大数据的清洗

《美国数学建模竞赛》

完整课程请长按下方二维码

缺失值处理：插值

插值的定义

- 我们经常会遇到大量的数据需要处理，而处理数据的关键就在于这些算法，例如数据拟合、参数估计、插值等数据处理算法。此类问题在MATLAB中有很多现成的函数可以调用，熟悉MATLAB，这些方法都能游刃有余的用好。
- 在实际中，常常要处理由实验或测量所得到的一些离散数据。插值与拟合方法就是要通过这些数据去确定某一类已知函数的参数或寻求某个近似函数，使所得到的近似函数与已知数据有较高的拟合精度。此类问题为插值问题。

插值的使用

- 当数据量不够，需要补充，且认定已有数据可信时，通常利用函数插值方法。
- 实际问题当中碰到的函数 $f(x)$ 是各种各样的，有的表达式很复杂，有的甚至给不出数学的式子，只提供了一些离散数据，如某些点上的函数值和导数值。
- MATLAB 实现：实现分段线性插值不需要编制函数程序，它自身提供了内部的功能函数
- interp1 (一维插值)　intep2 (二维)　interp3 (三维)　intern (n维)

一维插值函数：

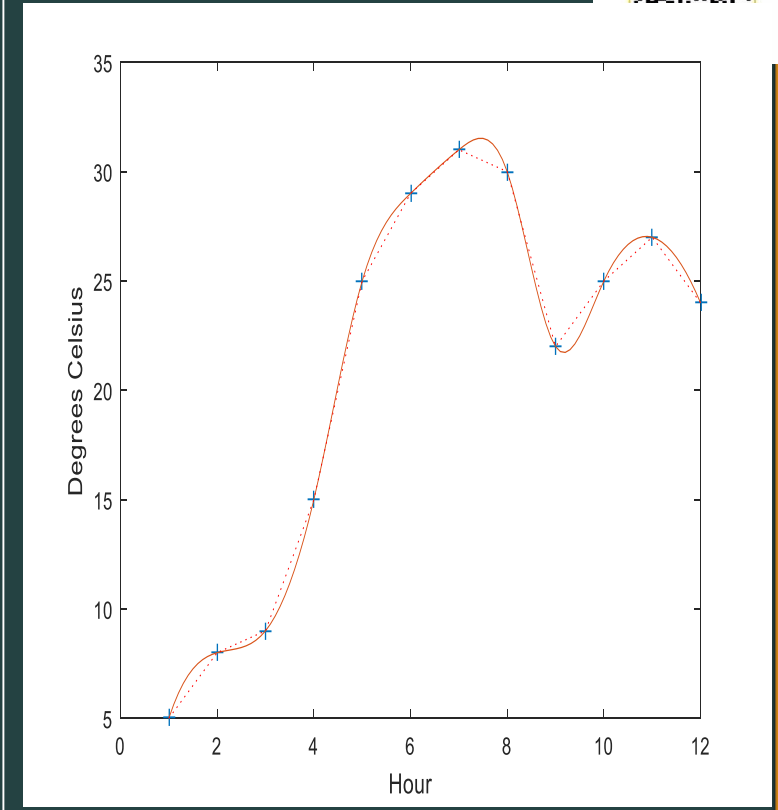
```
yi=interp1(x, y, xi, 'method')
```

- x_i 处的插值结果
- 插值节点
- 被插值点
- 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x_i不能够超过x的范围.

例：从1点12点的11小时内，每隔1小时测量一次温度，测得的温度的数值依次为：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.

- 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:')
- xlabel('Hour'),ylabel('Degrees Celsius')



用MATLAB作网格节点数据的插值

```
z=interp2(x0,y0,z0,x,y,'method')
```

- `x0,y0,z0`：被插值点的函数值
- `x0,y0`：插值节点
- `x,y`：被插值点
- `'method'`：插值方法

要求x_0,y_0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x_0,y_0的范围.

