldsymbol{A}^T\boldsymbol{A}$。

步骤 2：求解 KKT 方程组。应用 $\boldsymbol{QR}$ 分解方法求解 KKT 方程组 (9.3)。

以上算法的计算复杂性为 $2mn^2+2(n+p)^3$。

在约束最小二乘问题中，如果 $\boldsymbol{A}=\boldsymbol{I}$，$\boldsymbol{b}=\boldsymbol{0}$，则约束最小二乘问题转变为如下约束最小范数问题：

$$\begin{aligned} \min\ & f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 \\ \text{s.t.}\ & \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d} \end{aligned} \tag{9.5}$$

其中，向量 $\boldsymbol{x} \in R^n$、$\boldsymbol{I} \in R^{m\times n}$、$\boldsymbol{C} \in R^{p\times n}$、$\boldsymbol{d} \in R^p$。

约束最小范数问题 (9.5) 对应的 KKT 方程组为

$$\begin{bmatrix} 2\boldsymbol{I} & \boldsymbol{C}^T \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{x}} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{d} \end{bmatrix} \tag{9.6}$$

由式 (9.6) 得

$$2\hat{\boldsymbol{x}} + \boldsymbol{C}^T\boldsymbol{z} = \boldsymbol{0} \quad \Rightarrow \quad \hat{\boldsymbol{x}} = -(1/2)\boldsymbol{C}^T\boldsymbol{z}$$

代入 $\hat{\boldsymbol{x}}$ 到 $\boldsymbol{C}\hat{\boldsymbol{x}} = \boldsymbol{d}$ 得

$$-(1/2)\boldsymbol{C}\boldsymbol{C}^T\boldsymbol{z} = \boldsymbol{d}$$

由于矩阵 $\boldsymbol{C}$ 的行是线性独立的，因此，$\boldsymbol{C}\boldsymbol{C}^T$ 可逆，因而有

$$\boldsymbol{z} = -2(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{d}$$

将 $\boldsymbol{z}$ 代入 $\hat{\boldsymbol{x}}$ 的表达式得，

$$\hat{\boldsymbol{x}} = \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}\boldsymbol{d} \tag{9.7}$$

简写为

$$\hat{\boldsymbol{x}} = \boldsymbol{C}^{\dagger}\boldsymbol{d}$$

其中，$\boldsymbol{C}^{\dagger} = \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T)^{-1}$ 为行线性独立矩阵 $\boldsymbol{C}$ 的伪逆。

10　非线性最小二乘

10.1 非线性最小二乘问题

考虑一组 m 个非线性等式

$$f_i(\boldsymbol{x}) = 0，\ i = 1，\ \cdots，\ m，$$

其中，$\boldsymbol{x} \in R^n$，f_i：$R^n \to R$ 为实值函数，称作第 i 个残差。

将非线性等式组写成紧凑的向量形式为，

$$f(\boldsymbol{x}) = \boldsymbol{0}，\tag{10.1}$$

其中，$f(\boldsymbol{x}) = \left(f_1(\boldsymbol{x})，\ \cdots，\ f_m(\boldsymbol{x})\right)$ 为 m 维向量。可以认为 f 是一个从 n 维向量到 m 维向量的一个映射，即，f：$R^n \to R^m$，称 m 维向量 $f(\boldsymbol{x})$ 为残差向量。

如果不能找到非线性等式组 (10.1) 的解，则可以寻找近似解，即最小化残差平方和

$$f_1(\boldsymbol{x})^2 + \cdots + f_m(\boldsymbol{x})^2 = \|f(\boldsymbol{x})\|^2。$$

也就是寻找如下非线性最小二乘问题的最优解

$$\min \|f(\boldsymbol{x})\|^2 。\tag{10.2}$$

由最优化理论可知，如果 $\hat{\boldsymbol{x}}$ 为非线性最小二乘问题 (10.2) 的最优解，则 $\hat{\boldsymbol{x}}$ 满足如下条件：

$$\frac{\partial}{\partial x_i}\|f(\hat{\boldsymbol{x}})\|^2 = 0，\ i = 1，\ \cdots，\ n，$$

