第四章-插值与逼近

(4) 最佳平方逼近和最小二乘拟合

目录

问题背景

问题：找到一个函数 P_n（比如多项式）使得 $P_n \approx f$

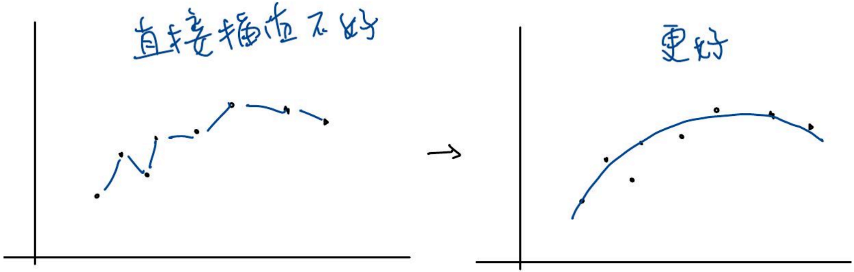
插值法的思路：我们对于部分节点数据 (x_i, f_i)，我们找到多项式 $P_n(x_i) = f_i$

我们需要量化距离，比如使用 $D_\infty(P_n, f)$，就此引出一致收敛的概念

插值法在实际问题中的限制：

- 对于带噪声的数据点不友好：如果 $\tilde{f}_i = f_i +$ 噪声，则插值 $P_n(x_i) = \tilde{f}_i$ 显得不是很合理
- 直接的插值法不太直接适合高维问题
- $\cdots$

例子



目标

逼近部分需要解决的问题：

- 对于既定的函数 f，找到最优的近似函数
- 基于数据，找到最优的近似函数

我们可以考虑不同类型的函数族，在本门课中，我们主要会集中于多项式

此处的知识可以延展到偏微分方程的数值解、以及机器学习问题

具体情景

问题：我们不想使用任何高次多项式（高次往往意味着高代价）；给定最高 n 次，最佳的近似函数是什么？

参数化： $P_n(x) = a_0 + a_1 x + \cdots + a_n x^n$；需要找到最好的系数 $a_0, a_1, \cdots, a_n$

方法： 求解如下的优化问题

$$\min_{a_0, a_1, \cdots, a_n} D(P_n, f)$$

其中 $D(\cdot, \cdot)$ 代表两个函数之间的距离/差异。

函数插值部分我们大量使用了 D_∞：

优势 代表最大可能的误差

缺点 该选择不是很好求导

出于易于求导数的原因，后续我们将使用 D_2

将所有次数不超过 n 次多项式集合为 H_n；我们希望求解

$$\begin{aligned}\|f-P_n^*\|_2 &= \min_{P\in H_n}\|f-P\|_2\\ &= \min_{P\in H_n}\sqrt{\int_a^b \big(f(x)-P(x)\big)^2\,\mathrm{d}x}\end{aligned}$$

最小化平方距离的函数 P_n^* 被称为 f 在区间 $[a,b]$ 上的**最佳平方逼近**

注：若我们使用 D_∞，此处类似的最佳多项式被称为最佳一致逼近多项式（课本 3.2 节）（不要求）。

目录

核心思路

考虑基函数 $\varphi_k(x) = x^k$，目标函数 $f \in C[0,1]$。我们解如下的优化问题：

$$\min_{\{a_k\}_{k=0}^n} \int_0^1 \left(f(x) - \sum_{k=0}^n a_k \varphi_k(x)\right)^2 \mathrm{d}x$$

$$= \min_{\{a_k\}_{k=0}^n} \int_0^1 f(x)^2 \mathrm{d}x - 2\sum_{k=0}^n a_k \int_0^1 f(x)\varphi_k(x) \mathrm{d}x + \sum_{j,k=0}^n a_j a_k \int_0^1 \varphi_k(x)\varphi_j(x) \mathrm{d}x$$

为了简化符号，我们标记 $G_{j,k} = \int_0^1 \varphi_k(x)\varphi_j(x) \mathrm{d}x$，$d_k = \int_0^1 f(x)\varphi_k(x) \mathrm{d}x$：

上述问题等价于求解 $$\min_{\{a_k\}_{k=0}^n} -2\sum_{k=0}^n a_k d_k + \sum_{j,k=0}^n a_j a_k G_{j,k}$$

