最小二乘法溯源追踪

Eason

目录

最小二乘法(Least square, LS) 最早由高斯提出。

高斯曰：

> The most probable value of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum.

1 引子

假设我们有四个观测点$(1,6),(2,5),(3,7),(4,10)$，我们希望找到一个线性函数$y=\beta_1+\beta_2 x$拟合这四个点。也就是说，我们希望找到β_1,β_2 适配这个系统。另外，我们在这里也假定

这个系统是线性系统，则有：

$$\begin{aligned} 6 &= \beta_1 + \beta_2 && (1)\\ 5 &= \beta_1 + 2\beta_2 && (2)\\ 7 &= \beta_1 + 3\beta_3 && (3)\\ 10 &= \beta_1 + 4\beta_3 && (4) \end{aligned}$$

以上有四个方程，两个未知数。如果系数矩阵的秩大于2，那么这个方程组就是过定方程组。

最小二乘的目的在于找到β_1, β_2使得式 (1)最小：

$$\begin{aligned} S(\beta_1, \beta_2) &= [6-(\beta_1+\beta_2)]^2 && (5)\\ &+ [5-(\beta_1+2\beta_2)]^2 && (6)\\ &+ [7-(\beta_1+3\beta_3)]^2 && (7)\\ &+ [10-(\beta_1+4\beta_3)]^2 && (8) \end{aligned}$$

$S(\beta_1, \beta_2)$进行求导，并令其等于零。得到：

$$\beta_1 = 3.5 \quad (9)$$

$$\beta_2 = 1.4 \quad (10)$$

最终的拟合结果为：$y = 3.5 + 1.4x$。

2 推广

考虑系统：

$$\sum_{j=1}^{n} A_{ij}\beta_j = y_i, i = 1, 2, \ldots, m \quad (11)$$

这个系统有m个观测值$y_i, i = 1, \ldots, m$，n个未知系数 $\beta_1, \ldots, \beta_n$。比如一个房子的属性有面积，单价，卧室数量，车位数量四个属性，根据这四个属性我们会对房子做一个估价y。我们想要拟合一个房价和这四个属性之间的线性关系。对应到式~(11)

$$y = \beta_1 H_1 + \beta_2 H_2 + \beta_3 H_3 + \beta_4 H_4 \quad (12)$$

其中H_1代表面积，H_2代表单价，H_3代表卧室数量， H_4代表车位数量，$\beta_1,\ldots,\beta_4$代表这四个属性的权重，y代表房价。这里我们假定这四个因素和总价y之间只存在线性关系。为了得到$\beta_1,\ldots,\beta_4$我们需要一些房子的数据表，根据这个表，我们来拟合这个关系。

对应到式 (11) ，我们有m个观测，每个观测涉及n个属性。把式~(11)写成矩阵的形式有：

$$\mathbf{Y}=\mathbf{A}\beta \tag{13}$$

其中：

$$\mathbf{Y}=[y_1,y_2,\ldots,y_m]$$

$\mathbf{Y}$是$m\times 1$的列向量。

$$\mathbf{A}=\begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ A_{21} & A_{22} & \ldots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mn} \end{bmatrix}$$

$\mathbf{A}$是$m\times n$的矩阵，m代表一共有m个观测数据， n 代表一共有n个属性。

$$\beta=[\beta_1,\beta_2,\ldots,\beta_n]$$

β是$n\times 1$的列向量，n代表n个属性。 β 向量代表了n个属性的权重。

通常，方程~(13) 没有解。我们的目的是找到一个最合适的 β 使得$\mathbf{Y}$与$\mathbf{A}\beta$的差值的平方最小。

$$\arg\min_{\beta} S(\mathbf{Y},\mathbf{A}\beta) \tag{14}$$

其中：

$$S(\mathbf{Y},\mathbf{A}\beta)=\sum_{i=1}^{m}|y_i-\sum_{j=1}^{n}A_{ij}\beta_j|^2=\|\mathbf{Y}-\mathbf{A}\beta\|^2 \tag{15}$$

