实验指导书： 插值与拟合模型

【实验目的】

教学目的：[1] 了解插值、最小二乘拟合的基本原理

[2] 掌握用 MATLAB 计算一维插值和两种二维插值的方法；

[3] 掌握用 MATLAB 作最小二乘多项式拟合和曲线拟合的方法

【实验相关知识】

插值与拟合

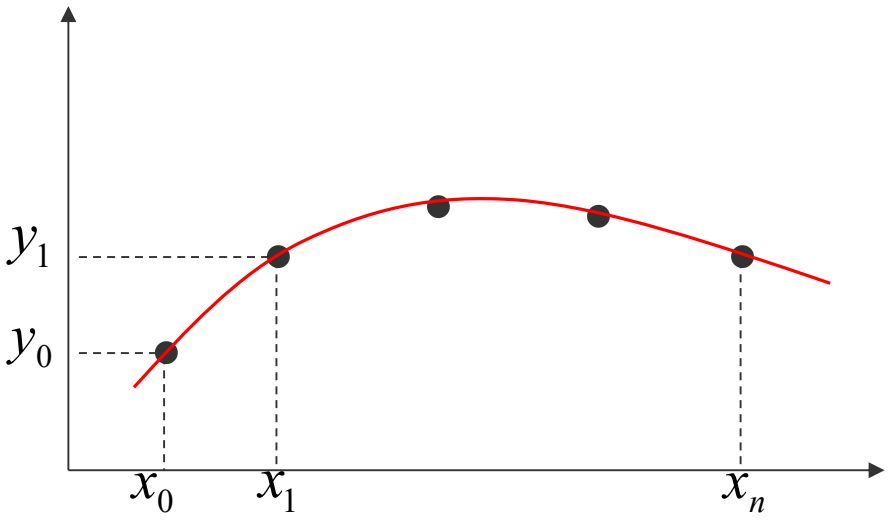
§1 多项式插值问题

已知函数 $y=f(x)$ 在 n+1 个互异结点处的函数值，如下表所示：

x_i	x_0	x_1	……	x_n
$y_i=f(x_i)$	y_0	y_1	……	y_n

求一个 n 次多项式 $P_n(x)$，使得 $P_n(x_i)=y_i$，$i=1,2,\cdots\cdots,n$。并利用 $P_n(x)$ 近似未知函数 $f(x)$。

从几何上看就是寻找一条 n 次多项式曲线 $y=P_n(x)$，使其通过平面上已知的 n+1 个点：



一、Lagrange 插值

$$P_n(x)=y_0l_0(x)+y_1l_1(x)+\cdots+y_nl_n(x)$$

其中，

$$l_i(x)=\frac{\prod\limits_{\substack{j=0\\ j\neq i}}^{n}(x-x_j)}{\prod\limits_{\substack{j=0\\ j\neq i}}^{n}(x_i-x_j)},\ i=1,2,\cdots,n$$

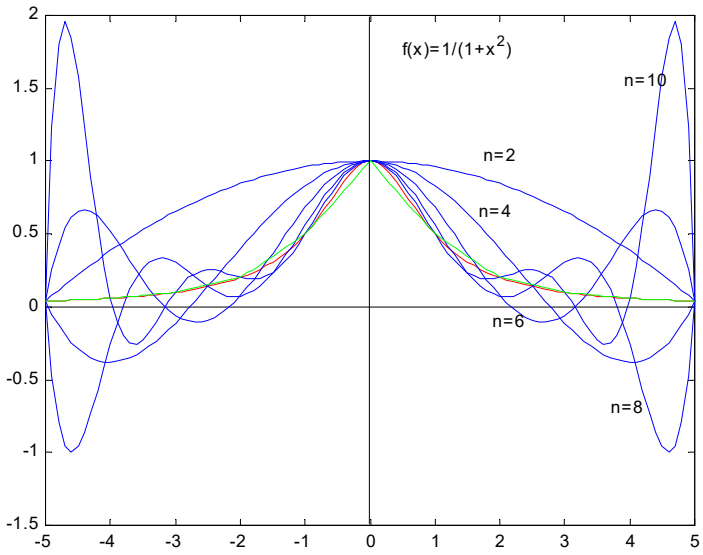
二、Newton 插值

$$P_n(x)=y_0+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+$$
$$\cdots+f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1})$$

其中，

$$f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0}$$

$$f[x_0,x_1,\cdots,x_k]=\frac{f[x_1,x_2,\cdots,x_k]-f[x_0,x_1,\cdots,x_{k-1}]}{x_k-x_0},k=1,2,\cdots,n$$