写成向量形式为 $\nabla\|f(\hat{\boldsymbol{x}})\|^2 = \boldsymbol{0}$。即，非线性最小二乘问题 (10.2) 的最优解 $\hat{\boldsymbol{x}}$ 的必要条件为：

$$\nabla\|f(\hat{\boldsymbol{x}})\|^2 = \nabla\left(\sum_{i=1}^{m} f_i(\hat{\boldsymbol{x}})^2\right) = 2\sum_{i=1}^{m} f_i(\hat{\boldsymbol{x}})\nabla f_i(\hat{\boldsymbol{x}}) = 2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}}) = \boldsymbol{0}，\quad (10.3)$$

其中，$Df(\hat{\boldsymbol{x}})$ 为函数 f 的 $m \times n$ 维导数矩阵，也称 Jacobian 矩阵，

$$Df(\hat{\boldsymbol{x}}) = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}_{\boldsymbol{x}=\hat{\boldsymbol{x}}}。$$

10.2 非线性最小二乘求解

求解非线性方程组或非线性最小二乘问题要比求解线性方程组困难得多，甚至有时判断解的存在性就已经非常困难。虽然有些高级算法能够精确地求解非线性最小二乘问题，但计算量巨大，在实际中很少应用。因此，本章介绍两种启发式算法，Gauss-Newton 算法和 Levenberg-Marquardt 算法，虽然不能保证解的最优性，但实际应用效果却非常好。

Gauss-Newton 算法和 Levenberg-Marquardt 算法均属于迭代算法。算法生成解的序列$\boldsymbol{x}^{(1)}$，$\boldsymbol{x}^{(2)}$，…。$\boldsymbol{x}^{(1)}$ 为算法初始解，$\boldsymbol{x}^{(k)}$ 为算法第 k 次迭代的解，从解 $\boldsymbol{x}^{(k)}$ 到 $\boldsymbol{x}^{(k+1)}$ 称作算法的一次迭代。当 $\|f(\hat{\boldsymbol{x}})\|$ 足够小、或者 $\boldsymbol{x}^{(k)}$ 和 $\boldsymbol{x}^{(k+1)}$ 足够接近、或者算法达到了预先设定的迭代代数时，算法停止运行。

10.2.1 Gauss-Newton 算法

Gauss-Newton 算法在第 k 次迭代时，用一阶 Taylor 函数 $\hat{f}$ 构造非线性函数 f 的近似函数，并应用最小二乘算法计算近似线性方程组的解。在当前解点 $\boldsymbol{x}^{(k)}$，函数 f 的一阶 Taylor 近似函数 $\hat{f}$ 为

$$\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)=f\left(\boldsymbol{x}^{(k)}\right)+Df\left(\boldsymbol{x}^{(k)}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{(k)}\right) \tag{10.4}$$

其中，$Df\left(\boldsymbol{x}^{(k)}\right)$ 为函数 f 在 $\boldsymbol{x}^{(k)}$ 点的 $m\times n$ 维 Jacobian 矩阵。当 $\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\|$ 足够小时，$\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)$ 是 $f(\boldsymbol{x})$ 很好的近似。如果 $Df\left(\boldsymbol{x}^{(k)}\right)$ 具有线性独立的列向量（$m\geq n$），求解最小二乘问题 $\left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2$ 得，

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\left(Df\left(\boldsymbol{x}^{(k)}\right)^T Df\left(\boldsymbol{x}^{(k)}\right)\right)^{-1} Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right) \tag{10.5}$$