问题：为什么 $\int_0^1 f(x)^2 \mathrm{d}x$ 在找最佳系数的过程可以被扔掉？

我们需要优化的函数其实是一个二次函数:

$$\min_{\{a_k\}_{k=0}^n} -2\sum_{k=0}^n a_k d_k + \sum_{j,k=0}^n a_j a_k G_{j,k} = L(a_0, a_1, \cdots, a_n)$$

关于 a_k 求导可知,

$$\frac{\mathrm{d}L(a_0, a_1, \cdots, a_n)}{\mathrm{d}a_k} = -2d_k + 2a_k G_{k,k} + 2\sum_{j\neq k} a_j G_{k,j} = -2d_k + 2\sum_{j=0}^n G_{k,j} a_j$$

局部极小值满足对于任何的 k, $\frac{\mathrm{d}L(a_0,a_1,\cdots,a_n)}{\mathrm{d}a_k} = 0$; 再通过简化可得,

$$\boldsymbol{G}\boldsymbol{a} = \boldsymbol{d}$$

$$\boldsymbol{G} = \begin{bmatrix} G_{00} & G_{01} & \cdots & G_{0n} \\ \vdots & \vdots & \vdots & \vdots \\ G_{n0} & G_{n1} & \cdots & G_{nn} \end{bmatrix} \qquad \boldsymbol{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} \qquad \boldsymbol{d} = \begin{bmatrix} d_0 \\ \vdots \\ d_n \end{bmatrix}$$

在这里我们需要计算

$$G_{j,k} = \int_0^1 \varphi_k(x)\varphi_j(x)\ \mathrm{d}x \qquad d_k = \int_0^1 f(x)\varphi_k(x)\ \mathrm{d}x$$

对于 $\varphi_k(x) = x^k$ 的例子，我们可知

$$G_{j,k} = \frac{1}{k+j+1}$$

对应的矩阵 $\boldsymbol{G} = \begin{bmatrix} \frac{1}{1} & \frac{1}{2} & \cdots & \frac{1}{1+n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{2+n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n+1} \end{bmatrix}$ 是 Hilbert 矩阵

练习

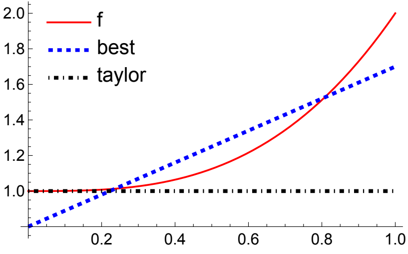
考虑 $f(x) = 1 + x^3$，求它在区间 $[0, 1]$ 上的一次最佳平方逼近多项式。请用基函数 $\varphi_0(x) = x^0$, $\varphi_1(x) = x^1$，计算法方程中的 $\boldsymbol{G}$ 和 $\boldsymbol{d}$。

练习

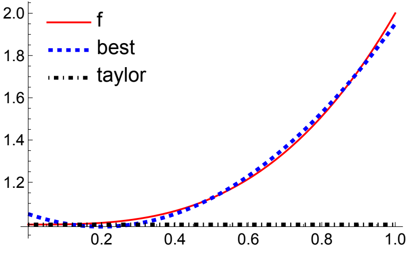
考虑 $f(x) = 1 + x^3$，求它在区间 $[0,1]$ 上的一次最佳平方逼近多项式。请用基函数 $\varphi_0(x) = x^0$, $\varphi_1(x) = x^1$，计算法方程中的 $\boldsymbol{G}$ 和 $\boldsymbol{d}$。

答案： $\boldsymbol{G} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}$, $\boldsymbol{d} = \begin{bmatrix} 5/4 \\ 7/10 \end{bmatrix}$

使用 $\{1, x\}$ 为基函数



使用 $\{1, x, x^2\}$ 为基函数



对于两个函数 f, g，计算

$$\int_a^b f(x)g(x)\ \mathrm{d}x$$

是一个常见的数学场景，它背后具有更多的基础理论

目录

函数的内积

根据对于不同节点 i 的权重，我们可以定义一般的内积

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_w = \sum_{i=1}^{n} x_i y_i w_i$$

其中 $w_i > 0$ 代表权重

向量的大小

$$\begin{aligned} &\|\boldsymbol{x}\|_2 \\ =&\sqrt{x_1^2 w_1 + x_2^2 w_2 + \cdots + x_n^2 w_n} \\ =&\sqrt{\langle x, x \rangle_w} \end{aligned}$$