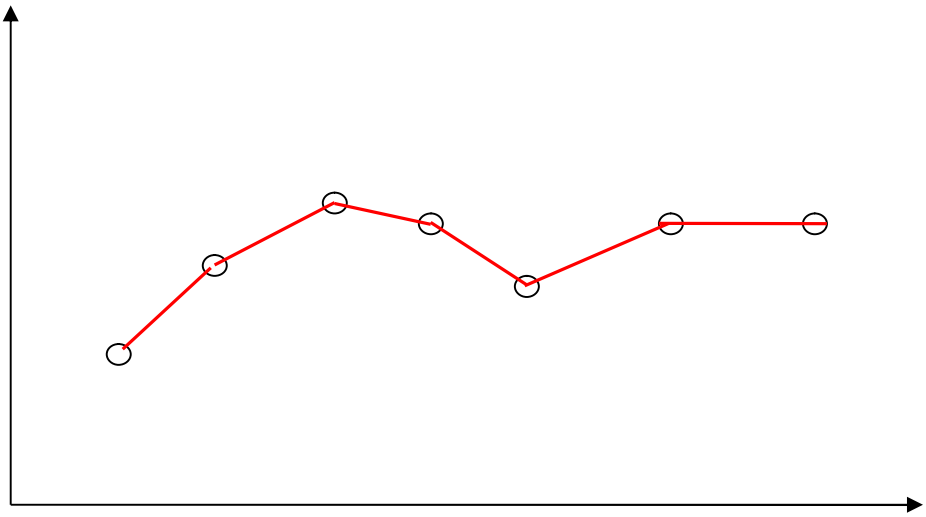


随着插值结点的增多，插值多项式的次数也增加。然而多项式次数越高，近似效果未必越好，反而容易出现高次插值的 Runge 现象，为此需要考虑下面的分段插值问题。

三、分段插值

1、分段线性插值

在相邻两个结点$[x_k, x_{k+1}]$内，求一条线段近似函数 $f(x)$，$x\in[x_k, x_{k+1}]$。



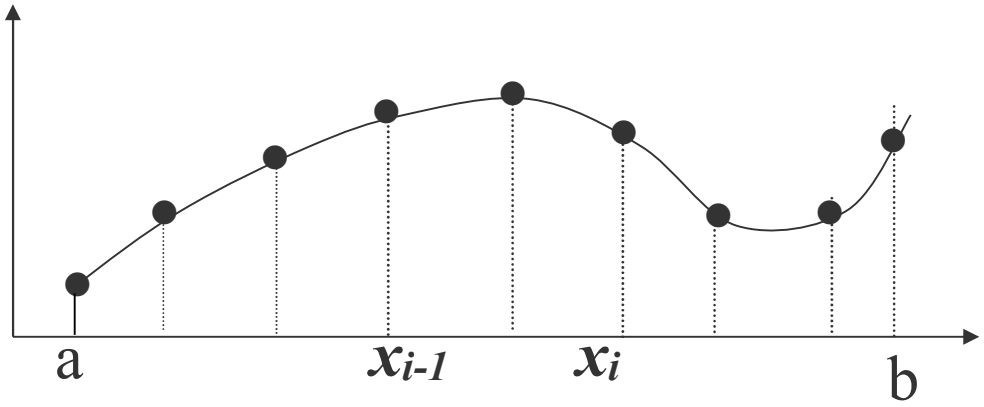
2、分段抛物插值

在相邻三个结点之间用抛物线近似未知函数。

四、样条插值

分段线性插值虽然避免了高次多项式插值的 Runge 现象，然而在插值结点处又产生了新问题：不光滑。为了克服这一现象，引入三次样条插值：在相邻两个结点之间用三次多项式函数 $s_i(x)$近似未知未知函数，并保证在插值结点处满足衔接条件：

$$s_i(x_i)=s_{i+1}(x_i),\ s_i'(x_i)=s_{i+1}'(x_i),\ s_i''(x_i)=s_{i+1}''(x_i)\ (i=1,\cdots,n-1)$$



五、Matlab 插值命令

yi=interp1(x，y，xi，'method')

(x，y)：插值节点；xi：被插值点；yi：xi 处的插值结果；

method：插值方法；‘nearest’ 最邻近插值；‘linear’ 线性插值；

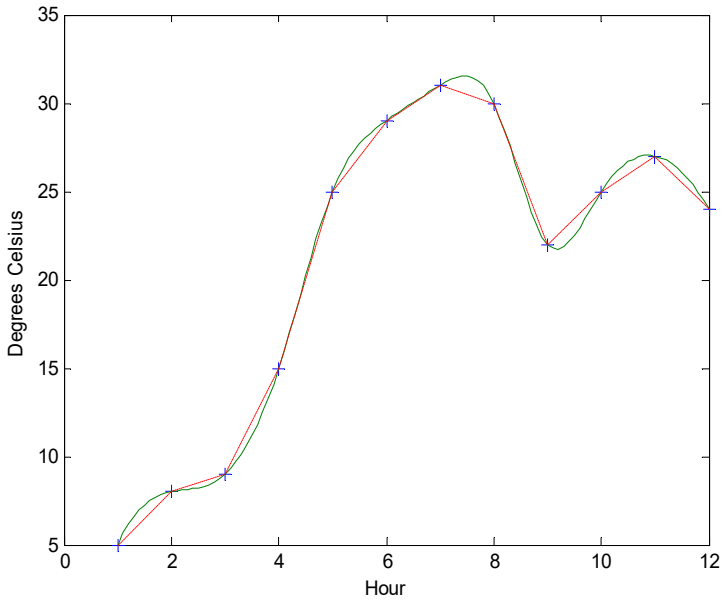
‘spline’ 三次样条插值；‘cubic’ 立方插值。缺省时：分段线性插值。

注意：所有的插值方法都要求 x 是单调的，并且 xi 不能够超过 x 的范围。

例 1：在 1-12 的 11 小时内，每隔 1 小时测量一次温度，测得的温度依次为：5，8，9，15，25，29，31，30，22，25，27，24。试估计每隔 1/10 小时的温度值。

Matlab 命令

```
hours=1:12;
temps=[5 8 9 15 25 29 31 30 22 25 27 24];
h=1:0.1:12;
t=interp1(hours,temps,h,'spline');
plot(hours,temps,'+',h,t,hours,temps,'r:')      %作图
xlabel('Hour'),ylabel('Degrees Celsius')
```



练习：下列数据表示从 1790 年到 2000 年的美国人口数据，利用这些数据给出 1790 年—2000 年每隔 5 年的美国人口数据，并预测 2005 年、2010 年人口数，与实际值比较。（人口单位：千人）

年份	1790	1800	1810	1820	1830	1840
人口	3929	5308	7240	9638	12866	17069
年份	1850	1860	1870	1880	1890	1900
人口	23192	31443	38558	50156	62948	75995
年份	1910	1920	1930	1940	1950	1960
人口	91972	105711	122755	131669	150697	179323
年份	1970	1980	1990	2000	2005	2010
人口	203212	226505	248710	281416	?	?