算法 10.1 求解非线性最小二乘问题的 Gauss-Newton 算法

步骤 1：给定可微函数 f：$R^n\to R^m$，选定初始解 $\boldsymbol{x}^{(1)}$。

步骤 2：**For** $k=1$，2，$\cdots$，$k^{\max}$

步骤 2.1：构造函数 f 的近似函数 $\hat{f}$，计算 Jacobian 矩阵 $Df\left(\boldsymbol{x}^{(k)}\right)$

$$\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)=f\left(\boldsymbol{x}^{(k)}\right)+Df\left(\boldsymbol{x}^{(k)}\right)\left(\boldsymbol{x}-\boldsymbol{x}^{(k)}\right)$$

步骤 2.2：求解最小二乘问题 $\min\left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2$，令

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\left(Df\left(\boldsymbol{x}^{(k)}\right)^T Df\left(\boldsymbol{x}^{(k)}\right)\right)^{-1} Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right)。$$

当 $\left\|f\left(\boldsymbol{x}^{(k+1)}\right)\right\|$ 足够小，或者 $\boldsymbol{x}^{(k)}$ 和 $\boldsymbol{x}^{(k+1)}$ 足够接近时，算法 10.1 提前终止。

Gauss-Newton 算法有两个假设前提：(1)$\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\|$ 足够小；(2)$Df\left(\boldsymbol{x}^{(k)}\right)$ 具有线性独立的列向量。当以上两点不满足时，Gauss-Newton 算法很可能失效。

10.2.2 Levenberg-Marquardt 算法

Levenberg-Marquardt 算法是 Gauss-Newton 算法的延伸。因此，有时也被称作 Gauss-Newton 算法。

在应用 Gauss-Newton 算法时，有时得到的最小二乘问题 $\left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2$ 的最优解 $x^{(k+1)}$ 会远离 $\boldsymbol{x}^{(k)}$，因而不能保证 $\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\approx f(\boldsymbol{x})$，意味着不满足条件 $\left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2\approx\|f(\boldsymbol{x})\|^2$。因此，选择 $x^{(k+1)}$ 时需要考虑两个目标：最小化 $\left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2$ 和最小化 $\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\|^2$。

求解如下函数：

$$\min \left\|\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right)\right\|^2+\lambda^{(k)}\left\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\right\|^2, \tag{10.6}$$

其中，$\lambda^{(k)}>0$ 为算法参数。问题 (10.6) 等价于如下问题：

$$\min \left\|\begin{bmatrix} Df\left(\boldsymbol{x}^{(k)}\right) \\ \sqrt{\lambda^{(k)}}\boldsymbol{I} \end{bmatrix}\boldsymbol{x}-\begin{bmatrix} Df\left(\boldsymbol{x}^{(k)}\right)\boldsymbol{x}^{(k)}-f\left(\boldsymbol{x}^{(k)}\right) \\ \sqrt{\lambda^{(k)}}\boldsymbol{x}^{(k)} \end{bmatrix}\right\|^2$$

因此，

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\left(Df\left(\boldsymbol{x}^{(k)}\right)^T Df\left(\boldsymbol{x}^{(k)}\right)+\lambda^{(k)}\boldsymbol{I}\right)^{-1} Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right), \tag{10.7}$$

当 $Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right)=\boldsymbol{0}$ 时，$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}$ 为最优解。

算法 10.2 求解非线性最小二乘问题的 Levenberg-Marquardt 算法

步骤 1：给定可微函数 f：$R^n \to R^m$，选定初始解 $\boldsymbol{x}^{(1)}$ 和初始参数 $\lambda^{(1)} > 0$。

步骤 2：**For** $k = 1, 2, \cdots, k^{\max}$

步骤 2.1：构造函数 f 的近似函数 $\hat{f}$，计算 Jacobian 矩阵 $Df\left(\boldsymbol{x}^{(k)}\right)$

$$\hat{f}\left(\boldsymbol{x};\ \boldsymbol{x}^{(k)}\right) = f\left(\boldsymbol{x}^{(k)}\right) + Df\left(\boldsymbol{x}^{(k)}\right)\left(\boldsymbol{x} - \boldsymbol{x}^{(k)}\right)$$

