## 数据预处理——异常值处理

异常值是指样本中明显异于其他值的点,好比与羊群走散的落单羊。 我们要做的就是找到这只落单的羊并把它赶回羊群。 MATLAB 提供了以下两个函数处理异常值:

(1) 检测异常值(isoutlier)

TF = isoutlier(A, method)

A 为输入数组,可以是向量、矩阵或多维数组等。

TF 是一个逻辑数组, 用 1 标识异常值的位置。

method: 字符型参数,指定异常值检测的方法。 方法选项如下表:

方法	说明
'median'	默认方法,与中位数相差超过五倍中位数绝
'mean'	偏差(MAD)的值定义为异常值。 3σ原则、偏离均值超过三倍标准差的值定义 异常值
'quartiles'	指高于上四分位数或低于下四分位数超过1 个四分位范围的值。
'grubbs'	应用 Grubbs 检验以检测离群值,即基于假 检验每次迭代删除一个离群值。 此方法假设 中的数据呈正态分布。
'gesd'	应用广义极端 Student 化偏差检验检测离值。 此迭代方法与 'grubbs' 类似,但当有个离群值互相遮盖时,此方法的执行效果好。

(2) 检测并替换异常值(filloutlier)

B = filloutliers (A, fillmethod, findmethod)

A 为输入数组,可以是向量、矩阵或多维数组等。

findmethod:字符型参数,指定异常值检测的方法。方法选项与上文 isoutlier 相同。

fillmethod: 填充方法, 数值或字符型变量, 方法选项如下表:

B为替换异常值后的数据。

填充方法	说明
数值标量	用指定的数值进行填充,如数字1。
'center'	使用由 find method 决定的中心值进行填充。
ʻclip'	对于比 findmethod 决定的下阈值还小的元素,用下阈值填充。 对于比 findmethod 决定的上阈值还大的元素,用上阈值填充。
'previous'	使用上一个非离群值进行填充。
'next'	使用下一个非离群值进行填充。
'nearest'	使用最接近的非离群值进行填充。
'linear'	使用相邻的非离群值的线性插值进行填充。
'spline'	使用分段三次样条插值进行填充。
'pchip'	使用保形分段三次样条插值进行填充

## % 检测并替换异常值示例

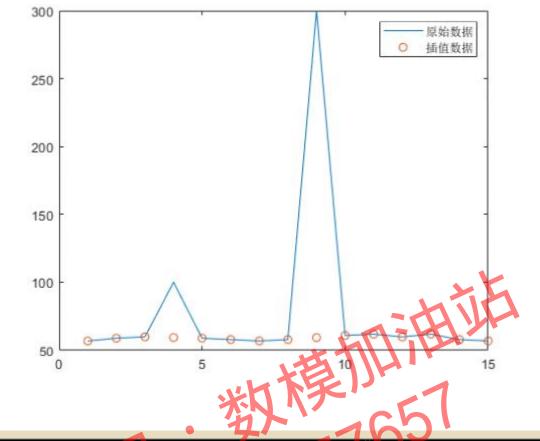
A = [57 59 60 100 59 58 57 58 300 61 62 60 62 58 57];

% 线性插值填充异常值

B = filloutliers(A, 'linear');

plot (1:15, A. 1:15, B. 'o'

legend('原始数据', '插值数据')



异常值处理方法	
删除含有异常值的记录	直接将含有异常值的记录
视为缺失值	将异常值视为缺失值, 和
平均值修正	可用前后两个观测值的平
不处理	直接在具有异常值的数据