7 Detecting Outliers
离群值

利用统计方法和机器学习算法筛出离群值

数学领域，提出问题比解决问题，更珍贵。

In mathematics the art of proposing a question must be held of higher value than solving it.

—— 格奥尔格·康托尔 (Georg Cantor) | 德国数学家 | 1845 ~ 1918

- `numpy.percentile()` 计算百分位
- `pandas.DataFrame()` 构造 pandas 数据帧
- `seaborn.boxplot()` 绘制箱型图
- `seaborn.histplot()` 绘制直方图
- `seaborn.kdeplot()` 绘制概率密度估计曲线
- `seaborn.pairplot()` 绘制成对分析图
- `seaborn.rugplot()` 绘制 rug 图像
- `seaborn.scatterplot()` 绘制散点图
- `sklearn.covariance.EllipticEnvelope()` 协方差椭圆法检测离群值
- `sklearn.ensemble.IsolationForest()` 孤立森林检测离群值
- `sklearn.svm.OneClassSVM()` 支持向量机检测离群值
- `stats.probplot()` 绘制 QQ 图

7.1 这几个数有点不合群？

离群值 (outlier)，又称逸出值、离群值，指的是样本数据中和其他数值差别较大的数值，也就是明显地偏大或偏小。

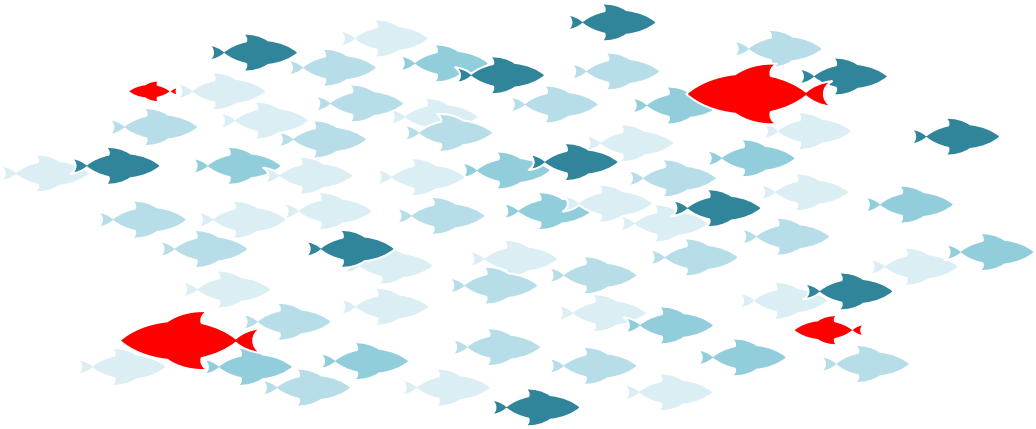


图 1. 离群点

离群值破坏力

离群值可以具有很强的破坏力。比如，离群值可能给最大值、最小值、取值范围、平均值、方差、标准差、分位等计算带来偏差。

图 2 所示为离群值对**线性回归** (linear regression) 的影响。再举个例子，实践中，大家会发现离群值对于时间序列相关性计算破坏力更大。这一章专门介绍各种发现离群值的工具。

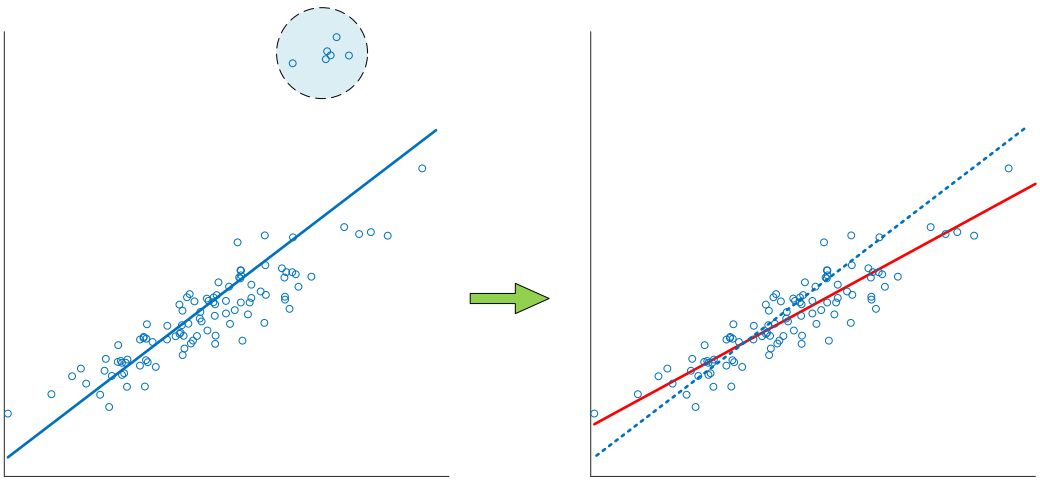


图 2. 离群点对回归分析的影响

工具

如图 3 所示，判断离群值的方法有很多。本章将围绕图 3 中主要方法展开。

最简单的方法是，观察样本数据的最大值和最小值，根据生活常识或专业知识判断，取值范围是否合理。比如，鸢尾花数据集中，如果出现某个样本点的花萼长度为 5.2 米，这显然是个离群点。再举例，鸢尾花任何特征数值肯定不能是负数。