步骤 2.2：求解最小二乘问题 (10.6) 得

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(Df\left(\boldsymbol{x}^{(k)}\right)^T Df\left(\boldsymbol{x}^{(k)}\right) + \lambda^{(k)}\boldsymbol{I}\right)^{-1} Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right)$$

步骤 2.3：判断是否接受 $\boldsymbol{x}^{(k+1)}$ 并更新参数 $\lambda^{(k+1)}$

如果 $\left\|f\left(\boldsymbol{x}^{(k+1)}\right)\right\|^2 < \left\|f\left(\boldsymbol{x}^{(k)}\right)\right\|^2$，更新当前解为 $\boldsymbol{x}^{(k+1)}$，更新 $\lambda^{(k+1)} = 0.8\lambda^{(k)}$；

否则，保持当前解为 $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$，更新 $\lambda^{(k+1)} = 2\lambda^{(k)}$。

当 $\left\|f\left(\boldsymbol{x}^{(k+1)}\right)\right\|$ 足够小时，算法 10.2 提前终止，此时，获得了问题 (10.6) 的近优解。当 $\|2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}})\|$ 足够小时，算法 10.2 也提前终止，但此时未必

获得问题 (10.6) 的近优解。

应用 Levenberg-Marquardt 算法时，如果求解的一系列问题类似，可以采用热启动的方法减少算法的迭代代数。也可以采用从多个初始点出发，多次运行算法的方法来提高算法的求解效果。

10.3 非线性最小二乘应用

10.4.1 非线性最小二乘回归

应用函数

$$f(x;\ \boldsymbol{\theta}) = \theta_1 e^{\theta_2 x}\cos(\theta_3 x + \theta_4)$$

生成 $N = 60$ 个点 $(x^{(i)},\ y^{(i)})$ 并加上扰动。利用非线性最小二乘模型拟合模型参数$\boldsymbol{\theta}$。

非线性最小二乘模型为

$$\min \sum_{i=1}^{N} \left(\theta_1 e^{\theta_2 x^{(i)}}\cos\left(\theta_3 x^{(i)} + \theta_4\right) - y^{(i)}\right)^2,$$

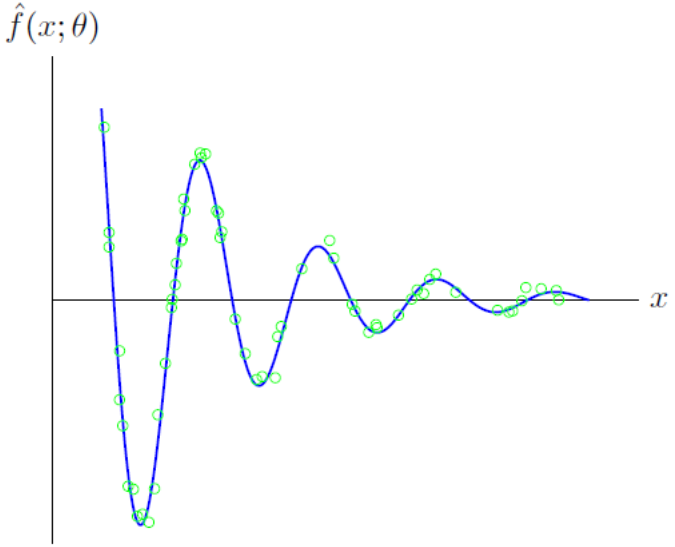

图 10.1 非线性最小二乘模型拟合函数 $\hat{f}(x;\boldsymbol{\theta}) = \theta_1 e^{\theta_2 x}\cos(\theta_3 x + \theta_4)$

10.4.2 非线性最小二乘分类

对于 2 分类 $(y = \pm 1)$ 问题，当 $y = 1$ 时，回归函数值应该接近1，当 $y = -1$ 时，回归函数值应该接近-1。因此，可以求解如下误差函数的最小值来得到分类模型

$$\sum_{i=1}^{N}\left(\hat{f}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2 = \sum_{i=1}^{N}\left(\mathbf{sign}\left(\tilde{f}\left(\boldsymbol{x}^{(i)}\right)\right) - y^{(i)}\right)^2, \tag{10.8}$$

