曲线拟合

《计算方法》课程笔记

2024 年 11 月 19 日

§1 引言

多项式插值的前提: 节点横纵坐标数值上是精确的. 但是科学实验和生产实践中, 很难满足! 一来, 观测数据数目往往很多, 不可避免带有测量误差; 此外, 插值曲线通过所有节点, 会使曲线保留着一切测量误差. 最后, 个别数据误差较大时, 导致插值实际效果不理想. 我们提出另一种函数逼近方法: 曲线拟合.

可由给定数据 (x_i, y_i), 构造出近似函数 $\varphi(x)$, 不求 $\varphi(x)$ 完全通过所有节点, 但求近似曲线能反映出数据的基本趋势. 这要满足

1. 数据点均在离此拟合曲线的不远处的上方或下方位置
2. 曲线能反映数据总体分布, 局部不会出现较大波动, 能反映被逼近函数的特性
3. 曲线函数与已知函数从总体上来说, 其偏差按某种方法度量达到最小

§2 最小二乘曲线拟合

例如, 测量或试验得到数值表

x_0	x_1	x_2	$\cdots$	x_n
y_0	y_1	y_2	$\cdots$	y_n

, 现在区分几个基本概念:

- 函数插值是插值函数 $P(x)$ 与被插函数 $f(x)$ 在节点 x_i 处函数值相同, 即 $P(x_i) = f(x_i)(i = 0, 1, \ldots, n)$;
- 拟合函数 $\varphi(x)$ 不要求曲线严格通过所有数据点 (x_i, y_i), 即 $\varphi(x)$ 在节点 x_i 处的偏差 (亦称残差): $\epsilon_i = \varphi(x_i) - f(x_i), i = 0, 1, \cdots, n$ 不都严格地等于零.

为使拟合曲线尽可能反映所给数据的变化趋势, 要求 $|\epsilon_i|$ 按某种度量标准最小. 记向量 $\epsilon = [\epsilon_0, \epsilon_1, \ldots, \epsilon_n]$, 即要求残差向量 ϵ 的某种向量范数 $||\epsilon||$ 最小.

一种想法是使用残差向量 $e = (\epsilon_1, \cdots, \epsilon_n)$ 的一范数. 即 $\frac{1}{n}||e||_1 = \frac{1}{n}\sum_{i=0}^{n} |\epsilon_i|$ 最小; 但是它太复杂. 如果考虑无穷维范数, 即 $||e||_\infty = \max_i |\varphi(x_i) - f(x_i)|$, 它并不可导. 因此求解困难.

一个简单的想法是使用二范数. 即

$$||e_2||^2 = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=0}^{n} (\varphi(x_i) - f(x_i))^2.$$

2.1. 直线拟合. 对于给定的数据点 $(x_i, y_i)(i = 1, 2, \ldots, m)$, 求拟合直线 $y_0 = a_0 + a_1 x$, 使得总误差最小.

也就是说决定 a_0, a_1, 使得 $F(a_0, a_1) = \sum_{i=1}^{m}(a_0 + a_1 x_i - y_i)^2$ 的取值最小.

根据多元函数的极值条件, 一个必要的条件是

$$\begin{cases} \frac{\partial F(a_0,a_1)}{\partial a_0} = 0 \\ \frac{\partial F(a_0,a_1)}{\partial a_1} = 0 \end{cases}$$

采用矩阵的语言来描述, 设模型的参数向量 $\mathbf{a} = (a_0, a_1)' \in \mathbb{R}^{2\times1}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \in \mathbb{R}^{m\times2}$ 这里 $\mathbf{X}$ 除了各个 x_i 之外, 左侧还有一列: 这是因为它将要根据 $\mathbf{a}$ 的参数, 对每一个输入 x_i, 输出模型认为应该有的值, 即, $\mathbf{Xa}$ 希望把每一个 x 变为 $a_0 + a_1 x$. 最后, 我们把观测值放在一个向量里面 $\mathbf{y} = (y_1, y_2, \cdots, y_m)' \in \mathbb{R}^{m\times1}$. 要优化的对象为

$$F(\mathbf{a}) = ||\mathbf{y} - \mathbf{Xa}||^2 = (\mathbf{y} - \mathbf{Xa})'(\mathbf{y} - \mathbf{Xa}).$$

要求 F 的最小值, 根据向量求导的法则, 可求得 $\frac{\partial F}{\partial \mathbf{a}} = 2 \cdot (-1)\mathbf{X}'(\mathbf{y} - \mathbf{Xa})$. 根据多元函数的极值条件, 就有 $\frac{\partial F}{\partial \mathbf{a}} = \mathbf{0}$, 就得到了当前函数极小值的必要条件.