确定离群值之后，需要合理处理。常见的办法有，比如通过设为 NaN 将其删除，或者填充。填充的方法很多，可以参考上一章内容。

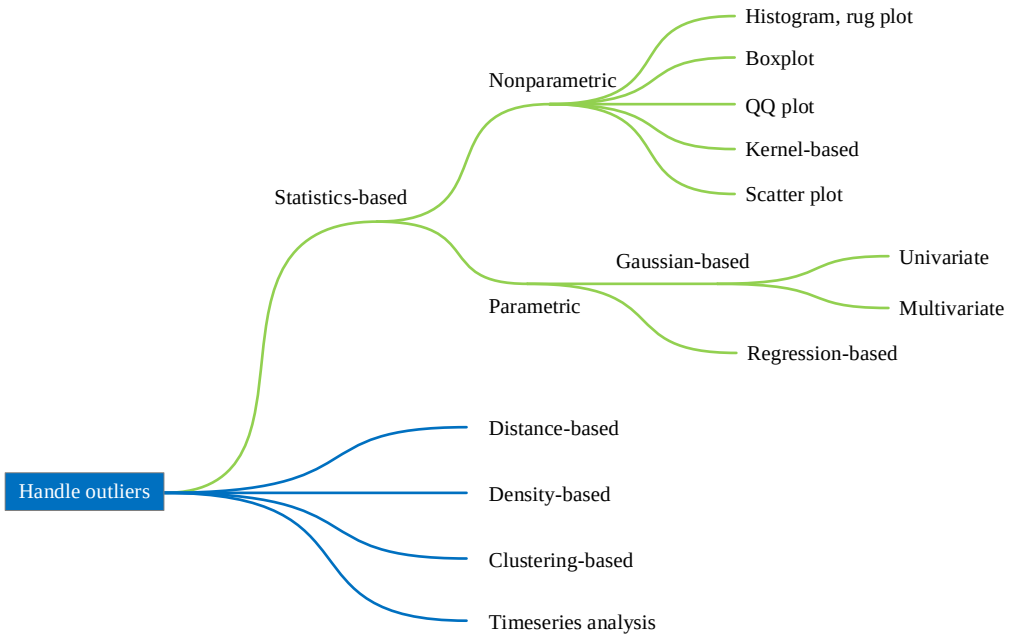


图 3. 发现离群点的常见方法

7.2 直方图：单一特征分布

丛书《统计至简》一本专门介绍过**直方图** (histogram)。可以通过观察数据的直方图来初步判断单一特征的分布情况以及可能存在的离群值。

百分位

图 4 所示鸢尾花四个特征数据的直方图。将数据顺序排列，离群值肯定出现分布的两端。比如，在图 4 上，绘制 1%和 99%百分位所在位置。可以用 1%和 99%百分位来界定数据分布的“左尾”和“右尾”。

另外，25%、50%和 75%这三个百分位也同样重要，图 5 给出了鸢尾花四个特征的这三个百分位所在位置。下一节讲解箱型图时，将使用 25%、50%和 75%这三个百分位。