其中，$\tilde{f}\left(\boldsymbol{x}^{(i)}\right)$ 为连续函数。

符号函数 **sign**() 不可微。无法使用 Levenberg-Marquardt 算法求解 (10.8) 描述的最小二乘问题。通常的替代方法是使用 sigmoid 函数替代符号函数。sigmoid 函数定义为

$$\phi(u)=\frac{e^{u}-e^{-u}}{e^{u}+e^{-u}},$$

函数图形如图 10.2 所示。sigmoid 函数 $\phi(u)$ 连续可微，可以应用 Levenberg-Marquardt 算法求解如下非线性最小二乘问题进行分类预测

$$\sum_{i=1}^{N}\left(\phi\left(\tilde{f}\left(\boldsymbol{x}^{(i)}\right)\right)-y^{(i)}\right)^{2}。\tag{10.9}$$

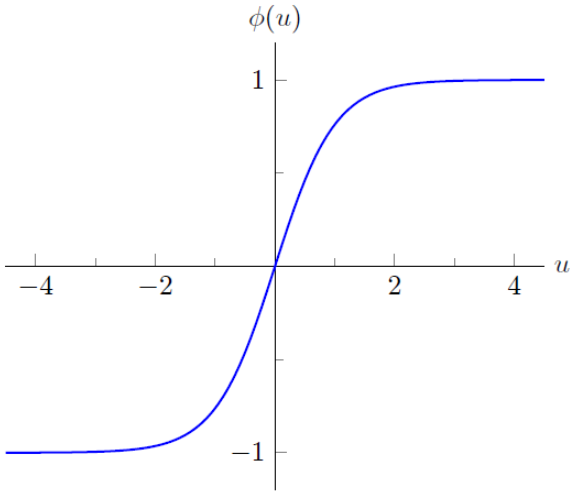

图 10.2 sigmoid 函数

例 10.1 利用非线性最小二乘分类求解例 7.4 的图像分类问题。将 $\boldsymbol{x}$ 设置为 493 维至少在 600 张图片中像素值不为 0 的像素，加上 1 维常量值 v，构成线性模型

$$\tilde{f}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + v,$$

通过求解如下非线性最小二乘问题确定参数 $\boldsymbol{\beta}$ 和 v，

$$\min \sum_{i=1}^{N} \left(\phi\left(\left(\boldsymbol{x}^{(i)}\right)^T\boldsymbol{\beta} + v\right) - y^{(i)}\right)^2 + \lambda\|\boldsymbol{\beta}\|^2 \text{。} \tag{10.10}$$

其中，ϕ 为 sigmoid 函数，λ 为正则化参数。

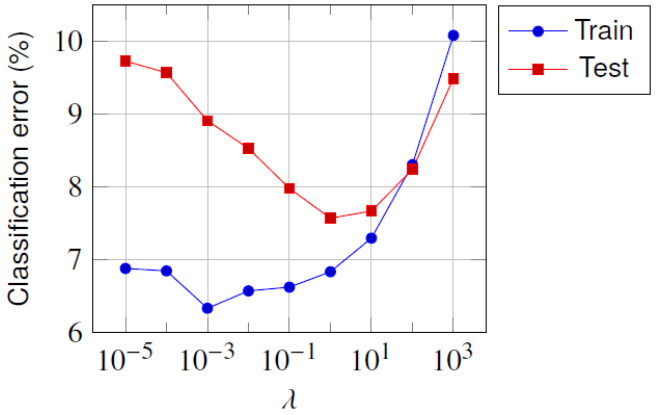

图 10.3 预测误差与正则化参数 λ 的关系

按照第 7.3 节的方法构造 5000 个随机特征。实验结果表明：添加 5000 个随机特征以后，非线性最小二乘分类精度显著提高。

作业 4：用你熟悉的计算机语言完成例题 10.1 的实验，给出训练集分类错误率和测试集分类错误率。**作业提交时间：**2019 年 11 月 23 日。

11　约束非线性最小二乘

11.1 约束非线性最小二乘问题

约束非线性最小二乘问题具有如下形式：

$$\begin{aligned} \min \ & \|f(\boldsymbol{x})\|^2 \\ \text{s.t.} \ & g(\boldsymbol{x}) = \boldsymbol{0} \end{aligned} \tag{11.1}$$