`'nearest'`	最邻近插值；
`'linear'`	双线性插值；
`'cubic'`	双三次插值；
缺省时	双线性插值.

- 例：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

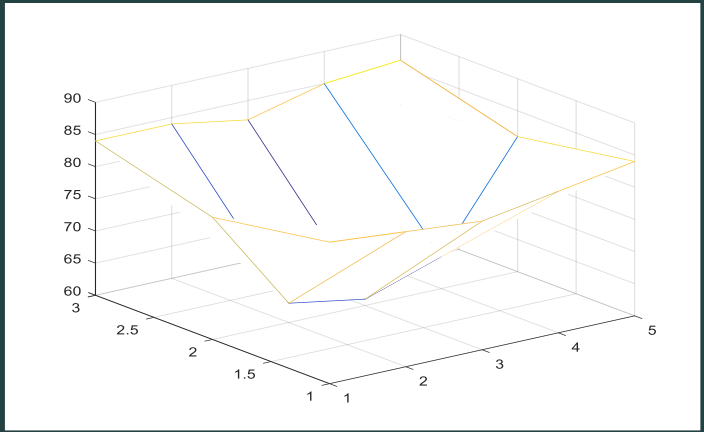
  试作出平板表面的温度分布曲面$z=f(x,y)$的图形.

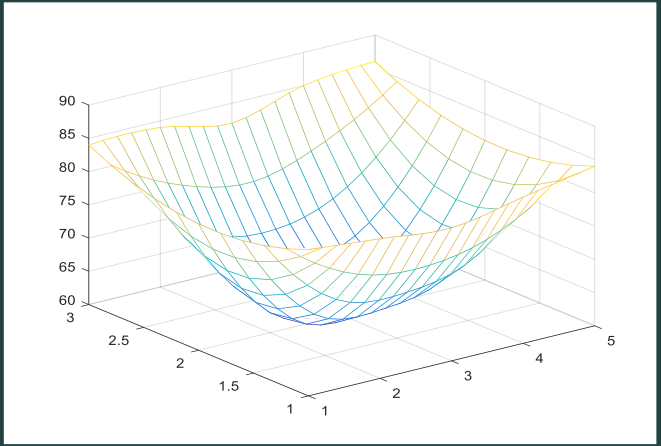
- 1. 先在三维坐标画出原始数据，画出粗糙的温度分布曲线图.
  - 输入以下命令：
  - x=1:5;
  - y=1:3;
  - temps=[82 81 80 82 84;79 63 61 65 81;84 84 82 85 86];
  - mesh(x,y,temps)

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)





插值函数griddata格式为:

cz =griddata（x，y，z，cx，cy，‘method’）

被插值点的函数值

插值节点

被插值点

插值方法

要求cx取行向量，cy取为列向量.

‘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 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.输入插值基点数据