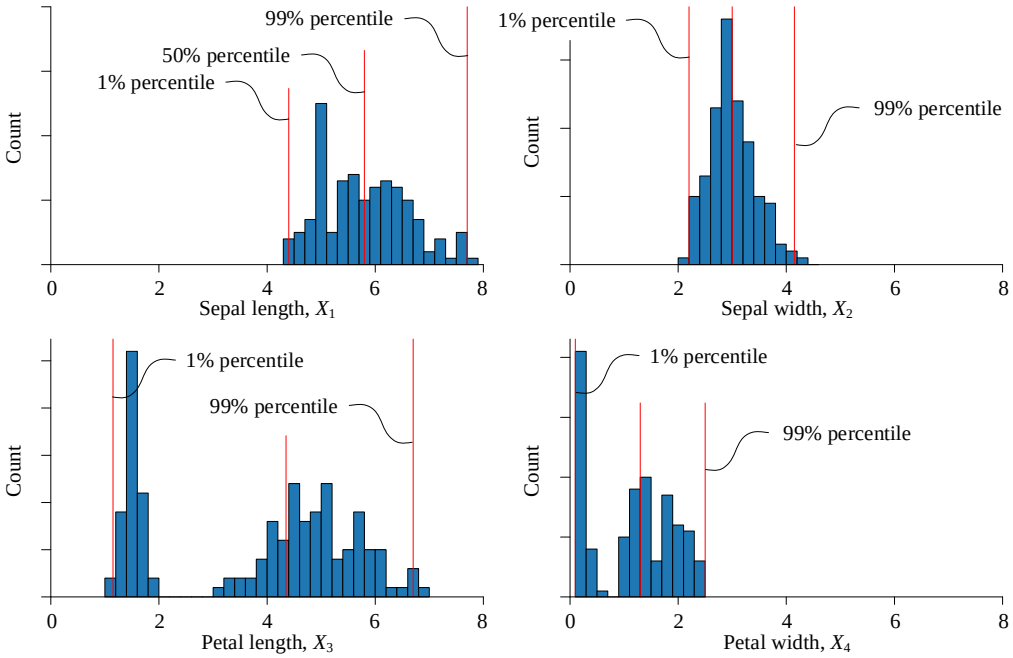


图 4. 鸢尾花数据直方图，以及 1%和 99%百分位

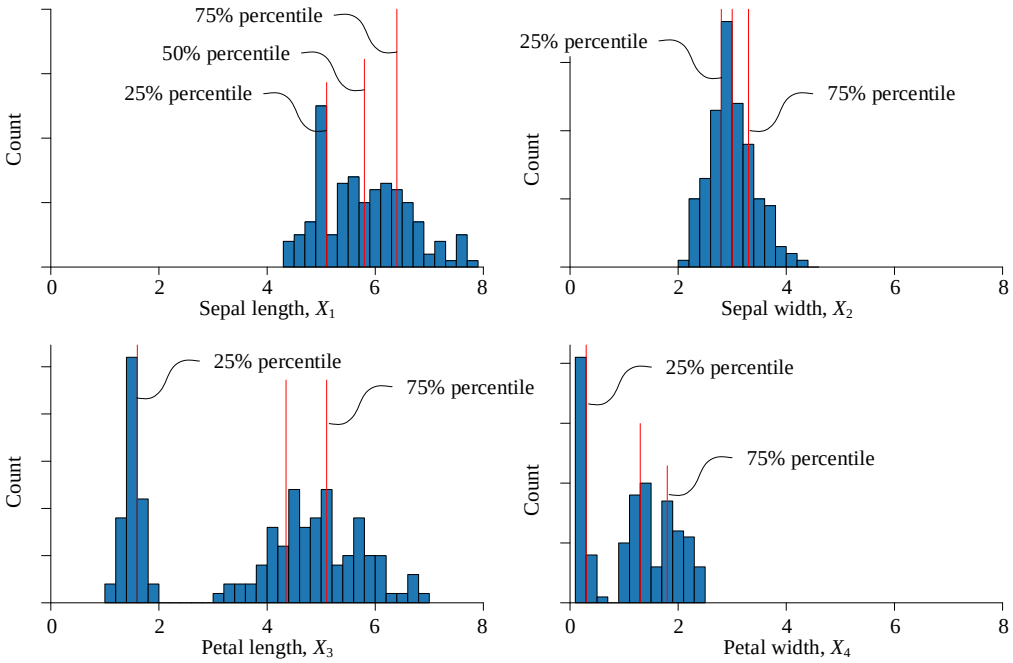


图 5. 鸢尾花数据直方图，以及 25%、50%和 75%百分位

山脊图

图 6 所示为采用 joypy 绘制的山脊图，也可以用来发现分类数据中潜在离群值。

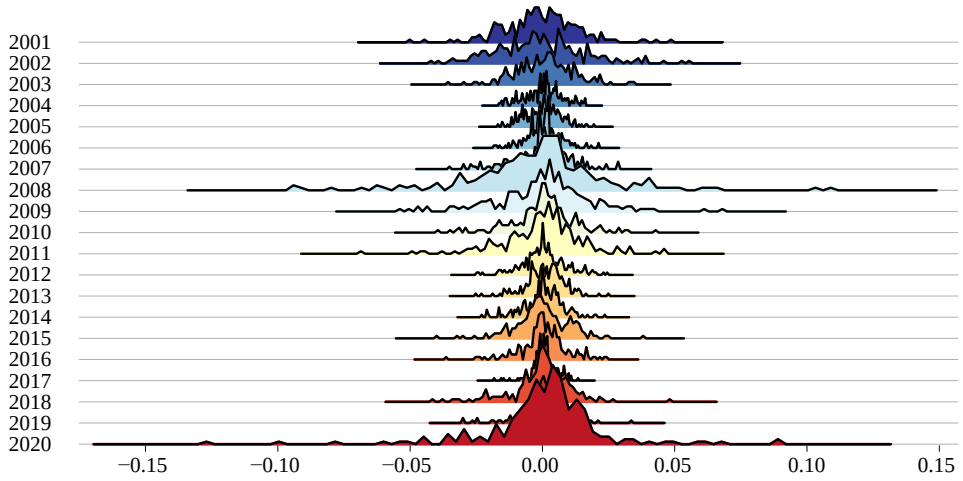


图 6. 标普 500 日收益率数据

概率密度估计 + rug 图

概率密度估计图像也可以用来观察异常值。图 7 所示为 KDE 图像，叠加 rug 图。图上同样标出 1% 和 99%百分位点位置。

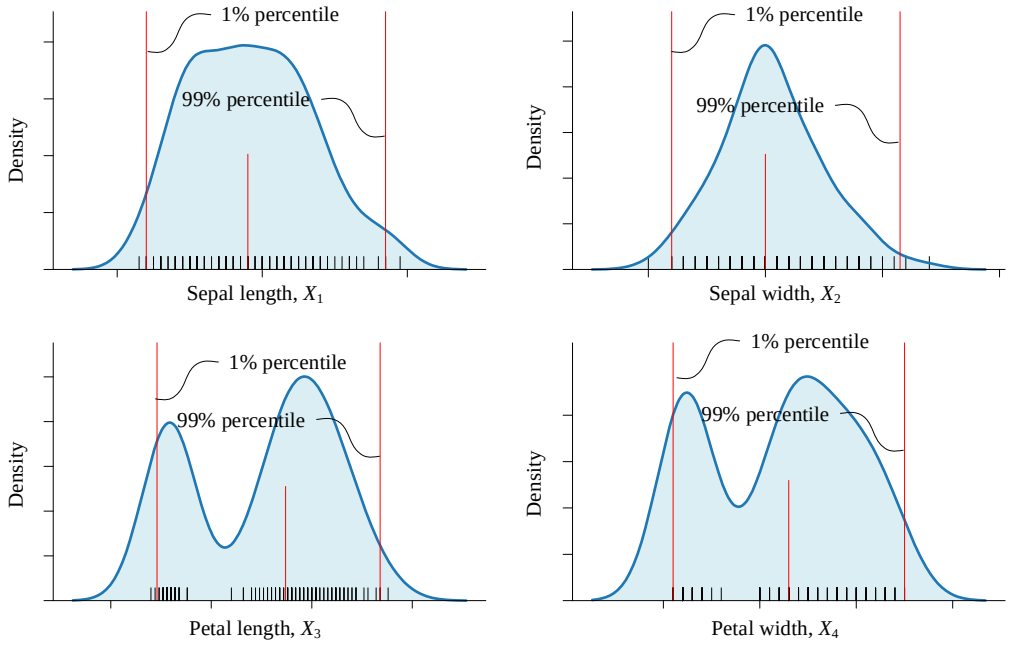


图 7. KDE 密度估计，叠加 rug 图

缩尾调整

缩尾调整 (winsorize) 是将超出变量特定百分位范围的数值替换为其特定百分位数值的方法。请读者参考如下链接学习如何使用 scipy.stats.mstats.winsorize() 函数进行缩尾调整：