其中，向量 $\boldsymbol{x} \in R^n$，$f(\boldsymbol{x})$ 为 m 维向量函数，$g(\boldsymbol{x})$ 为 p 维向量函数。问题 (11.1) 可以展开写成如下形式

$$\begin{aligned} \min \ & f_1(\boldsymbol{x})^2 + \cdots + f_m(\boldsymbol{x})^2 \\ \text{s.t.} \ & g_i(\boldsymbol{x}) = 0，\ i = 1，\ \cdots，\ p。 \end{aligned}$$

当约束条件函数 $g(\boldsymbol{x})$ 为仿射函数时，$g(\boldsymbol{x}) = \boldsymbol{0}$ 可以写成 $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$ 的形式。此时，可以应用 Levenberg-Marquardt 算法求解问题 (11.1)。

称满足 $g(\boldsymbol{x}) = \boldsymbol{0}$ 的 $\boldsymbol{x}$ 为问题 (11.1) 的可行解。如果 $\hat{\boldsymbol{x}}$ 是可行解，并且对任意可行解$\boldsymbol{x}$，均有 $\|f(\boldsymbol{x})\|^2 \geq \|f(\hat{\boldsymbol{x}})\|^2$，则称 $\hat{\boldsymbol{x}}$ 是问题 (11.1) 的解。

构造问题 (11.1) 的拉格朗日函数

$$L(\boldsymbol{x},\ \boldsymbol{z}) = \|f(\boldsymbol{x})\|^2 + z_1 g_1(\boldsymbol{x}) + \cdots + z_p g_p(\boldsymbol{x}) = \|f(\boldsymbol{x})\|^2 + g(\boldsymbol{x})^T \boldsymbol{z}, \quad (11.2)$$

其中，$\boldsymbol{z} \in R^p$ 为拉格朗日乘子。对于问题 (11.1) 的解 $\hat{\boldsymbol{x}}$，一定存在 $\hat{\boldsymbol{z}}$ 满足如下最优性条件：

$$\frac{\partial L}{\partial x_i}(\hat{\boldsymbol{x}},\ \hat{\boldsymbol{z}}) = 0,\ i = 1,\ \cdots,\ n, \qquad \frac{\partial L}{\partial z_i}(\hat{\boldsymbol{x}},\ \hat{\boldsymbol{z}}) = 0,\ i = 1,\ \cdots,\ p。$$

在此，假设 $\nabla g_1(\hat{\boldsymbol{x}}),\ \cdots,\ \nabla g_p(\hat{\boldsymbol{x}})$ 是线性独立的。

最优性条件的第 2 组等式可以写成 $g_i(\hat{\boldsymbol{x}}) = 0,\ i = 1,\ \cdots,\ p$。写成向量形式为

$$g(\boldsymbol{x}) = \boldsymbol{0}, \quad (11.3)$$

表明 $\hat{\boldsymbol{x}}$ 是可行解。第 1 组等式写成向量形式为

$$2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}}) + Dg(\hat{\boldsymbol{x}})^T \hat{\boldsymbol{z}} = \boldsymbol{0}, \quad (11.4)$$

条件 (11.3) 和 (11.4) 构成了问题 (11.1) 的最优解 $\hat{\boldsymbol{x}}$ 的必要条件。

11.2 惩罚算法

通过对不可行解施加惩罚的方法，可以将约束非线性最小二乘问题 (11.1) 转化为如下形式求得近优解：

$$\min \ \|f(\boldsymbol{x})\|^2 + \mu\|g(\boldsymbol{x})\|^2 \tag{11.5}$$

其中，$\mu > 0$ 为惩罚系数。

问题 (11.5) 的解可以通过应用 Levenberg-Marquardt 算法求解如下问题得到

$$\min \left\| \begin{matrix} f(\boldsymbol{x}) \\ \sqrt{\mu} g(\boldsymbol{x}) \end{matrix} \right\|^2 \tag{11.6}$$