§2 二维插值问题

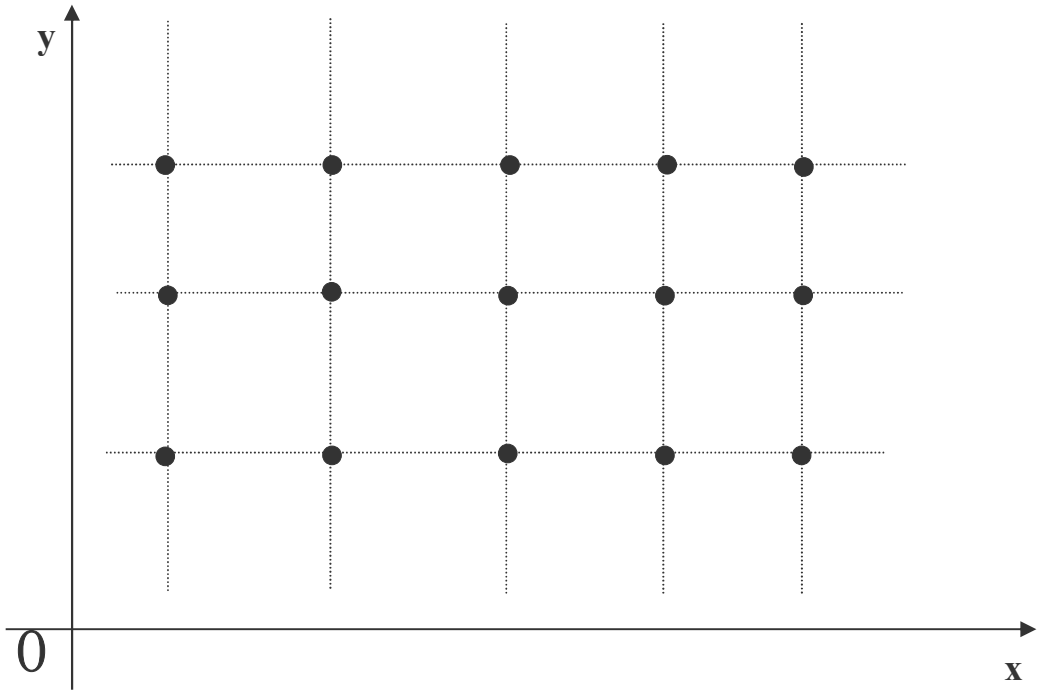
已知函数 $z = f(x, y)$ 在若干结点处的函数值，

$$z_{ij} = f(x_i, y_j), i = 1,2,\cdots,m; j = 1,2,\cdots,n,$$

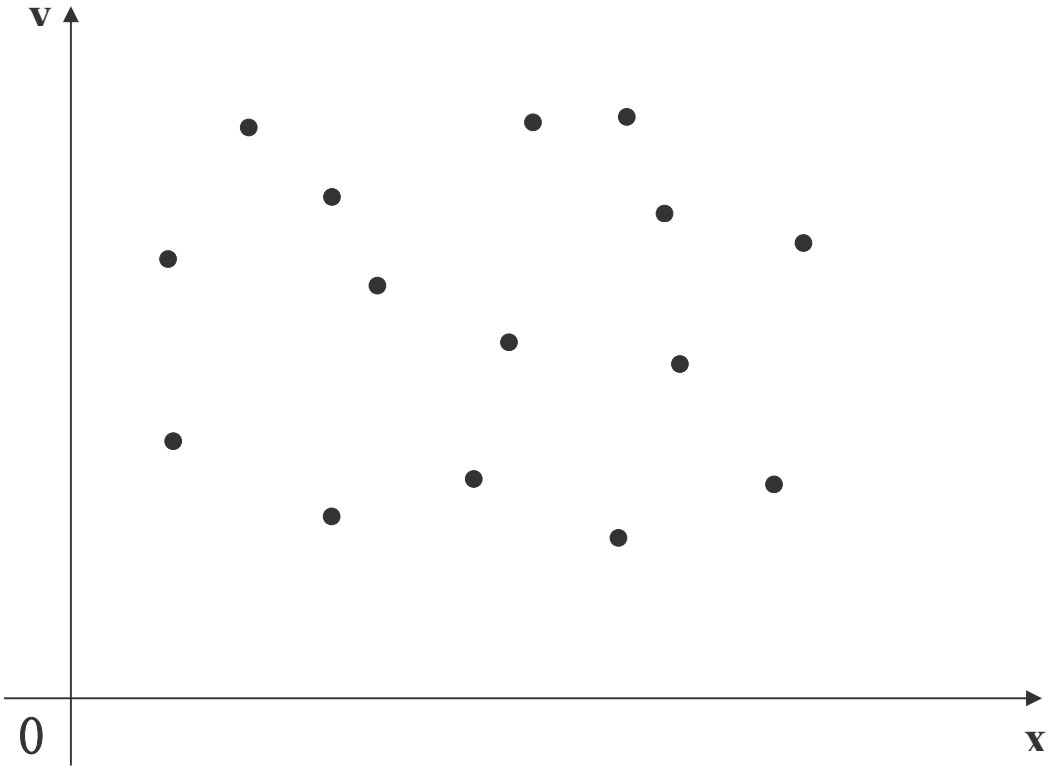
找一个简单二元函数使其通过已知的 m×n 个结点，并用此函数近似 $f(x, y)$。