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.winsorize.html

7.3 散点图：成对特征分布

本章前文所讲的可视化方案均用来发现单一特征可能存在的离群值。采用散点图，发现成对特征数据可能存在的离散点。图 8 所示为鸢尾花数据花萼长度、花萼宽度散点图。图 8 中还绘制了单一特征的 rug 图。

此外，也可以使用如图 9 成对特征数据来观察数据分布，以及可能存在的离群值。

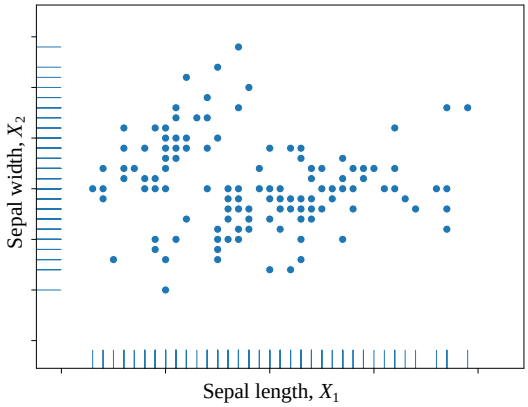


图 8. 散点图，横轴花萼长度，纵轴花萼宽度

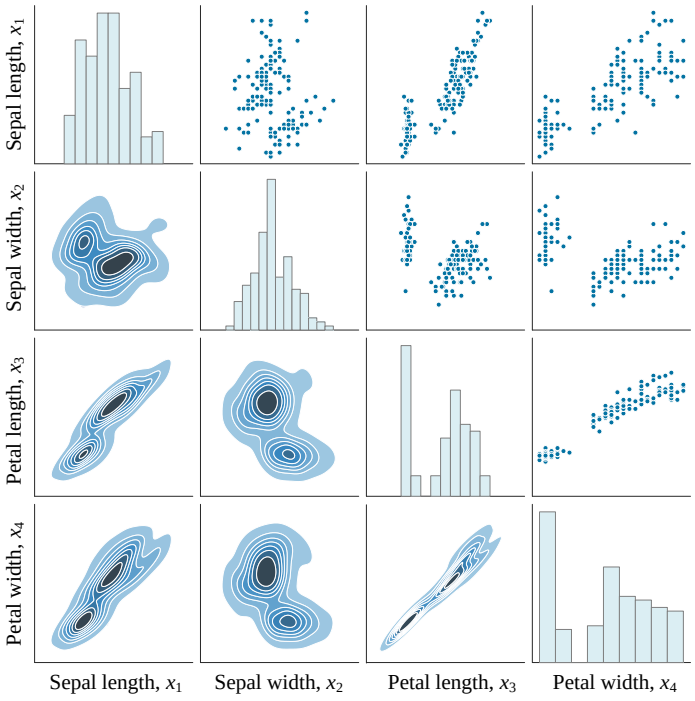


图 9. 鸢尾花数据成对特征分析图

7.4 QQ 图：分位数–分位数

《统计至简》第 9 章介绍过 QQ 图。QQ 图是散点图，也可以用来发现离群值。相信大家已经清楚，QQ 图的横坐标为某一样本的分位数，纵坐标则是另一样本的分位数。QQ 图的纵坐标一般是正态分