当惩罚系数 μ 足够大时，会得到解 $\boldsymbol{x}$ 使得 $\|g(\boldsymbol{x})\|^2$ 非常小，$\|f(\boldsymbol{x})\|^2$ 也很小，认为是问题 (11.1) 的近似解。

算法 11.1 求解约束非线性最小二乘问题的惩罚算法

步骤 1：给定可微函数 f：$R^n \to R^m$ 和 g：$R^n \to R^p$，初始解 $\boldsymbol{x}^{(1)}$ 和初始参数 $\mu^{(1)} = 1$。

步骤 2：**For** $k = 1, 2, \cdots, k^{\max}$

步骤 2.1：从 $\boldsymbol{x}^{(k)}$ 出发，应用 Levenberg-Marquardt 算法求解非线性最小二乘问题

$$\min \ \|f(\boldsymbol{x})\|^2 + \mu^{(k)}\|g(\boldsymbol{x})\|^2$$

获得解 $\boldsymbol{x}^{(k+1)}$。

步骤 2.2：更新参数 $\mu^{(k)}$：$\mu^{(k+1)} = 2\mu^{(k)}$。

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(Df\left(\boldsymbol{x}^{(k)}\right)^T Df\left(\boldsymbol{x}^{(k)}\right) + \lambda^{(k)}\boldsymbol{I}\right)^{-1} Df\left(\boldsymbol{x}^{(k)}\right)^T f\left(\boldsymbol{x}^{(k)}\right)$$

当 $\left\|g\left(\boldsymbol{x}^{(k)}\right)\right\|$ 足够小时，算法 11.1 提前终止，此时，获得了问题 (11.1) 的近优解。惩罚算法的一个缺点是惩罚系数 $\mu^{(k)}$ 会随着迭代代数的增加迅速变大，导致算法需要很多次迭代，甚至算法失败。

11.3 增广拉格朗日算法

增广拉格朗日算法是对惩罚算法的改进，以解决惩罚算法在惩罚系数 μ 变大时遇到的困难。该算法由 Magnus Hestenes 和 Michael Powell 在 1960 年代提出。

问题 (11.1) 的增广拉格朗日函数为

$$\begin{aligned} L_\mu(\boldsymbol{x},\ \boldsymbol{z}) &= L(\boldsymbol{x},\ \boldsymbol{z}) + \mu\|g(\boldsymbol{x})\|^2 \\ &= \|f(\boldsymbol{x})\|^2 + g(\boldsymbol{x})^T\boldsymbol{z} + \mu\|g(\boldsymbol{x})\|^2 \end{aligned} \tag{11.7}$$

即，在函数 $L(\boldsymbol{x},\ \boldsymbol{z})$ 的基础上增加了 $\mu\|g(\boldsymbol{x})\|^2$。

增广拉格朗日函数与以下问题的拉格朗日函数相同

$$\begin{aligned} \min\ & \|f(\boldsymbol{x})\|^2 + \mu\|g(\boldsymbol{x})\|^2 \\ \text{s.t.}\ & g(\boldsymbol{x}) = \boldsymbol{0} \end{aligned} \tag{11.8}$$