这就表明 $\mathbf{X}'(\mathbf{y} - \mathbf{Xa}) = \mathbf{0}$, 实际上我们发现残差向量 $(\mathbf{y} - \mathbf{Xa})$ 和 $\mathbf{X}$ 是垂直的. 展开上述的向量形式, 就得到了

$$\begin{cases} a_0 m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \\ a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i \end{cases}$$

只要我们解答这个方程组, 即可得到直线的拟合方程.

当然, 上述内容只是一个必要条件. 更多的数学知识可以说明这个条件是必要的, 我们这里不展开叙述.

例 1. 假设有以下的实验数据

i	1	2	3	4
x_i	1.36	1.73	1.95	2.28
y_i	14.094	16.844	18.475	20.963

此例子容易在大题出现. 建议直接使用标量形式, 绕过矩阵进行解答.

使用最小二乘法用直线拟合上述函数.

根据推导过程, 我们发现 $\mathbf{X} = \begin{bmatrix} 1 & 1.36 \\ 1 & 1.73 \\ 1 & 1.95 \\ 1 & 2.28 \end{bmatrix}$, $\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 14.094 \\ 16.844 \\ 18.745 \\ 20.963 \end{bmatrix}$. 残差

向量为

$$\mathbf{y}-\mathbf{X}\mathbf{a}=\begin{bmatrix}14.094\\16.844\\18.745\\20.963\end{bmatrix}-\begin{bmatrix}a_0+1.36a_1\\a_0+1.73a_1\\a_0+1.95a_1\\a_0+2.28a_1\end{bmatrix}=\begin{bmatrix}14.094-a_0-1.36a_1\\16.844-a_0-1.73a_1\\18.745-a_0-1.95a_1\\20.963-a_0-2.28a_1\end{bmatrix}.$$

并且

$$\mathbf{X}'=\begin{bmatrix}1&1&1&1\\1.36&1.73&1.95&2.28\end{bmatrix}$$

$\mathbf{X}'\mathbf{Y}=\mathbf{0}$ 意味着 $\begin{cases}-4a_0-7.32a_1+70.376&=0\\-7.32a_0-13.8434a_1+132.12985&=0\end{cases}$, 解得 $a_0=3.9374, a_1=7.4626$. 于是拟合得到 $y=3.9374+7.4626x$.

2.2. 多项式拟合. 数据点的分布并不一定呈一条直线, 此时可用多项式拟合. 对于给定的一组数据 $(x_i,y_i)(i=1,2,\ldots,n)$, 寻求次数不超过 $m(m<n)$ 的多项式 $y(x)=a_0+a_1x+a_2x^2+\ldots+a_mx^m$ 来拟合所给定的数据, 与线性拟合类似, 使偏差的平方和

$$Q=\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{m}a_jx_i^j\right)^2$$

最小.

同样适用向量的形式表示这些数据. 我们设 $\mathbf{a}=(a_0,a_1,a_2,\cdots,a_m)'\in\mathbb{R}^{(m+1)\times1}$ 为多项式前面的系数; $\mathbf{y}=(y_1,y_2,\cdots,y_n)'\in\mathbb{R}^{n\times1}$. 实际上, 只要将样本的矩阵 $\mathbf{X}$ 改变为 $\mathbf{X}=\begin{bmatrix}1&x_1&x_1^2&\cdots&x_1^m\\1&x_2&x_2^2&\cdots&x_2^m\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&x_n&x_n^2&\cdots&x_n^m\end{bmatrix}\in\mathbb{R}^{n\times(m+1)}$, 这样子 $\mathbf{Xa}$ 就是把某一个 a_i 转换为 $\sum_{j=0}^{m}a_jx_i^j$ 了. 和刚刚一样, 我们只要展开 $\mathbf{X}'(\mathbf{Y}-\mathbf{Xa})=\mathbf{0}$ 为标量形式, 就得到了最小二乘法对应的多项式拟合形式.

实际上, 我们发现, 只要 $\mathbf{X}$ 可以被恰当选择, 我们就可以用最小二乘法进行拟合. 但是, 当 n 较大时 $(n\geq7)$, 系数矩阵条件数较大, 法方程组往往是病态的, 给求解工作带来了困难.

例 2. 有一组实验数据如下:

i	1	2	3	4	5	6
x_i	0	1	2	3	4	5
y_i	5	2	1	1	2	3

此例子容易在大题出现. 建议直接使用标量形式, 绕过矩阵进行解答.

使用最小二乘法以抛物线形式拟合这组数据.