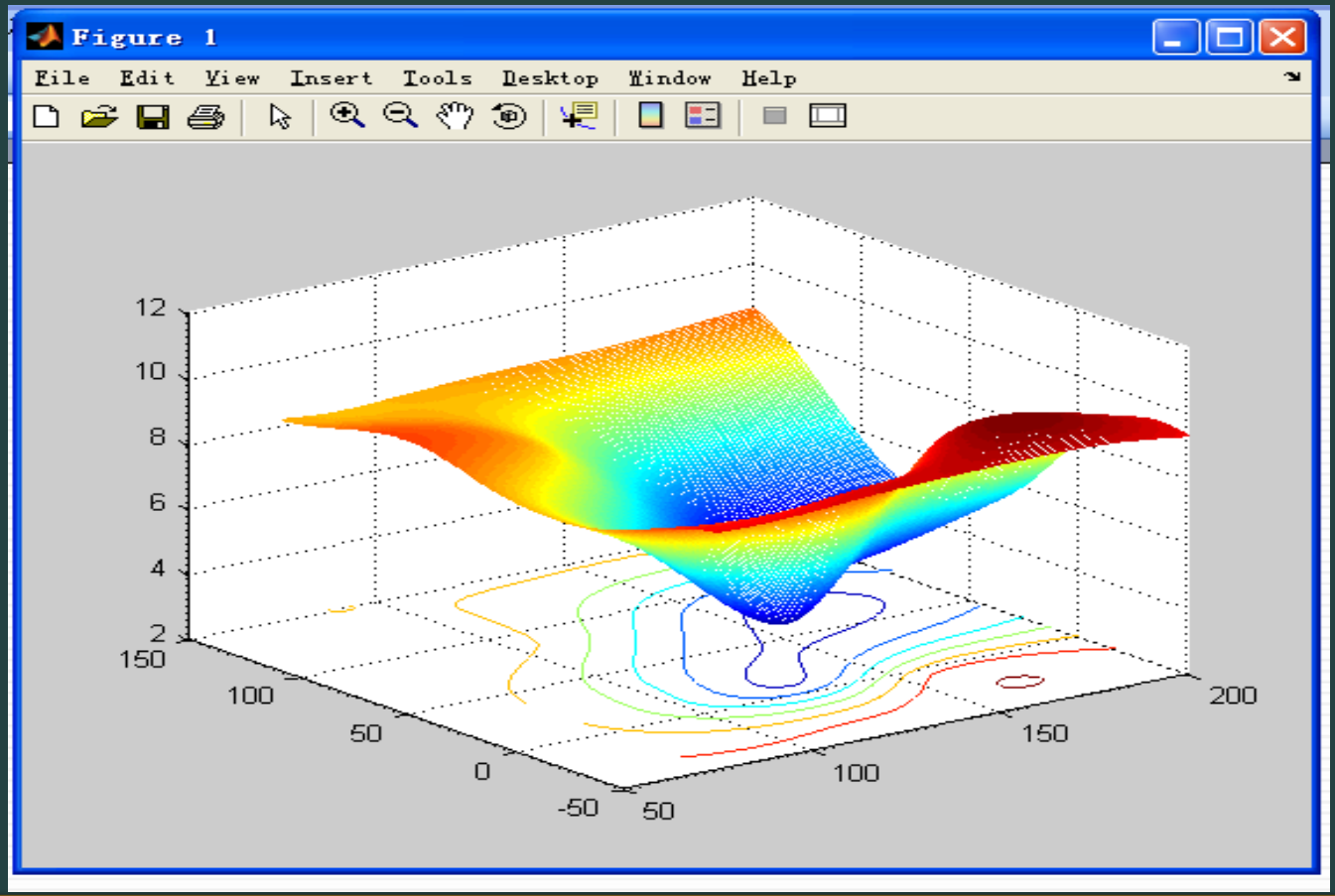
2.在矩形区域(75,200)×(-50,150)进行插值。

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.