布，当然也可以是其他分布。如果两分布相似，散点在 QQ 图上趋近于落在一条直线上。图 10 所示为 QQ 图原理，图中横轴为正态分布的分位数。

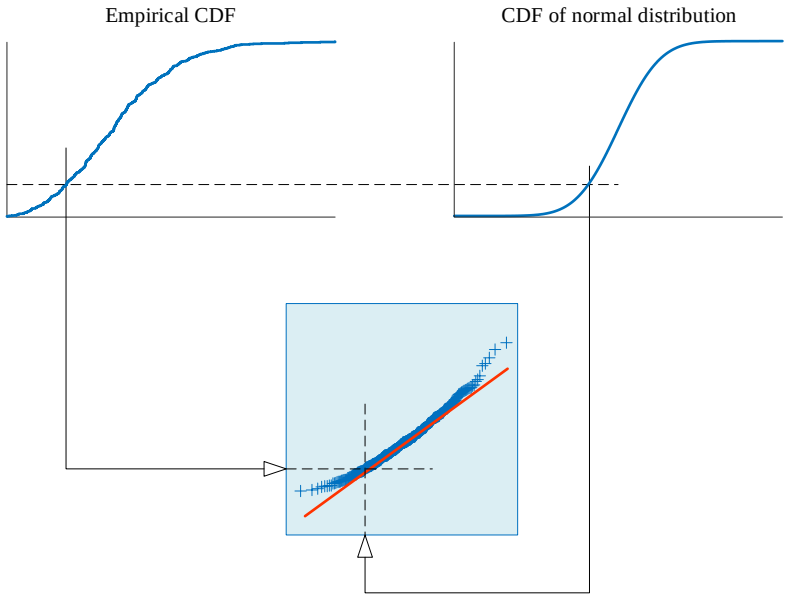


图 10. QQ 图原理

图 11 到图 14 分别给出鸢尾花四个特征数据的直方图和 QQ 图。容易发现不同的数据分布，对应特定的 QQ 图分布特点。《统计至简》第 9 章介绍过如何通过 QQ 图形态判断原始数据分布特点，请大家自行回顾，本节不再重复。