假设在区间 $[a, b]$ 上，ρ 是一个非负函数，则 ρ 可以类似承担权重的作用。函数 f, g 的内积可以被定义为：

$$\langle f, g \rangle_\rho = \int_a^b f(x) g(x) \rho(x) \, \mathrm{d}x.$$

函数的 L^2 范数：

$$\|f\|_2 = \sqrt{\int_a^b f^2(x) \rho(x) \, \mathrm{d}x} = \sqrt{\langle f, f \rangle_\rho}$$

注：当权重 ρ 可从上下文看出，或权重可以是任意的，或 $\rho = 1$ 时，我们为简化有时用 $\langle f, g \rangle$ 来代替 $\langle f, g \rangle_\rho$

对于向量，

$$
\begin{aligned}
\langle \boldsymbol{x}, \boldsymbol{y} \rangle &= \langle \boldsymbol{y}, \boldsymbol{x} \rangle \\
\langle c\boldsymbol{x}, \boldsymbol{y} \rangle &= c\langle \boldsymbol{x}, \boldsymbol{y} \rangle \\
\langle \boldsymbol{x}_1 + \boldsymbol{x}_2, \boldsymbol{y} \rangle &= \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle + \langle \boldsymbol{x}_2, \boldsymbol{y} \rangle \\
\langle \boldsymbol{x}, \boldsymbol{x} \rangle &\geq 0
\end{aligned}
$$

当且仅当$\boldsymbol{x} = \mathbf{0}$, $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$.

Cauchy-Schwarz 不等式

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$$

三角不等式

$$\|\boldsymbol{x} + \boldsymbol{y}\|_2 \leq \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2$$

对于函数，

$$
\begin{aligned}
\langle f, g \rangle &= \langle g, f \rangle \\
\langle cf, g \rangle &= c\langle f, g \rangle \\
\langle f_1 + f_2, g \rangle &= \langle f_1, g \rangle + \langle f_2, g \rangle \\
\langle f, f \rangle &\geq 0
\end{aligned}
$$

当且仅当$f = 0$, $\langle f, f \rangle = 0$.

Cauchy-Schwarz 不等式

$$|\langle f, g \rangle| \leq \|f\|_2 \|g\|_2$$

三角不等式

$$\|f + g\|_2 \leq \|f\|_2 + \|g\|_2$$

对于向量，若 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$，我们称 $\boldsymbol{x}$，$\boldsymbol{y}$ 正交（或者垂直）

对于函数，若 $\langle f, g \rangle = 0$，我们称 f，g 正交（或者垂直）

为了强调权重 ρ，我们也会讲f，g 在 $[a, b]$ 上带权 ρ 正交

若向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots$ 满足两两互相垂直，则称为正交向量组

若函数 $\varphi_0, \varphi_1, \cdots$ 满足两两互相垂直，则称为**正交函数族**

若我们进一步有 $\langle \varphi_j, \varphi_k \rangle = \delta_{j,k}$，则称 $\{\varphi_k\}$ 为**标准正交函数族**

例子：在区间 $[0, 1]$ 上，$\varphi_k(x) = \cos(2\pi kx)$，$k = 0, 1, 2, \cdots$，组成正交函数族。

一些性质

若

$$a_1\boldsymbol{x}_1+\cdots+a_n\boldsymbol{x}_n=0$$

当且仅当 $a_1=a_2=\cdots=a_n=0$，则称向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 线性无关

假设 $\varphi_0,\varphi_1,\cdots,\varphi_{n-1}$ 在区间 $[a,b]$ 上连续，若

$$a_0\varphi_0+a_1\varphi_1+\cdots+a_{n-1}\varphi_{n-1}=0$$

当且仅当 $a_0=a_1=\cdots=a_{n-1}=0$ 时成立，则称 $\varphi_0,\varphi_1,\cdots,\varphi_{n-1}$ 在 $[a,b]$ 上线性无关。

定理

$\varphi_0, \varphi_1, \cdots, \varphi_{n-1}$ 在 $[a, b]$ 上线性无关的充要条件是 *Cramer* 行列式 $det(\boldsymbol{G}_{n-1}) \neq 0$,

$$\boldsymbol{G}_{n-1} = \begin{bmatrix} \langle \varphi_0, \varphi_0 \rangle & \langle \varphi_0, \varphi_1 \rangle & \cdots & \langle \varphi_0, \varphi_{n-1} \rangle \\ \langle \varphi_1, \varphi_0 \rangle & \langle \varphi_1, \varphi_1 \rangle & \cdots & \langle \varphi_1, \varphi_{n-1} \rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle \varphi_{n-1}, \varphi_0 \rangle & \langle \varphi_{n-1}, \varphi_1 \rangle & \cdots & \langle \varphi_{n-1}, \varphi_{n-1} \rangle \end{bmatrix}$$