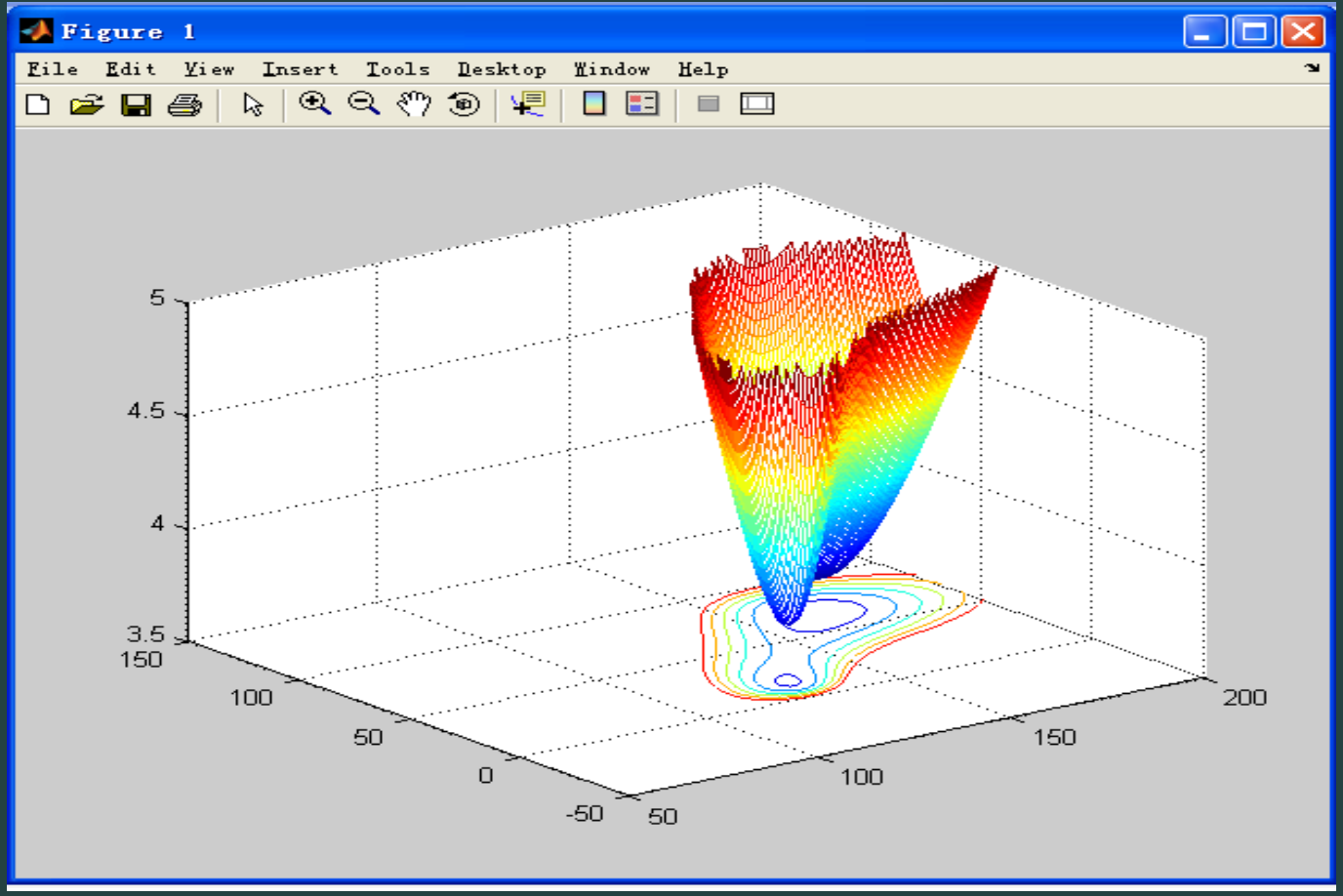
```
%程序1：插值并作海底曲面图
 x  =[129.0  140.0  103.5  88.0  185.5  195.0  105.5 157.5  107.5  77.0  81.0
162.0  162.0  117.5 ];
y =[ 7.5  141.5  23.0  147.0  22.5  137.5  85.5      -6.5  -81  3.0  56.5  -66.5
84.0  -33.5 ];
z =[ 4  8  6  8  6  8  8  9  9  8  8  9  4  9 ];
x1=75:1:200;
y1=-50:1:150;
[x1,y1]=meshgrid(x1,y1);
z1=griddata(x,y,z,x1,y1,'v4');
meshc(x1,y1,z1)
```