假定$\mathbf{A}$的各列是相互独立的，也就是说n个属性互不相关，通过正则化方程~(13)：

$$\mathbf{A}^T\mathbf{Y}=\mathbf{A}^T\mathbf{A}\beta \tag{16}$$

$\mathbf{A}^T\mathbf{A}$是一个半正定的矩阵，其逆一定存在（假设 $\mathbf{A}$ 的各列是相互独立的）。最终，有：

$$\beta=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \tag{17}$$

式~(17)就是最小二成法的解。

3 推导正则方程

在上一节，我们不加说明的给出：式~(14)的解就是~(16)的解。现在，我们给出式 (16)的由来。

首先定义拟合值和观测值的差：

$$r_i = y_i - \sum_{j=1}^{n} A_{ij}\beta_j \tag{18}$$

那么：

$$S = \sum_{i=1}^{m} r_i^2 \tag{19}$$

为了求得S最小时的β，我们需要依S对 β求偏导。

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial \beta_j}, j = 1, 2, \ldots, n \tag{20}$$

另外：

$$\frac{\partial r_i}{\beta_j} = -H_{ij} \tag{21}$$

综上有：

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m}(y_i - \sum_{k=1}^{n} A_{ik}\beta_k)(-A_{ij}) \tag{22}$$

令：

$$2\sum_{i=1}^{m}(y_i - \sum_{k=1}^{n} A_{ik}\beta_k)(-A_{ij}) = 0 \tag{23}$$

重新组织式 (23)，有：

$$\sum_{i=1}^{m}\sum_{k=1}^{n} A_{ij}A_{ik}\beta_k = \sum_{i=1}^{m} A_{ij}y_i, j = 1, 2, \ldots, m \tag{24}$$

写成矩阵的形式有：

$$(\mathbf{A}^T\mathbf{A})\beta = \mathbf{A}^T\mathbf{Y} \tag{25}$$

啊哈，现在我们总算把最小二成准则和对应的方程联系起来了。锦上添花，我们再给出一种获取式 (25)的方式。这次我们以矢量的方式给出来。

定义：

$$S(\mathbf{Y}, \beta) = \|\mathbf{Y} - \mathbf{A}\beta\|^2 \tag{26}$$

展开有：

$$(\mathbf{Y}-\mathbf{A}\beta)^T(\mathbf{Y}-\mathbf{A}\beta)=\mathbf{Y}^T\mathbf{Y}-\beta^T\mathbf{A}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{A}\beta+\beta^T\mathbf{A}^T\mathbf{A}\beta \tag{27}$$

对式 (27) 就β求导，有：

$$\mathbf{A}^T\mathbf{Y}=(\mathbf{A}^T\mathbf{A})\mathbf{A}\beta \tag{28}$$

4 例子

假设有一个电阻，但是阻值没有标明，我们需要通过万用表测量以得到正确的阻值。但是，万用表也是有误差的，每次测量都会存在误差。假设我们测量了 k次，则有：

$$y_1 = x+n_1 \tag{29}$$

$$y_2 = x+n_2 \tag{30}$$

$$\vdots \quad \vdots \tag{31}$$

$$y_k = x+n_k \tag{32}$$

$$\tag{33}$$

矩阵形式有：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} x + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_k \end{bmatrix} \tag{34}$$

根据$\hat{x}=(A^TA)^{-1}A^Ty$，则：

$$\hat{x}=\frac{1}{k}\sum_{i=1}^{k}y_i \tag{35}$$

可以看出来，在这个例子中，最小二乘法结果与多次测量求均值的直观感觉一致。更进一步：

$$\hat{x}=\frac{1}{k}\sum_{i=1}^{k}y_i=x+\frac{1}{k}\sum_{i}^{k}n_i \tag{36}$$

在通信系统里$y=x+n$是AWGN信道的模型。多测测量等效于多次发送同一信息，然后经过不同的噪声污染。多次发送求平均之后，信号功率没有变化，噪声功率却变为原来的 $\frac{1}{n}$。噪声是误差的来源。在统计学中，误差不可能被消除，只能被降低。多次测量求均值的过

程就是降低误差的过程，即降低噪声的过程。在通信系统中，噪声不能被消除，只能被抑制。无论在通信系统还是统计学领域，其背后的数学原理都是一样的。

注意在这里我们的目标是获取x的最优估计，x是我们要估计的量。之前我们给出来的曲线拟合的例子里，我们要获取的是多项式的系数。虽然场景不同，但是其背后的数学原理相同。

5 加权的最小二乘法

之所以会出现加权的最小二乘法，是因为并不是每次的观测可靠度都是一样的。比如之前的测量电阻阻值的例子，有可能有些测量是用昂贵的仪器测量的，有些测量是用便宜的仪器测量的。这两类仪器的测量结果当然不同，我们不能对其平等看待。所以就要对那些可靠的结果加上较大的权重，对那些稍微不可靠的结果加上较小的权重。