取 $\mathbf{X}=\begin{bmatrix}1 & x_1 & x_1^2\\ 1 & x_2 & x_2^2\\ \vdots & \vdots & \vdots\\ 1 & x_n & x_n^2\end{bmatrix}$, $\mathbf{y}=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$, $\mathbf{a}=\begin{bmatrix}a_0\\ a_1\\ a_2\end{bmatrix}$. 实际上,$\mathbf{X}'=\begin{bmatrix}1 & 1 & \cdots & 1\\ x_1 & x_2 & \cdots & x_n\\ x_1^2 & x_2^2 & \cdots & x_n^2\end{bmatrix}$,

$\mathbf{y}-\mathbf{Xa}=\begin{bmatrix}y_1-(a_0+a_1x_1+a_2x_1^2)\\ y_2-(a_0+a_1x_2+a_2x_2^2)\\ \vdots\\ y_n-(a_0+a_1x_n+a_2x_n^2)\end{bmatrix}$, 我们留意到每乘积结果的形态如同

$$\begin{bmatrix}\sum_{i=1}^n y_i-na_0-a_1\sum_{i=1}^n x_i-a_2\sum_{i=1}^n x_i^2\\ \sum_{i=1}^n x_iy_i-a_0\sum_{i=1}^n x_i-a_1\sum_{i=1}^n x_i^2-a_2\sum_{i=1}^n x_i^3\\ \sum_{i=1}^n x_i^2y_i-a_0\sum_{i=1}^n x_i^2-a_1\sum_{i=1}^n x_i^3-a_2\sum_{i=1}^n x_i^4\end{bmatrix}.$$

于是让他们全都等于 0 就得到了方程组 $\begin{cases}6a_0+15a_1+55a_2=14\\ 15a_0+55a_1+225a_2=30\\ 55a_0+225a_1+979a_2=122\end{cases}$. 于是得到的方程为 $y(x)=4.7143-2.7857x+0.5000x^2$.

实际上, 对于一般的次数, 上述规律可以进行推广. 具体就是在 $\sum_{i=1}^n y_i$ 项里面插入高次的 x_i^k, 形成 $\sum_{i=1}^n x_i^k y_i$ 的样子; 后续的 $a_k\sum_{i=1}^n x_i^l$ 中继续插入 x_i, 形成 $a_k\sum_{i=1}^n x_i^{l+1}$ 的形式.

2.3. 做变量代换. 有些非线性拟合曲线可以通过适当的变量替换转化为线性曲线, 从而用线性拟合进行处理. 如下所示.

曲线拟合方程	变换关系	变换后的线性拟合方程
$y=ax^b$	$Y=\ln y, X=\ln x$	$Y=\ln a+bX$
$y=ax^\mu+c$	$X=x^\mu$	$Y=aX+c$
$y=\frac{x}{ax+b}$	$Y=\frac{1}{y}, X=\frac{1}{x}$	$Y=a+bX$
$y=\frac{1}{ax+b}$	$Y=\frac{1}{y}$	$Y=b+ax$
$y=\frac{1}{ax^2+bx+c}$	$Y=\frac{1}{y}$	$Y=ax^2+bx+c$
$y=\frac{x}{ax^2+bx+c}$	$Y=\frac{x}{y}$	$Y=ax^2+bx+c$

例 3. 设某实验数据如下所示. 取指数函数 $y=ae^{-bx}$ 作为拟合函数, 使用最小二乘法拟合.

此例子容易在大题出现.

i	1	2	3	4	5	6
x_i	0	0.5	1	1.5	2	2.5
y_i	2.0	1.0	0.9	0.6	0.4	0.3

对于 ae^{-bx} 而言, 两边取对数得 $\underbrace{\ln y}_{Y}=\underbrace{\ln a}_{a_0}\underbrace{-b}_{a_1}x$. 可得到线性模型. 由于

$n = 6, m = 1$, 法方程组为

$$\begin{cases} 6a_0 + a_1 \sum_{i=1}^{6} x_i = \sum_{i=1}^{6} Y_i \\ a_0 \sum_{i=1}^{6} x_i + a_1 \sum_{i=1}^{6} x_i^2 = \sum_{i=1}^{6} x_i Y_i \end{cases}$$

计算相应的数据, 得到 $\begin{cases} 6a_0 + 7.5a_1 = -2.43302 \\ 7.5a_0 + 13.75a_1 = -5.714112 \end{cases}$. 得 $a_0 = 0.562302, a_1 = -0.772282$. 于是 $a = e^{-bx}$, 得到 $a = e^{0.562302} = 1.754708$,$b = 0.772282$. 其拟合指数函数为 $y = 1.754708e^{-0.772282x}$.

2.4. 对于使用计算器的附注. 使用 FX-999 CN CW 的计算器时候, 可以进入双变量统计模式, 并且可以从参数列表中直接读出对应的表达式.

倘若没有, 就要自己做变量代换, 化为一次或二次拟合的形式. 最后把数据换为原始形式.