海底曲面图



```
%程序2：插值并作出水深小于5的海域范围。
x  =[129.0  140.0  103.5  88.0  185.5  195.0  105.5 157.5  107.5  77.0  81.0  162.0
162.0  117.5 ];
y =[ 7.5  141.5  23.0  147.0  22.5  137.5  85.5     -6.5  -81  3.0  56.5  -66.5  84.0  -
33.5 ];
z =[ 4  8  6  8  6  8  8  9  9  8  8  9  4  9 ];
x1=75:1:200;
y1=-50:1:150;
[x1,y1]=meshgrid(x1,y1);
z1=griddata(x,y,z,x1,y1,'v4');  %插值
z1(z1>=5)=nan;  %将水深大于5的置为nan，这样绘图就不会显示出来
meshc(x1,y1,z1)
```

水深小于5的海域范围



- 插值是数学建模前的准备工作，当实际问题提供的数据较少或比较混乱，而建模的目的又需要从已给的数据找出相应的时，就需要使用插值方法，特别是在地形地貌的绘制中，插值具有举足重轻的作用，另外在缺失值的补全中也可以使用插值方法。一般来讲，各种插值方法没有优劣之分，因此可以选用任意一种插值方法。

异常值处理

内容使用的背景

- 以往多从宏观角度进行住院费用研究。通过住院费用的异常数据挖掘算法的研究，使得从微观的角度，针对单个病例的住院费用的监管成为可能。一方面，可以有效的找到不合理的医疗费用支出，找出不规范的医疗行为，控制医疗费用不合理的上涨；另一方面，可以找出一些违规的操作和错误的录入信息，规范新农合的信息管理，加强新农合的监管力度，这样就提出了一个如何发掘“异常（outlier）”数据的问题。

内容使用的背景

- 现有数据挖掘研究大多集中于发现适用于大部分数据的常规模式，在许多应用领域中，异常数据通常作为噪音而忽略，许多数据挖掘算法试图降低或消除异常数据的影响。而在有些应用领域识别异常数据是许多工作的基础和前提，异常数据会带给我们新的视角。 如在欺诈检测中，异常数据可能意味欺诈行为的发生，在入侵检测中异常数据可能意味入侵行为的发生。

什么是异常数据？

- 异常是在数据集中偏离大部分数据的数据，使人怀疑这些数据的偏离并非由随机因素产生，而是产生于完全不同的机制。
- 异常挖掘（outlier mining）问题由两个子问题构成：
- (1)如何度量异常；
- (2)如何有效发现异常。
- 不同的异常点挖掘方法就是通过不同的异常度量方法，构造异常点得分(outlier score)，从而发现异常点。

方法理论简介

- **(1)基于统计的异常检测**
- 假定用一个参数模型来描述数据的分布（如正态分布），应用基于统计分布的异常点检测方法依赖于三个因素：数据分布，参数分布（如均值或方差），期望异常点的数目。
- 异常点是一个对象，在数据的概率分布模型中，它具有低概率，而数据的概率分布模型通过估计用户指定的分布及参数，由数据创建。如，假定数据具有高斯分布，则基本分布的均值和标准差可以通过计算数据的均值和标准差来估计，然后可以估计每个对象在该分布下的概率。

- 基于统计的异常点检测方法，就是在估计出数据的分布后，找出那些小概率出现的数据点，例如，如果数据来自标准正态分布N（0,1），则对象落在3标准差的中心区域以外的概率仅有0.0027，在统计意义下，则可以认为这样的数据点异常程度为1-0.0027。更一般地，如果x是属性值，则$|x|>c$的概率随c增加而迅速减小。

- 假设数据服从标准化正态分布，则异常点检测步骤如下：
- 1）给定阈值 α ，用来衡量异常程度
- 2）找出 $P(|x| \geq c) = \alpha$ ，对应的常数c
- 3）$|x| \geq c$ 的点即为异常点
- 这些内容可参考统计学中的参数估计的置信区间以及大数定律、中心极限定理等相关内容。

- 基于统计方法异常点检测技术的优缺点：
- 优点：
- 1）异常点检测的统计学方法具有坚实的基础，建立在标准的统计学技术(如分布参数的估计)之上。
- 2）当存在充分的数据和所用的检验类型的知识时，这些检验可能非常有效。
- 缺点：
- 1）大部分统计方法都是针对单个属性的，对于多元数据技术方法较少。
- 2）在许多情况下，数据分布是未知的。
- 3）对于高维数据，很难估计真实的分布。