目录

问题 1：改写函数逼近的一般情况

考虑基函数 φ_k，权重函数 ρ，目标函数 $f \in C[a,b]$。希望解如下的优化问题：

$$\begin{aligned}
&\min_{\{a_k\}_{k=0}^n} \int_a^b \left(f(x) - \sum_{k=0}^n a_k \varphi_k(x)\right)^2 \rho(x)\,\mathrm{d}x \\
=&\min_{\{a_k\}_{k=0}^n} \langle f, f\rangle_\rho - 2\sum_{k=0}^n a_k \langle f, \varphi_k\rangle_\rho + \sum_{j,k=0}^n a_j a_k \langle \varphi_j, \varphi_k\rangle_\rho
\end{aligned}$$

为了简化符号，标记 $G_{j,k} = \langle \varphi_j, \varphi_k\rangle_\rho$，$d_k = \langle f, \varphi_k\rangle_\rho$：

上述问题等价于求解 $$\min_{\{a_k\}_{k=0}^n} -2\sum_{k=0}^n a_k d_k + \sum_{j,k=0}^n a_j a_k G_{j,k}$$

最后通过求导可知，我们仅需求解

$$\sum_{j=0}^{n} G_{k,j} a_j = d_k \qquad \text{法方程}$$

即求解

$$\boldsymbol{G}\boldsymbol{a} = \boldsymbol{d}$$

问题 2：计算误差

若已经求解到最佳的系数 $a_0^*, a_1^*, \cdots$ ，且最佳的近似函数标记为 P_n^*

$$\begin{aligned}\|f - P_n^*\|_2^2 &= \langle f - P_n^*, f - P_n^* \rangle \\ &= \langle f, f \rangle - 2\boldsymbol{d}^T \boldsymbol{a}^* + \left(\boldsymbol{a}^*\right)^T \boldsymbol{G}\boldsymbol{a}^*,\end{aligned}$$

由于最佳系数满足法方程

$$\boldsymbol{G}\boldsymbol{a}^* = \boldsymbol{d}$$

因而

$$\|f - P_n^*\|_2^2 = \langle f, f \rangle - \boldsymbol{d}^T \boldsymbol{a}^* = \langle f, f \rangle - \sum_{k=0}^{n} d_k a_k^*$$

目录

最小二乘

假设我们有数据点 (x_i, y_i)，$i = 0, 1, 2, \cdots, m$，我们希望找到一个函数 S 使得误差 $\delta_i = S(x_i) - y_i$ 最小化

$\|\boldsymbol{\delta}\|_\infty$ 的问题：计算困难

$$\|\boldsymbol{\delta}\|_\infty = \max_{0 \le i \le m} |\delta_i| \to \text{导数不好求}$$

$$\|\boldsymbol{\delta}\|_2 = \sqrt{\delta_0^2 + \delta_1^2 \cdots + \delta_m^2} \to \text{导数好求}$$

我们假设 $S(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x)\quad(n<m)$

$$\min_S\|\boldsymbol{\delta}\|_2^2=\min_S\sum_{i=0}^m\left[S\left(x_i\right)-y_i\right]^2=\min_S\sum_{i=0}^m\left[\sum_{j=0}^n a_j\varphi_j\left(x_i\right)-y_i\right]^2$$

最优的函数（即最小化误差）标记为 S^*

更一般地，

$$\|\boldsymbol{\delta}\|_2^2=\sum_{i=0}^m\underbrace{\omega\left(x_i\right)}_{\text{权重}}\left(\sum_{j=0}^n a_j\varphi_j(x_i)-y_i\right)^2=L\left(a_0,a_1,\ldots,a_n\right)$$

ω 的选择取决于具体的问题；比如可以根据数据的可信度、重复实验的次数等因素来附上权重 ω

$$\frac{\partial L}{\partial a_k} = 2\sum_{i=0}^{m} \omega\left(x_i\right)\left(\sum_{j=0}^{n} a_j \varphi_j\left(x_i\right) - y_i\right)\varphi_k\left(x_i\right) = 0$$

我们引入一些符号来简化一下：

$$\langle \varphi_j, \varphi_k \rangle \overset{\text{定义为}}{=} \sum_{i=0}^{m} \omega(x_i)\varphi_j(x_i)\varphi_k(x_i)$$

$$\langle f, \varphi_k \rangle \overset{\text{定义为}}{=} \sum_{i=0}^{m} \omega(x_i) y_i \varphi_k(x_i) \overset{\text{标记为}}{=} d_k$$