一、结点类型

1、网格型结点

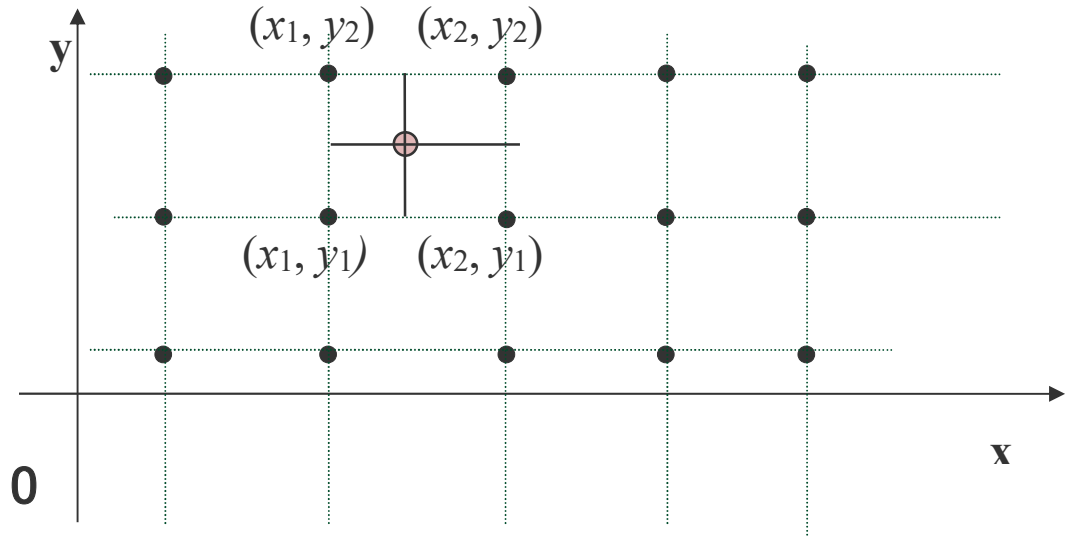


2、散乱结点



二、插值方法

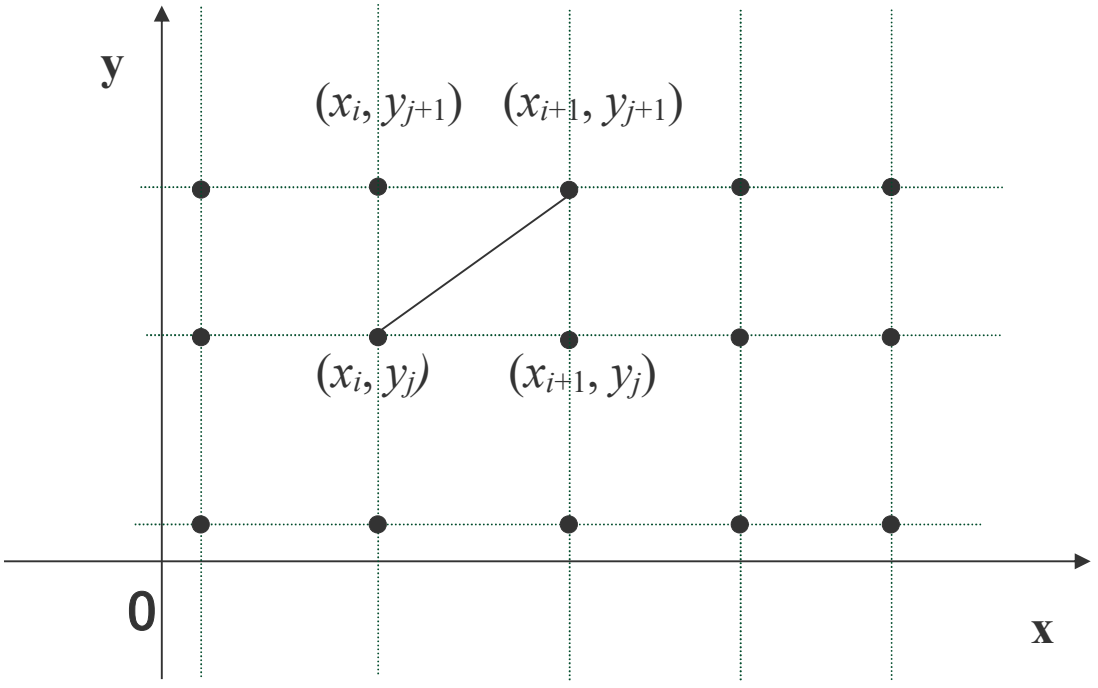
1、最邻近插值



二维或高维情形的最邻近插值，与被插值点最邻近的节点的函数值即为所求。

注意：最邻近插值一般不连续。具有连续性的最简单的插值是分片线性插值。

2、分片线性插值



将四个插值点（矩形的四个顶点）处的函数值依次简记为：

$f\ (x_i,\ y_j)=f_1$，$f\ (x_{i+1},\ y_j)=f_2$，$f\ (x_{i+1},\ y_{j+1})=f_3$，$f\ (x_i,\ y_{j+1})=f_4$