我们先把这个电阻的例子放在一边。考虑加权的最小二乘法的数学描述。假设 x是一个矢量，每个元素都是常数。系统方程为：

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_k \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} & \dots & H_{1n} \\ H_{21} & H_{22} & \dots & H_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k1} & H_{k2} & \dots & H_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \dots \\ n_k \end{bmatrix} \tag{37}$$

且$E(n_i^2) = \sigma_i^2$，即n_i是均值为零，方差为 σ_i^2的高斯噪声。假设$n_i, i = 1, \dots, k$是相互独立的，则有：

$$R = E[\mathbf{n}\mathbf{n}^T] = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \sigma_k^2 \end{bmatrix} \tag{38}$$

不同的σ_i^2意味着其对应的测量结果的权重不同。σ_i^2越小，则权重越大；反之，越小。

那么现在我们的损失函数（优化目标）是：

$$J = \epsilon_{y_1}^2 / \sigma_1^2 + \dots + \epsilon_{y_k}^2 / \sigma_k^2 \tag{39}$$

其中$\epsilon_{y_i} = y_i - \sum_{j=1}^{n} H_{ij}x_j$. 所以$J$又可以被写为：

$$\begin{aligned} J &= \epsilon_y^T R^{-1} \epsilon_y && (40)\\ &= (y - H\hat{x})^T R^{-1}(y - H\hat{x}) && (41)\\ &= (y^T R^{-1} - \hat{x}^T H^T R^{-1})(y - H\hat{x}) && (42)\\ &= y^T R^{-1} y - y^T R^{-1} H\hat{x} - x^T H^T R^{-1} y + \hat{x}^T H^T R^{-1} H\hat{x} && (43) \end{aligned}$$

上式以$\hat{x}$求微分，令其等于零，有：

$$\frac{\partial J}{\partial \hat{x}} = -y^T R^{-1} H + \hat{x}^T H^T R^{-1} H = 0 \tag{44}$$

进而，有：

$$\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} y \tag{45}$$

注意这个解要求R^{-1}存在，也就是说每一次测量都要求y_i受点噪声干扰。但是，也可以看出来，如果这个干扰很小的话，会出现R^{-1}对角线上的值非常大，在实际应用中会出现溢出现象。

现在让我们回到测量电阻的那个例子。系统方程仍然不变：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} x + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_k \end{bmatrix} \tag{46}$$

考虑噪声的协方差矩阵：

$$R = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2) \tag{47}$$

根据式 (45)，有：

$$\begin{aligned} \hat{x} &= (H^T R^{-1} H)^{-1} H^T R^{-1} y && (48)\\ &= (\sum_{i=1}^{k} 1/\sigma_i^2)^{-1} \sum_{j=1}^{k} \frac{y_j}{\sigma_j^2} && (49) \end{aligned}$$

6 迭代的最小二乘法

之前，我们对最小二乘法的原理和正则方程进行了推导。在之前推导最小二乘的过程中，我们假定不会有新的观测数据，所以系数矩阵都是固定大小的。这样的方法有个问题：如果有新的数据到来，那么之前的计算需要重新来过。

在本节，我们探讨如何迭代的更新$\hat{x}$，即：当我们已经根据$k-1$个估计数据得到了$\hat{x}$的一个估计，现在又来了一个数据y_k，如何不用重新计算 (45)更新$\hat{x}$.

一个线性迭代估计可以表示为：

$$y_k = H_k x + n_k \tag{50}$$

$$\hat{x}_k = \hat{x}_{k-1} + K_k(y_k - H_k \hat{x}_{k-1}) \tag{51}$$

我们根据$\hat{x}_{k-1}$和y_k更新$\hat{x}_k$。K_k是估计算子的增益矩阵，将在后文推导其计算过程。$y_k - H_k\hat{x}_{k-1}$是修正项。如果$K_k = 0$或者修正项为零，则$\hat{x}_k = \hat{x}_{k-1}$。在计算$K_k$之前，我们首先考虑迭代估计的误差期望：

$$\begin{aligned}
E[\epsilon_{x,k}] &= E[x - \hat{x}_k] && (52)\\
&= E[x - \hat{x}_{k-1} - K_k(y_k - H_k\hat{x}_{k-1})] && (53)\\
&= E[x - \hat{x}_{k-1} - K_k(H_k x + n_k - H_k\hat{x}_{k-1})] && (54)\\
&= E[\epsilon_{x,k-1} - K_k H_k(x - \hat{x}_{k-1}) - K_k n_k] && (55)\\
&= E[(\epsilon_{x,k-1} - K_k H_k \epsilon_{x,k-1}) - K_k n_k] && (56)\\
&= (I - K_k H_k)E[\epsilon_{x,k-1}] - K_k E[n_k] && (57)
\end{aligned}$$