上述的方程可以写成矩阵的形式 $\boldsymbol{G}\boldsymbol{a}=\boldsymbol{d}$

$$\boldsymbol{a}=\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} \quad \boldsymbol{d}=\begin{pmatrix} d_0 \\ d_1 \\ \vdots \\ d_n \end{pmatrix} \quad \boldsymbol{G}=\begin{pmatrix} \langle\varphi_0,\varphi_0\rangle & \langle\varphi_0,\varphi_1\rangle & \cdots & \langle\varphi_0,\varphi_n\rangle \\ \langle\varphi_1,\varphi_0\rangle & \langle\varphi_1,\varphi_1\rangle & \cdots & \langle\varphi_1,\varphi_n\rangle \\ \vdots & \vdots & & \vdots \\ \langle\varphi_n,\varphi_0\rangle & \langle\varphi_n,\varphi_1\rangle & \cdots & \langle\varphi_n,\varphi_n\rangle \end{pmatrix}$$

由于基函数 $\varphi_0,\varphi_1,\cdots,\varphi_n$ 线性无关，因而 $|\boldsymbol{G}|\neq 0$，故而方程组有唯一解

备注（不要求）：若函数的最佳平方逼近中 $\rho=\sum_{i=0}^{n}\omega(x_i)\delta(x-x_i)$，则变成了这里的最小二乘逼近；课本 (3.4) 节和 (3.6) 节讲的是一回事。

非多项式的最小二乘拟合

若我们希望拟合 $y(x) = ae^{b/x}$， a, b 为参数

$$\begin{aligned}\ln y &= \ln a + \frac{b}{x} \\ \text{令 } t &= \frac{1}{x} \\ \ln y &= \ln a + bt\end{aligned}$$

可以先将数据变成 $\left(\frac{1}{x_i}, \ln y_i\right)$ 再做线性拟合 (详见例题 3.6)。

目录

多元函数的最小二乘拟合

给定 $m+1$ 个 d 维的数据 $\boldsymbol{x}_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,d} \end{bmatrix}$ 和函数值 y_i $(0 \le i \le m)$

我们可以类似推广：对于一个多元函数 $f(x_1, x_2, \cdots, x_d)$，希望找到最佳的

$$S_n(x_1, x_2, \cdots, x_d) = \sum_{k=0}^{n} a_k \varphi_k(x_1, x_2, \cdots, x_d), \qquad n \le m$$

使得如下的误差最小

$$\min_{a_0, a_1, \cdots, a_n} \sum_{i=0}^{m} \omega_i \Big(y_i - S_n(x_{1,i}, x_{2,i}, \cdots, x_{d,i}) \Big)^2$$

此处的最小二乘拟合的参数化方式对于高维数据并不高效。对于高维问题，所需要的基函数的个数可能关于 d 呈现指数递增关系（维数灾难）

给定 $m+1$ 个 d 维的数据 $\boldsymbol{x}_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,d} \end{bmatrix}$ 和函数值 y_i $(0 \le i \le m)$

例子：d 是维数，$\boldsymbol{x}_i$ 是一张图片，y_i 是某个属性（比如图片的类别）；则我们希望找到一个带参数的函数族 S_θ 使得如下的误差最小化：

$$\min_\theta \sum_{i=0}^{m} \omega_i \Big(y_i - S_\theta(x_{1,i}, x_{2,i}, \cdots, x_{d,i}) \Big)^2$$

常见的例子是 $\omega_i = \frac{1}{m+1}$。这其实就是一个十分典型的机器学习问题；只是 S_θ 一般是用非线性的神经网络来构造。

前面提到的最小二乘可以被理解为一个参数线性化的网络。

目录

总结

- 我们希望找到 n 次多项式 P_n^* 来近似 f：

$$\|f - P_n^*\|_2 = \min_P \|f - P\|_2$$

- 该问题变成求解法方程（一个线性方程组）：

$$\boldsymbol{Ga} = \boldsymbol{d}$$

- 基于数据的最小二乘拟合具有类似的形式

解线性方程组如果想要解比较精确，我们会希望 $\mathrm{cond}(\boldsymbol{G})$ 不是很大。而 $\varphi_k(x) = x^k$ 对应的 Hilbert 矩阵 $\boldsymbol{G}$ 的条件数随着 n 快速变大，那么我们该如何解决？