分两片的函数表达式如下：

第一片(下三角形区域)：$(x,\ y)$满足

$$y \le \frac{y_{j+1}-y_j}{x_{i+1}-x_i}(x-x_i)+y_j$$

插值函数为：

$$f(x,y)=f_1+(f_2-f_1)(x-x_i)+(f_3-f_2)(y-y_j)$$

第二片(上三角形区域)：(x, y)满足

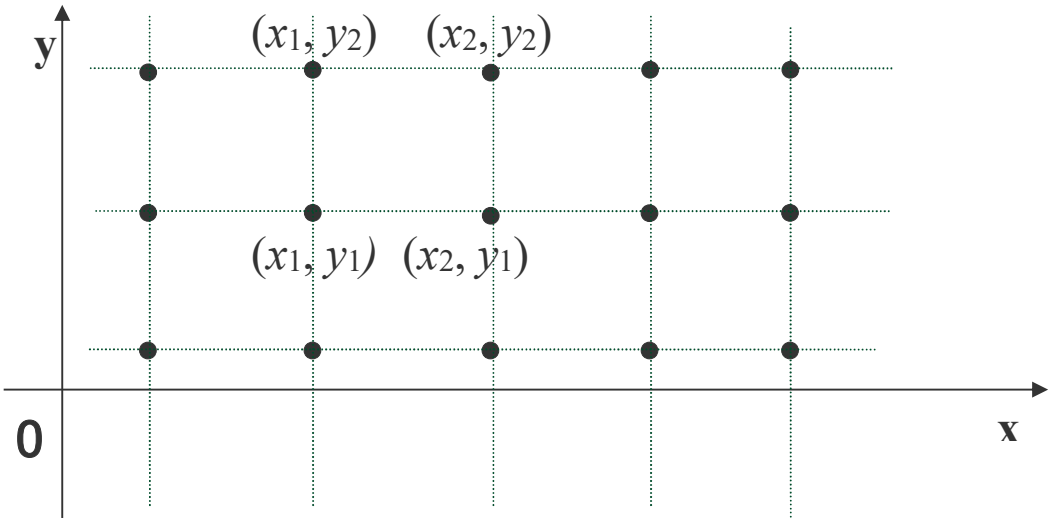
$$y \ge \frac{y_{j+1}-y_j}{x_{i+1}-x_i}(x-x_i)+y_j$$

插值函数为：

$$f(x,y)=f_1+(f_4-f_1)(y-y_j)+(f_3-f_4)(x-x_i)$$

注意：(x, y)当然应该是在插值节点所形成的矩形区域内。显然，分片线性插值函数是连续的；

3、双线性插值



双线性插值是一片一片的空间二次曲面构成。

双线性插值函数的形式如下：

$$f(x,y)=(ax+b)(cy+d)$$

其中有四个待定系数，利用该函数在矩形的四个顶点（插值节点）的函数值，得到四个代数方程，正好确定四个系数。

三、用 MATLAB 作网格节点数据的插值

z=interp2(x0,y0,z0,x,y,’ method’)

method：插值方法

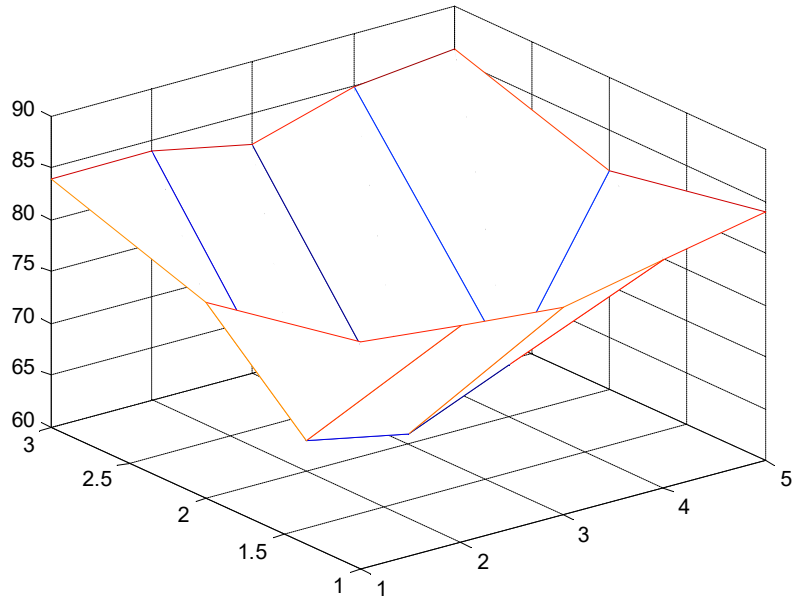
‘nearest’表示最邻近插值；‘linear’表示双线性插值；

‘cubic’表示双三次插值；　缺省时表示双线性插值；
（x0,y0,z0）表示插值节点；（x，y）表示被插值点；
z 表示被插值点的函数值。
要求 x0,y0 单调；x，y 可取为矩阵，或 x 取行向量，y 取为列向量，x,y 的值分别不能超出 x0,y0 的范围。
例 2：测得平板表面 3*5 网格点处的温度分别为：

82	81	80	82	84
79	63	61	65	81
84	84	82	85	86

试作出平板表面的温度分布曲面 z=f(x,y)的图形。
1.先在三维坐标画出原始数据，画出粗糙的温度分布曲图.
输入以下命令：

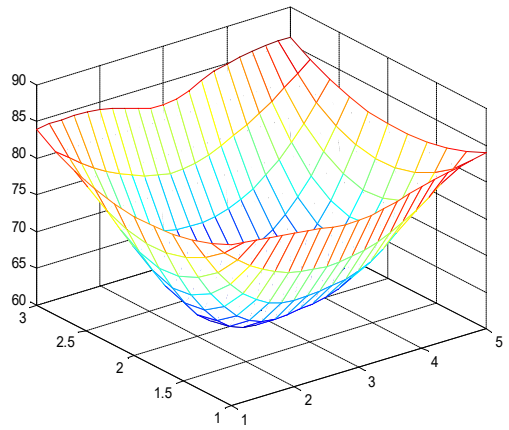
```
x=1:5;
y=1:3;
temps=[82 81 80 82 84;79 63 61 65 81;84 84 82 85 86];
mesh(x,y,temps)
```



2. 以平滑数据,在 x、y 方向上每隔 0.2 个单位的地方进行插值.
再输入以下命令:

```
xi=1:0.2:5;yi=1:0.2:3;zi=interp2(x,y,temps,xi',yi,'cubic');
mesh(xi,yi,zi)
```

画出插值后的温度分布曲面图.