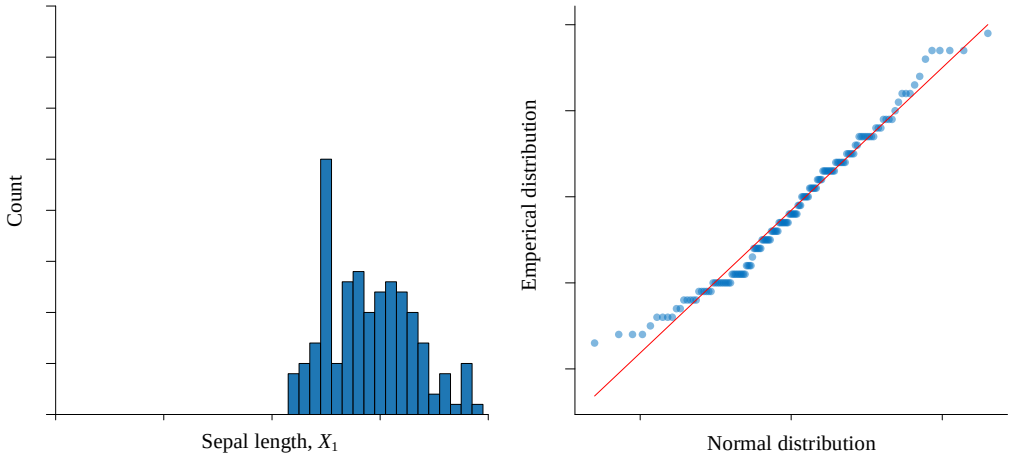


图 11. 花萼长度直方图和 QQ 图

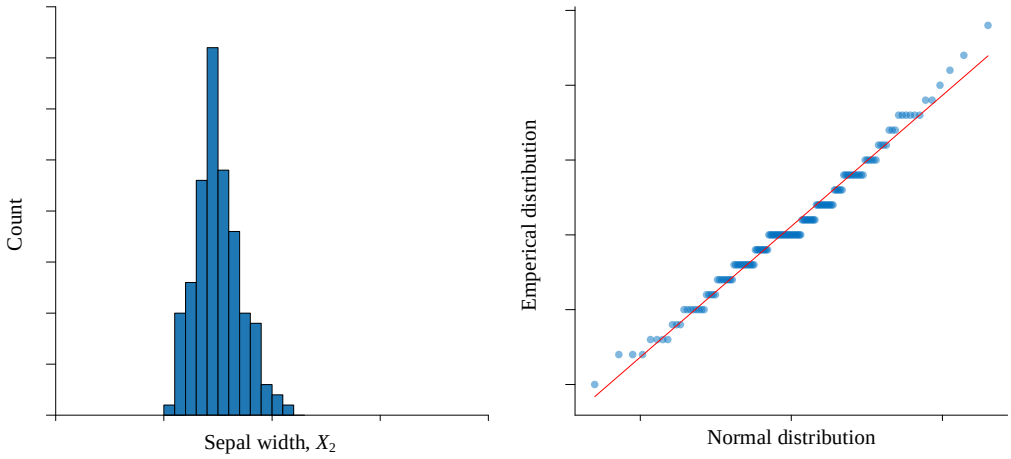


图 12. 花萼宽度直方图和 QQ 图

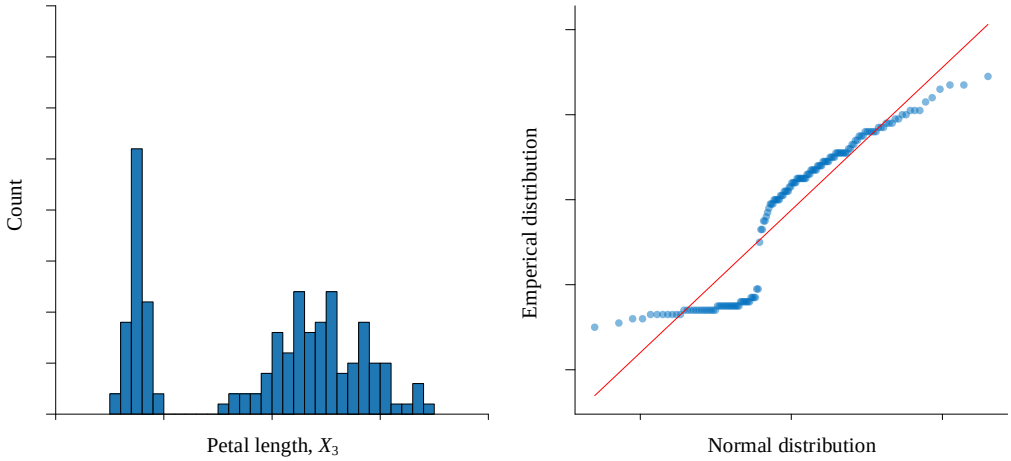


图 13. 花瓣长度直方图和 QQ 图

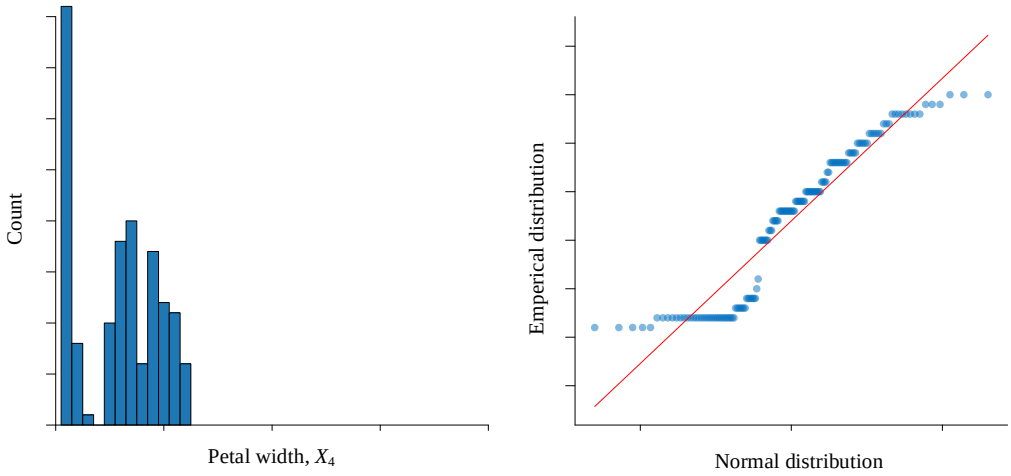


图 14. 花瓣宽度直方图和 QQ 图

7.5 箱型图：上界、下界之外样本

《统计至简》也专门介绍箱型图 (box plot)，箱型图也可以用来分析离群点。图 15 所示为箱型图原理。箱型图利用第一 (Q_1)、第二 (Q_2) 和第三 (Q_4) 四分位数展示数据分散情况。Q_1 也叫下四分位，Q_2 也叫中位数，Q_3 也称上四分位。

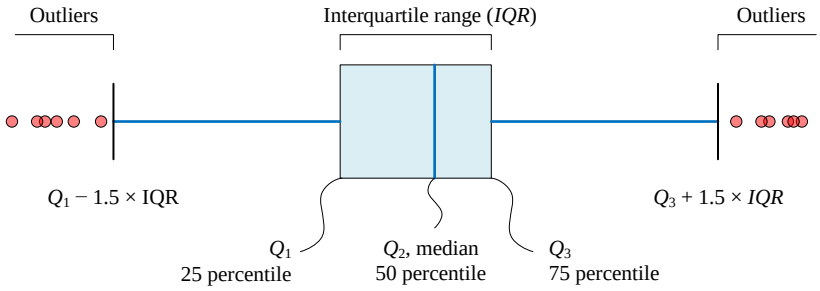


图 15. 箱型图原理

箱型图的四分位间距 (interquartile range) 的定义为：

$$IQR = Q_3 - Q_1 \tag{1}$$

在 $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ 之外的样本数据则可能是离群点。图 16 所示为鸢尾花数据的箱型图。$Q_3 + 1.5 \times IQR$ 也称上界，$Q_1 - 1.5 \times IQR$ 叫下界。

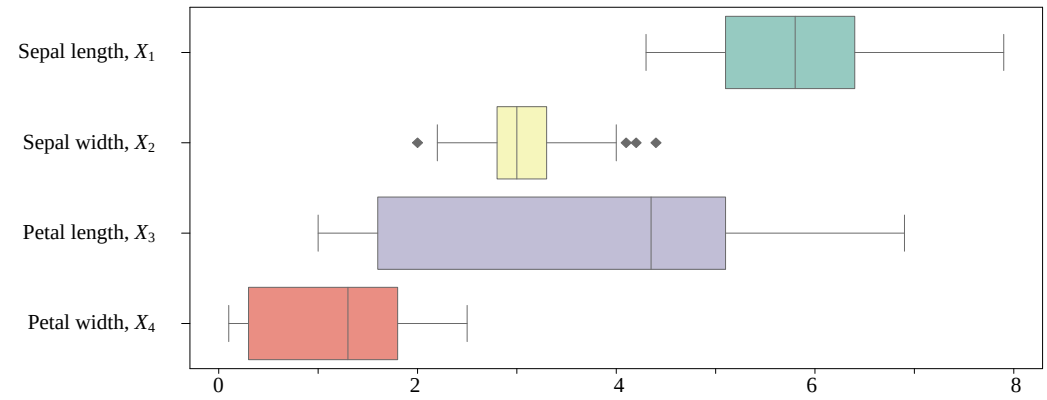


图 16. 鸢尾花箱型图

观察图 16，我们会发现用 Seaborn 绘制的箱型图左须距离 Q_1、右须距离 Q_3 宽度并不相同。这一点，我们在《编程不难》曾经提过。根据 Seaborn 的技术文档，左须、右须延伸至该范围 $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ 内最远的样本点，具体如图 17 所示。更为极端的样本会被标记为异常值。