问题 (11.8) 与 (11.1) 有相同的最优解。

增广拉格朗日算法通过计算一系列 μ、$\boldsymbol{z}$ 值下的关于 $\boldsymbol{x}$ 的最小化增广拉格朗日函数求得问题的最优解。增广拉格朗日函数具有如下等价形式：

$$L_\mu(\boldsymbol{x},\ \boldsymbol{z}) = \|f(\boldsymbol{x})\|^2 + g(\boldsymbol{x})^T\boldsymbol{z} + \mu\|g(\boldsymbol{x})\|^2$$

$$= \|f(\boldsymbol{x})\|^2 + \mu \left\| g(\boldsymbol{x}) + \frac{1}{2\mu}\boldsymbol{z} \right\|^2 - \frac{1}{2\mu}\|\boldsymbol{z}\|^2$$

$$= \left\| \begin{bmatrix} f(\boldsymbol{x}) \\ \sqrt{\mu} g(\boldsymbol{x}) + \frac{1}{2\sqrt{\mu}}\boldsymbol{z} \end{bmatrix} \right\|^2 - \frac{1}{2\mu}\|\boldsymbol{z}\|^2$$

当给定 μ、$\boldsymbol{z}$ 时，求解关于 $\boldsymbol{x}$ 的最小化增广拉格朗日函数等价于求解以下问题

$$\left\| \begin{bmatrix} f(\boldsymbol{x}) \\ \sqrt{\mu} g(\boldsymbol{x}) + \frac{1}{2\sqrt{\mu}}\boldsymbol{z} \end{bmatrix} \right\|^2 \tag{11.9}$$

问题 (11.9) 可以应用 Levenberg-Marquardt 算法求解，其最优解 $\hat{\boldsymbol{x}}$ 满足如下最优性条件：

$$\begin{aligned} \mathbf{0} &= 2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}}) + 2\mu Dg(\hat{\boldsymbol{x}})^T \left(g(\hat{\boldsymbol{x}}) + \frac{1}{2\mu}\boldsymbol{z} \right) \\ &= 2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}}) + Dg(\hat{\boldsymbol{x}})^T (2\mu g(\hat{\boldsymbol{x}}) + \boldsymbol{z}) \end{aligned}$$

定义

$$\hat{\boldsymbol{z}} = 2\mu g(\hat{\boldsymbol{x}}) + \boldsymbol{z}$$

则有

$$2Df(\hat{\boldsymbol{x}})^T f(\hat{\boldsymbol{x}}) + Dg(\hat{\boldsymbol{x}})^T \hat{\boldsymbol{z}} = \mathbf{0}$$

如果 $g(\hat{\boldsymbol{x}}) = \mathbf{0}$ 成立，满足了 (11.4) 的问题 (11.1) 最优解的必要条件。

算法 11.2 求解约束非线性最小二乘问题的增广拉格朗日算法

步骤 1：给定可微函数 f：$R^n \to R^m$ 和 g：$R^n \to R^p$，初始解 $\boldsymbol{x}^{(1)}$，设置初始参数 $\boldsymbol{z}^{(1)} = \boldsymbol{0}$，$\mu^{(1)} = 1$。

步骤 2：**For** $k = 1, 2, \cdots, k^{\max}$

步骤 2.1：从 $\boldsymbol{x}^{(k)}$ 出发，应用 Levenberg-Marquardt 算法求解非线性最小二乘问题

$$\min \|f(\boldsymbol{x})\|^2 + \mu^{(k)} \left\| g(\boldsymbol{x}) + \frac{1}{2\mu^{(k)}} \boldsymbol{z}^{(k)} \right\|^2$$

获得解 $\boldsymbol{x}^{(k+1)}$。

步骤 2.2：更新乘子：

$$\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)} + 2\mu^{(k)} g\left(\boldsymbol{x}^{(k+1)}\right)$$

步骤 2.3：更新惩罚参数：

如果 $\left\|g\left(\boldsymbol{x}^{(k+1)}\right)\right\| < 0.25\left\|g\left(\boldsymbol{x}^{(k)}\right)\right\|$，设置 $\mu^{(k+1)} = \mu^{(k)}$；

否则设置 $\mu^{(k+1)} = 2\mu^{(k)}$。

当 $\left\|g\left(\boldsymbol{x}^{(k)}\right)\right\|$ 足够小时，算法 11.2 提前终止，获得问题 (11.1) 的近优解。