练习 山区地貌：

在某山区测得一些地点的高程如下表。平面区域为

1200≤x≤4000,1200≤y≤3600)

试作出该山区的地貌图和等高线图，并对几种插值方法进行比较。

X / Y	1200	1600	2000	2400	2800	3200	3600	4000
1200	1130	1250	1280	1230	1040	900	500	700
1600	1320	1450	1420	1400	1300	700	900	850
2000	1390	1500	1500	1400	900	1100	1060	950
2400	1500	1200	1100	1350	1450	1200	1150	1010
2800	1500	1200	1100	1550	1600	1550	1380	1070
3200	1500	1550	1600	1550	1600	1600	1600	1550
3600	1480	1500	1550	1510	1430	1300	1200	980

通过此例对最邻近点插值、双线性插值方法和双三次插值方法的插值效果进行比较。

四、用 MATLAB 作散点数据的插值计算

插值函数 griddata 格式为:

cz =griddata（x，y，z，cx，cy，‘method’）

（cx，cy）表示被插值结点;（x，y，z）表示插值节点;cz 表示被插值结点的函数值;method:插值方法

‘nearest’最邻近插值；‘linear’双线性插值；‘cubic’双三次插值；

’v4’- Matlab 提供的插值方法；缺省时,双线性插值

例 3 在某海域测得一些点(x,y)处的水深 z 由下表给出，船的吃水深度为 5 英尺，在矩形区域（75，200）*（-50，150）里的哪些地方船要避免进入。

x	129 140 103.5 88 185.5 195 105
y	7.5 141.5 23 147 22.5 137.5 85.5
z	4 8 6 8 6 8 8
x	157.5 107.5 77 81 162 162 117.5
y	-6.5 -81 3 56.5 -66.5 84 -33.5
z	9 9 8 8 9 4 9

1、输入插值结点数据;

2、在矩形区域（75，200）*（-50，150）作二维插值，采用三次插值法；
3、作海底曲面图；
4.作出水深小于 5 的海域范围,即 z=5 的等高线。
命令如下

```
clear
x=[129 140 103.5 88 185.5 195 105.5 157.5 107.5 77 81 162 162 117.5];
y=[7.5 141.5 23 147 22.5 137.5 85.5 -6.5 -81 3 56.5 -66.5 84 -33.5];
z=[-4 -8 -6 -8 -6 -8 -8 -9 -9 -8 -8 -9 -4 -9];
cx=75:0.5:200;
cy=-70:0.5:150;
cz=griddata(x,y,z,cx,cy','cubic');
meshz(cx,cy,cz),rotate3d
xlabel('X'),ylabel('Y'),zlabel('Z')
figure(2),contour(cx,cy,cz,[-5 -5]);grid
hold on
plot(x,y,'+')
xlabel('X'),ylabel('Y')
```

练习 山区地貌：
在某山区测得一些地点的高程如下表：(平面区域1200<=x<=4000,1200<=y<=3600)，试作出该山区的地貌图和等高线图，并对几种插值方法进行比较。

3600	1480 1500 1550 1510 1430 1300 1200 980
3200	1500 1550 1600 1550 1600 1600 1600 1550
2800	1500 1200 1100 1550 1600 1550 1380 1070
2400	1500 1200 1100 1350 1450 1200 1150 1010
2000	1390 1500 1500 1400 900 1100 1060 950
1600	1320 1450 1420 1400 1300 700 900 850
1200	1130 1250 1280 1230 1040 900 500 700
Y/x	1200 1600 2000 2400 2800 3200 3600 4000

§3 曲线拟合

在科学实验的统计方法研究中，往往要从一组实验数据(x_i,y_i)中寻找自变量 x 与因变量 y 之间的函数关系$y=f(x)$。

在第一节中我们已经给出了一种方法：多项式插值。这种方法要求用来近似未知函数 f 的 n 次插值多项式准确无误地经过已知的 n+1 个结点。然而当结点数据是由某种实验或者计算方法得出的，就难免带有误差，要求多项式严格经过这些结点，无形中就会将误差保留下来，而且如果每一个结点都有误差的话，由于误差累积的效应，也会导致整体的近似效果较差。

这促使我们寻求一种新的方法近似未知函数，这一方法并不要求用来近似的函数$y=s(x)$严格通过已知结点，而只要求在结点x_j处误差$\delta_j=y_j-s(x_j)$，$j=0,1,\cdots,m$按某一标准最小。为了计算方便，通常就采用误差的平方和最小作为度量误差的标准。