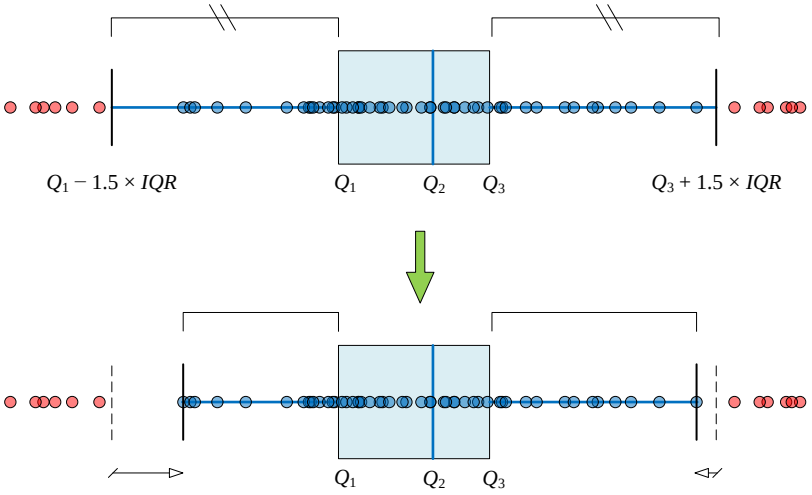


图 17. Seaborn 绘制箱型图左须、右须位置

7.6 Z 分数：样本数据标准化

从大到小排列一组 n 个样本数据，离群值肯定出现在序列的两端。首先计算出数据的样本均值 $\bar{x}$，和样本标准差 s。若任何数据点与均值的偏差绝对值大于三倍标准差，则可以判定数据点为离群点，即满足下式的 x 可能是离群值：

$$|x - \bar{x}| > 3s \tag{2}$$

此外，大家需要注意极大的离群值会“污染”样本均值。因此，实践中，也常用样本中位数作为基准。

三倍标准差$\pm 3s$ 相当于 99.7%置信度，对应显著性水平 $\alpha = 0.003$。此外，也可以采用两倍标准差 $\pm 2s$，这相当于 95%置信度，即 $\alpha = 0.05$。