- **(2)基于聚类的异常检测**
- 物以类聚——相似的对象聚合在一起，基于聚类的异常点检测方法有两个共同特点：
- （1）先采用特殊的聚类算法处理输入数据而得到聚类，再在聚类的基础上来检测异常。
- （2）只需要扫描数据集若干次，效率较高，适用于大规模数据集。

- 基于聚类的异常点检测方法计算如下：
- （1）把所有样本按某个聚类方法进行聚类，假设聚为k类：C_1 ，C_2 …C_K
- （2）对于每个对象p，计算该对象到每个类之间的距离$d(p, C_i)$
- （3）计算每个对象p的异常因子得分$OF(p)=\sum_{j=1}^{k}\frac{|C_j|}{|D|}\bullet d(p,C_j)$
- （4）计算所有对象的因子异常得分的平均值Ave_OF及标准差Dev_OF。
- （5）奇异值标定：若$OF(p)\geq Ave_OF+\beta\cdot Dve_OF(1\leq\beta\leq 2)$，则为奇异值。
- 通常取β=1或1.285 。

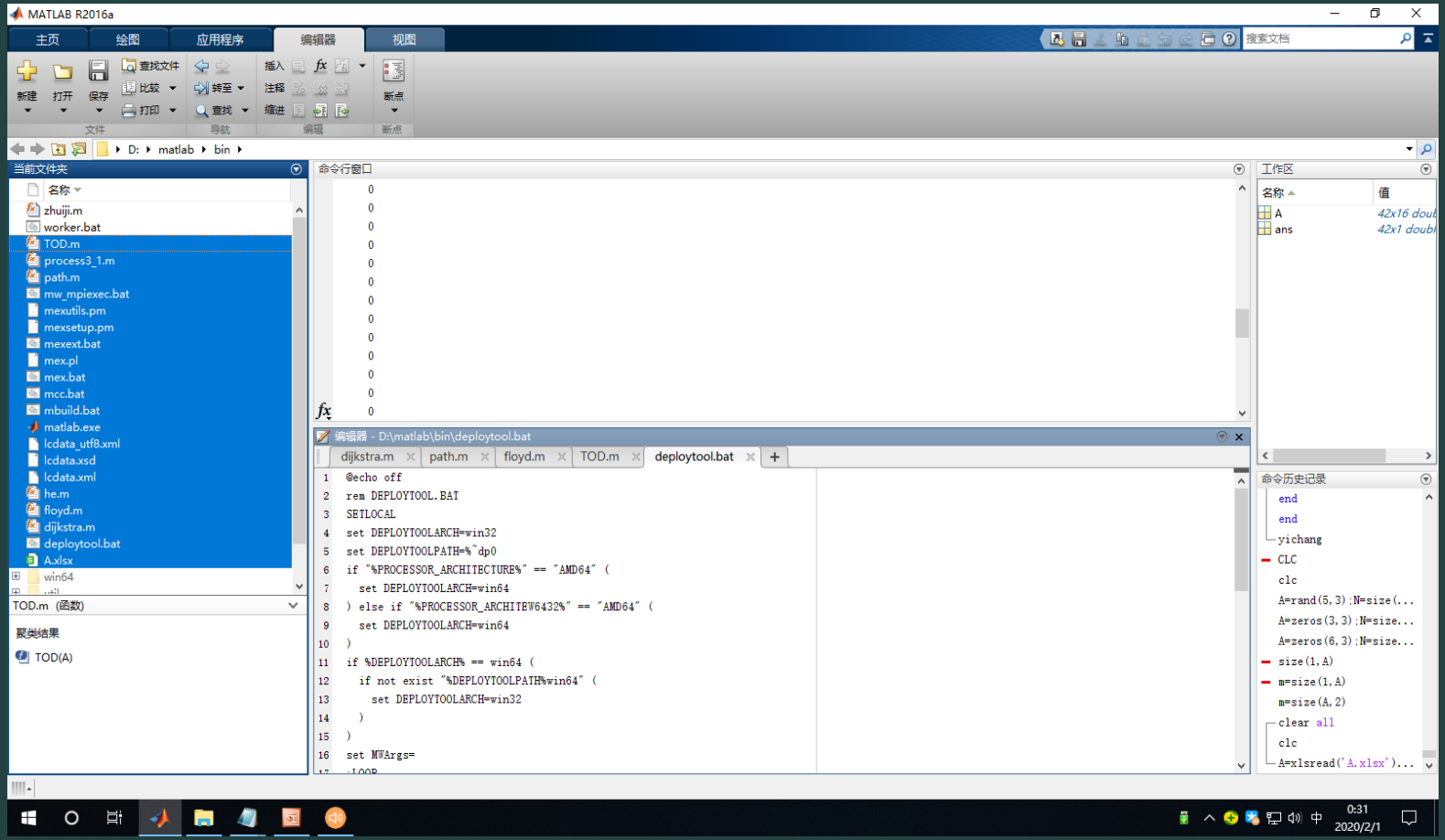
- 其中对象p到每个类之间的距离$d(p, C_i)$有两种计算方法：
- （1） p与类C_i的重心之间的距离；
- （2） p与类C_i中每个样本之间的距离的平均值。
- 以上异常点检测方法也称为两阶段法，也简称为TOD。