一、基于最小二乘原理的曲线拟合问题

已知函数已知函数 $y=f(x)$ 在 m+1 个互异结点处的函数值，如下表所示：

x_i	x_0	x_1	……	x_m
$y_i=f(x_i)$	y_0	y_1	……	y_m

在给定的函数类

$$\varphi=\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$$

中求一个函数

$$S(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x),$$

使得

$$Q=\sum_{i=0}^{m}(y_i-S(x_i))^2$$

取得最小值。并利用 S(*x*)近似未知函数 *f*(*x*)。称 S(*x*)为最小二乘拟合函数。

特别地当 S 是一个 n 次多项式时，称为最小二乘多项式拟合。

注意：

（1）插值和拟合都是要根据一组数据构造一个函数作为近似，由于近似的要求不同，二者的数学方法上是完全不同的。而面对一个实际问题，究竟应该用插值还是拟合，有时容易确定，有时则并不明显。

（2）用最小二乘原理拟合数据时，首先要确定拟合函数S(*x*)的形式，这往往是最重要也是最困难的一步，它不是一个单纯的数学问题，还与所研究的变化规律及所得观测数据有关；通常应从问题的变化规律、给定数据的散点图以及实际问题的背景综合分析加以确定。

（3）当发现有多种类型曲线均符合样本点变化特征时，可以分别采用不同类型曲线进行拟合，最后通过某一评价指标确定最优的拟合结果。

（4）加权的最小二乘曲线拟合方法

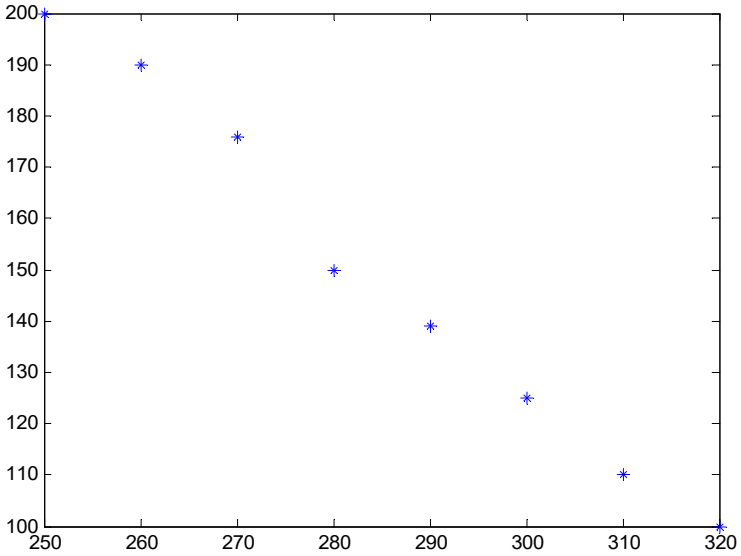
$$Q=\sum_{i=0}^{m}(y_i-S(x_i))^2\Rightarrow\sum_{i=0}^{m}\omega(x_i)(y_i-S(x_i))^2$$

其中 $\omega(x)\geq 0$ 是权重函数，它表示不同点处的数据比重（对最优拟合曲线的影响程度）不同。例如，权重可以用在点 (x_i,y_i) 处的重复观测次数表示。

例 4： 根据统计资料，水泥预期销售量与价格的关系如下表：

单价	250	260	270	280	290	300	310	320
售量（$\times 10^4$）	200	190	176	150	139	125	110	100

确定水泥预期销售量与价格的近似函数关系。



例5：在某化学反应里，根据实验所得生成物的浓度与时间的关系数据见下表，求浓度y与时间t的拟合曲线 $y = F(t)$

t（时间）	1	2	3	4	5	6	7	8
y（浓度）	4.00	6.40	8.00	8.80	9.22	9.50	9.70	9.86
t（时间）	9	10	11	12	13	14	15	16
y（浓度）	10.00	10.20	10.32	10.42	10.50	10.55	10.58	10.60

例6：CUMCM 92A　　例7：CUMCM 2004C

二、Matlab作曲线拟合

1、多项式拟合

p=polyfit(x,y,n)—其中x,y为给出的数据，n为多项式的次数。 输出变量p是一个向量，其元素给出了拟合多项式的系数（按从高到低排列）

多项式在x处的值y可用以下命令计算： y=polyval（a,x）

2. 用MATLAB作非线性最小二乘拟合

Matlab的提供了两个求非线性最小二乘拟合的函数：curvefit和leastsq。两个命令都要先建立M-文件fun.m，在其中定义拟合函数S(x)。

（1）c = curvefit（‘fun’,[a,b],x,y);

（2）c=leastsq(‘fun’,[a,b]);

例题：水泥价格、广告费与利润问题

推销商品的重要手段之一是做广告，而做广告要出钱，利弊得失如何估计，需要利用有关数学模型作定量的讨论。

某建材公司有一大批水泥需要出售，根据以往统计资料，零售价增高，则销量减少，具体数据见表三；如果做广告，可使销售量增加，具体增加量以售量提高因子 k 表示，k 与广告费

的关系列于表四。现在，已知水泥的进价是每吨 250 元。问如何确定该批水泥的出售价格和花多少广告费，可使公司获利最大？

表 1：水泥预期销售量与售价的关系

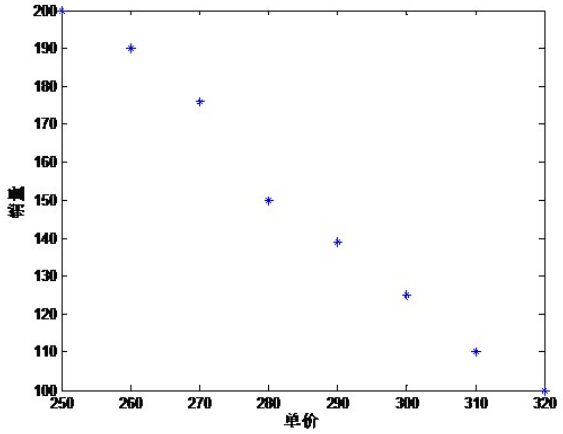
单价（元/吨）	250	260	270	280	290	300	310	320
售量（万吨）	200	190	176	150	139	125	110	100