如果$E[n_k] = 0$，$E[\epsilon_{x,k-1}] = 0$，那么$E[\epsilon_{x,k}] = 0$。也就是说，如果估计噪声是零均值的，并且初始的估计$x_0 = x$，那么此后所有的估计都是x。从这点看来，式 (50)给出来的估计是无偏的。并且这个特性与K_k无关，这是一个无偏估计算子该有的样子。接下来我们计算K_k。

6.1 RLS的K_k

由于K_k是无偏估计，我们需要小心的选择K_k。我们选择最小估计误差方差和作为估计准则。即：

$$\begin{aligned}
J_k &= E[(x_1 - \hat{x}_k)^2] + \ldots + E[(x_k - \hat{x}_k)^2] && (58)\\
&= E[\epsilon_{x_1,k}^2 + \epsilon_{x_2,k}^2 + \ldots + \epsilon_{x_n,k}^2] && (59)\\
&= E[\epsilon_{x,k}^T \epsilon_{x,k}] && (60)\\
&= E[Tr[\epsilon_{x,k}\epsilon_{x,k}^T]] && (61)\\
&= E[P_k] && (62)
\end{aligned}$$

其中P_k是误差的协方差矩阵。其迭代计算过程为：

$$\begin{aligned} P_k &= E[\epsilon_{x,k}^T \epsilon_{x,k}] && (63) \\ &= E\{[(I-K_kH_k)\epsilon_{x,k-1} - K_k n_k][\ldots]^T\} = \\ & (I-K_kH_k)E(\epsilon_{x,k-1}\epsilon_{x,k-1}^T)(I-K_kH_k)^T - \\ & K_kE(n_k\epsilon_{x,k-1}^T)(I-K_kH_k)^T - (I-K_kH_k)E(\epsilon_{x,k-1}n_k^T)K_k^T + \\ & K_kE(v_kv_k^T)K_k^T && (64) \end{aligned}$$

注意估计误差和观测噪声是相互独立的，所以：

$$E(n_k\epsilon_{x,k-1}^T) = E(n_k)E(\epsilon_{x,k-1}) = 0 \tag{65}$$

所以式 (63) 变为：

$$P_k = (I-K_kH_k)P_{k-1}(I-K_kH_k)^T + K_kR_kK_k^T \tag{66}$$

其中，R_k是噪声协方差。式 (66)是最小二成估计误差的迭代公式。根据这个公式，当n_k方差变大时，R_k中对角线上的值也会变大，进而P_k变大，这个过程符合直觉。现在我们需要找到K_k使得式 (58)最小。对于任何K_k，估计误差的期望都是零，所以当我们选定K_k使得J_k最小时，不仅估计误差期望为零，而且会一直为零。所以：

$$\frac{\partial J_k}{\partial K_k} = 2(I-K_kH_k)P_{k-1}(-H_k^T) + 2K_kR_k \tag{67}$$

令上式为零，我们可以得到：

$$K_kR_k = (I-K_kH_k)P_{k-1}H_k^T \tag{68}$$

整理，得：

$$K_k = P_{k-1}H_k^T(H_kP_{k-1}H_k^T + R_k)^{-1} \tag{69}$$

现在我们整理一下迭代最小二乘法的步骤：

1. 初始化估计子：

$$\begin{aligned} \hat{x}_0 &= E(x) && (70) \\ P_0 &= E[(x-\hat{x}_0^T)(x-\hat{x}_0^T)] && (71) \end{aligned}$$

如果没有关于x的任何先验信息，则令$P_0 = \infty I$；如果对 x有比较好的猜测，则可以令$P_0 = 0$

2. 对于$k=1,2,\ldots$有：

$$
\begin{aligned}
y_k &= H_k x_k + n_k && (72)\\
H_k &= P_{k-1}H_k^T(H_k P_{k-1} H_k^T + R_k)^{-1} && (73)\\
\hat{x}_k &= \hat{x}_{k-1} + K_k(y_k - H_k\hat{x}_{k-1}) && (74)\\
P_k &= (I - K_k H_k)P_{k-1}(I - K_k H_k)^T + K_k R_k K_k^T && (75)
\end{aligned}
$$