- **例** 用基于聚类的方法找出附件数据表A中的异常数据。
- 解：该数据表给出了42个样本，每个样本有16个属性。基于聚类的方法：具体程序见程序TOD.m,β=1.28 。

- 具体操作方法如下
- 第一步，将TOD.m文件以及数据文件A.xlsx一起放入默认路径下

```
function yichang=TOD(A)
N=size(A,1);
D=N;
yichang=zeros(N,1);
K=3;
[u re]=kmeans(A,K);  %K-均值聚类
tt=sort(u);
%聚类结果
    [t1,t2]=find(u==1);
    class1=A(t1,:);
    [t1,t2]=find(u==2);
    class2=A(t1,:);
    [t1,t2]=find(u==K);
    classK=A(t1,:);
    N1=size(class1,1);
    N2=size(class2,1);
    NK=size(classK,1);
  for i=1:size(A,1)
        D1=0;
        D2=0;
        DK=0;
        for j=1:N1
          D1=sqrt((A(i,:)-class1(j,:))*(A(i,:)-class1(j,:))')+D1;
        end
        temp1=N1/N*D1;
        for j=1:N2
          D2=sqrt((A(i,:)-class2(j,:))*(A(i,:)-class2(j,:))')+D2;
        end
        temp2=N2/N*D2;
for j=1:NK
          DK=sqrt((A(i,:)-classK(j,:))*(A(i,:)-classK(j,:))')+DK;
end
       tempK=NK/N*DK;

op(i)=temp1+temp2+tempK;
   end
xmean=mean(op);
xvar=std(op);
bata=1.28;
for i=1:N
    if op(i)>xmean+bata*xvar
      yichang(i)=1;
    end
end
 yichang
```



- 第二步，在命令窗口输入
- clear all
- clc
- A=xlsread('A.xlsx'); yichang=TOD(A)
- 运行，得到运行结果：yichang显示1的就是异常
- 找出的异常点为35，38，40号样本。

总结与体会

- 异常点的检测是在数学建模前很重要的数据预处理步骤，各种检测方法各有优缺点，目前尚无充分的理论证明哪一种检测方法最好，通常可用Kappa检验方法确定不同检测方法之间的一致性，但在数学建模前先检测可能的异常点数据是必不可少的步骤之一。