表 2：售量提高因子 k 与广告费的关系

广告费（万元）	0	60	120	180	240	300	360	420
提高因子 k	1.00	1.40	1.70	1.85	1.95	2.00	1.95	1.80

将表 1（水泥预期销售量与价格的关系）中数据绘图如下

图 1



可以看出销量与单价近似呈现线性关系，因此可设

$$y = a + bx$$

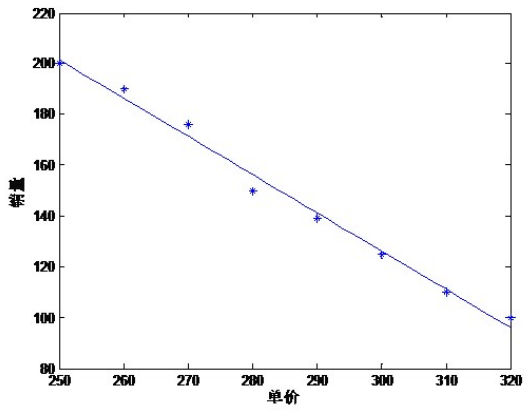
用最小二乘法，根据表 1 的数据可得正规方程组

$$\begin{bmatrix} 8 & 2280 \\ 2280 & 65400 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1190 \\ 332830 \end{bmatrix}$$

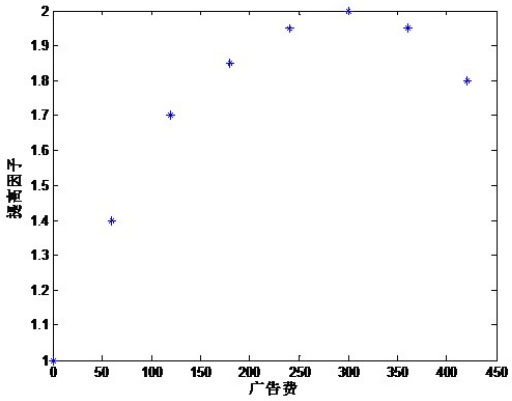
解此方程组，得到系数 $a = 577.6071, b = -1.5048$。即销量与单价的近似关系为：

$y = 577.6071 - 1.5048x$，将此结果与原数据点画在同一幅图形中比较如下

图 2：



将表 2（销售量提高因子与广告费的关系）中数据画图如下：
图 3：



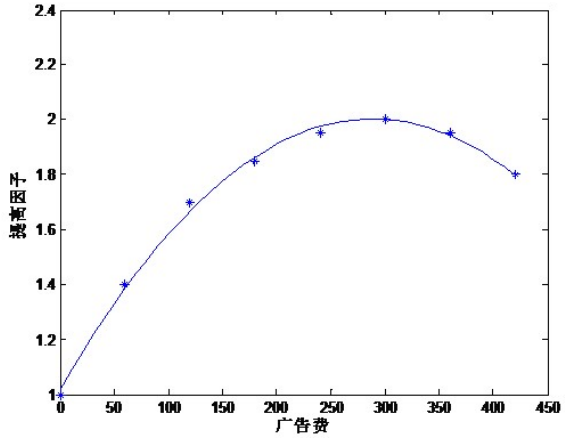
可以看出提高因子与广告费近似成二次关系，因此可设：

$$k = d + ez + fz^2$$

用最小二乘法，根据表 2 中数据，可得正规方程组

$$\begin{bmatrix} 8 & 1680 & 504000 \\ 1680 & 504000 & 169344000 \\ 504000 & 169344000 & 60600959990 \end{bmatrix} \begin{bmatrix} d \\ e \\ f \end{bmatrix} = \begin{bmatrix} 13.65 \\ 3147 \\ 952020 \end{bmatrix}$$

解此方程组得：$d = 1.02000, e = 6.807 \times 10^{-3}, f = -1.17973$。将拟合多项式图形与原数据画在同一幅图形中观察拟合效果：
图 4：



设实际销量为 S，则 $S = ky$。于是利润 L 可以表示为：

$$L = Sx - Sc - z = ky(x - c) - z$$
$$= (d + ez + fz^2)(a + bx)(x - c) - z$$

要求出利润最大值，只需令：

$$\frac{\partial L}{\partial x}=(d+ez+fz^2)(a-bc+2bx)=0$$
$$\frac{\partial L}{\partial z}=(e+2fz^2)(a+bx)(x-c)-1=0$$

解得：

$$\begin{cases} x=\dfrac{1}{2b}(bc-a)=316.93 \\ z=\dfrac{1}{2f}\left[\dfrac{1}{(a+bx)(x-c)}-e\right]=282.21 \end{cases}$$

进一步求出利润 L 的二阶偏导数得：

$$A=\frac{\partial^2 L}{\partial x^2}=2b(d+ez+fz^2)$$
$$B=\frac{\partial^2 L}{\partial x\partial z}=(e+2fz)(a-bc+2bx)$$
$$C=\frac{\partial^2 L}{\partial z^2}=2f(a+bx)(x-c)$$

当 $x=316.93, z=282.21$ 时，显然有 $A<0, B=0, C<0$，根据多元函数极值理论可知，在 $x=316.93, z=282.21$ 处，利润最大，最大利润为 $L\max=13209.6(\text{万元})$。

从而，可以预计，将单价定为 316.93 元/t，广告费花费 282.21 万元，实际销售量可望达到 2105800 吨，利润可望达到 13209.6（万元）。