图 18 展示了《统计至简》一册介绍的 68–95–99.7 法则，请大家回顾。注意，图中并不区分总体标准差 σ 和样本标准差 s，并假设均值为 0。

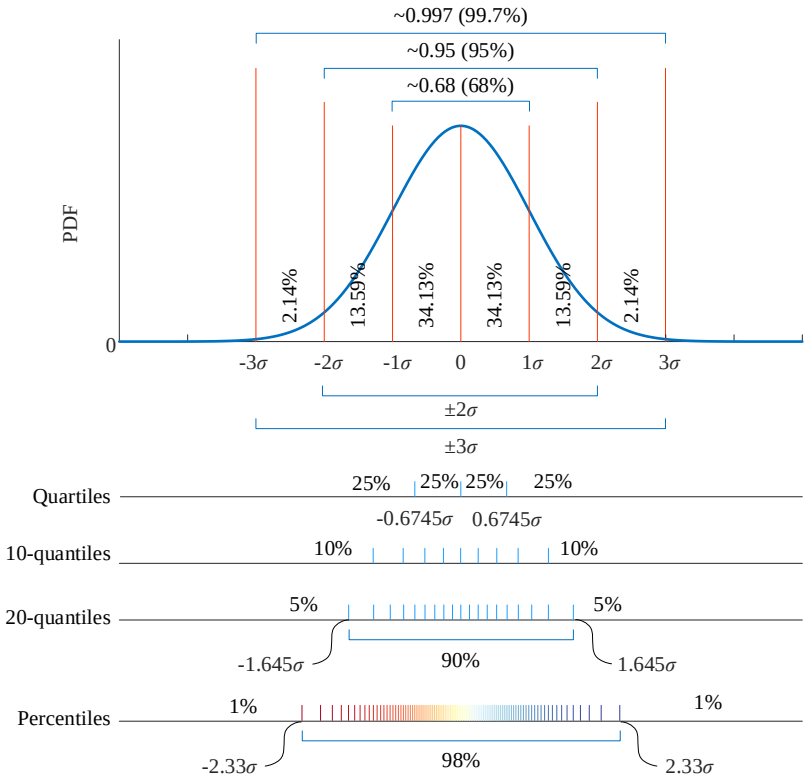


图 18. 标准差，注意图中并不区分总体标准差 σ 和样本标准差 s

Z 分数

从 Z 分数 (z score) 角度，(2) 相当于：

$$z = \frac{|x - \bar{x}|}{s} > 3 \tag{3}$$

也就是任何数据点的 Z 分数绝对值大于 3，即 Z 分数大于 3 或小于−3，可以判定数据点为离群点。图 19 所示为鸢尾花数据四个特征的 Z 分数。

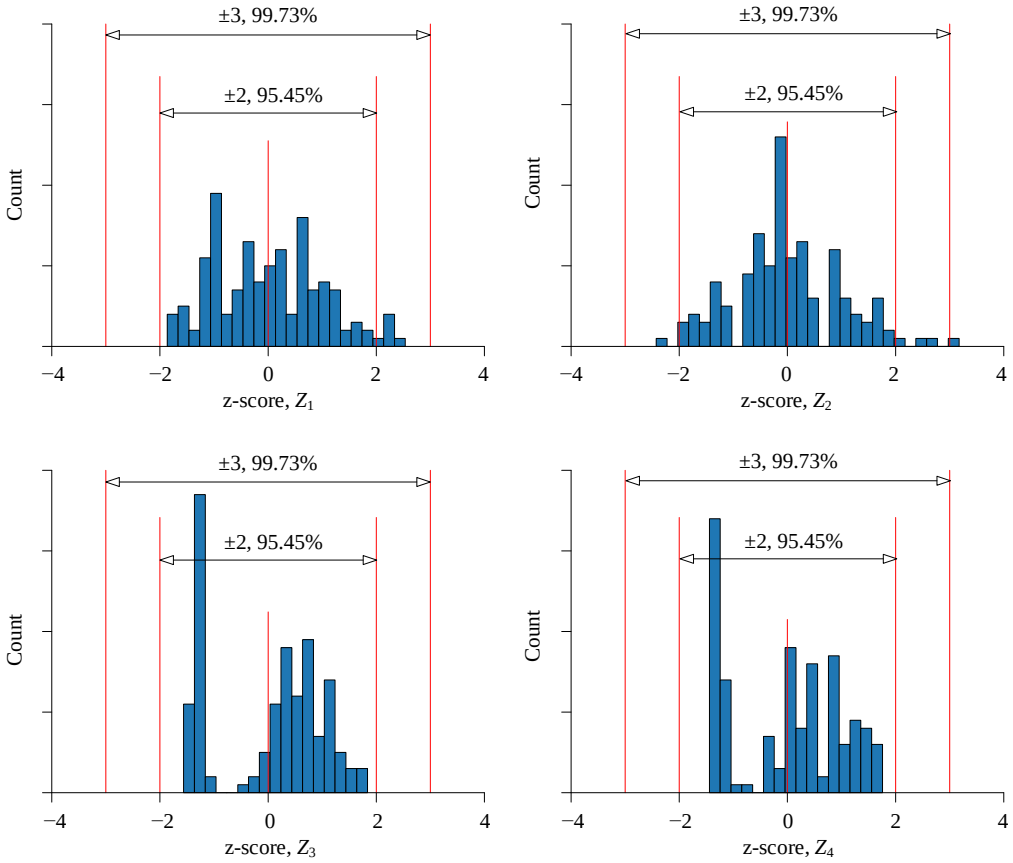


图 19. 鸢尾花 Z 分数

7.7 马氏距离和其他方法

对于二维乃至多维的情况，我们也可以使用 Z 分数。这个 Z 分数就是马氏距离 (Mahalanobis distance)，《统计至简》一册专门讲解过马氏距离。具体定义如下：

$$d(\boldsymbol{x},\boldsymbol{q}) = \sqrt{(\boldsymbol{x}-\boldsymbol{q})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{q})} \tag{4}$$

其中，$\boldsymbol{\Sigma}$ 为样本数矩阵 $\boldsymbol{X}$ 方差协方差矩阵。

如果样本数据分布近似服从多元高斯分布，马氏距离则可以作为判定离群值的有效手段。图 20 (a) 所示为，不同的马氏距离等高线对应不同的置信区间。图 20 (b) 而所示为$\pm\sigma$ ~ $\pm 4\sigma$ 置信区间。

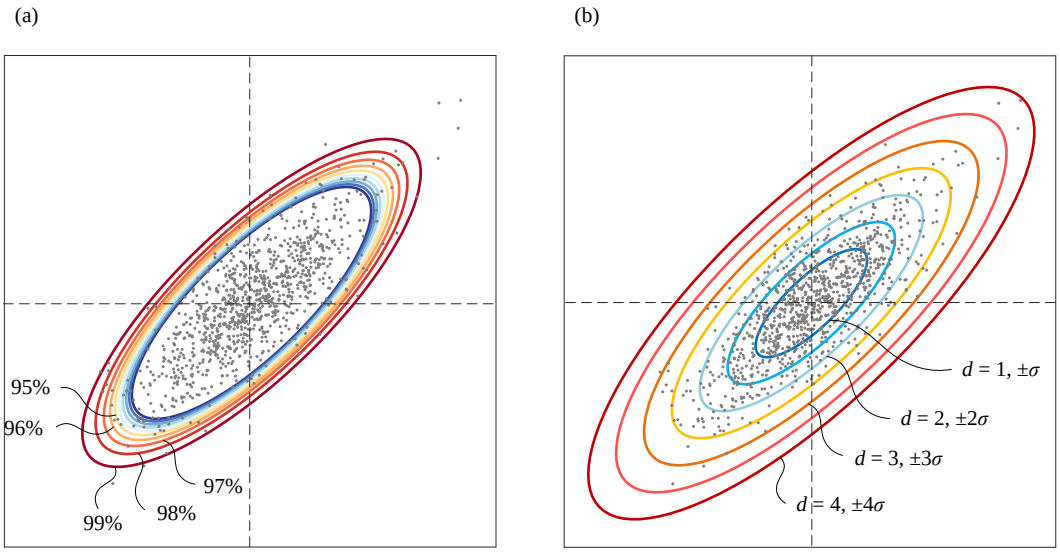


图 20. 协方差椭圆：(a) 95% ~ 99% 置信区间；(b) ±σ ~ ±4σ 置信区间

Scikit-learn 提供一个 covariance.EllipticEnvelope 对象，它就是利用马氏距离椭圆来判断离群点。图 21 所示为鸢尾花花萼长度、花萼宽度的散点图，和马氏距离为 2 的旋转椭圆。这个旋转椭圆之外的样本点可能是离群值。

有关马氏距离、卡方分布、置信区间关系，请大家参考《统计至简》第 23 章。

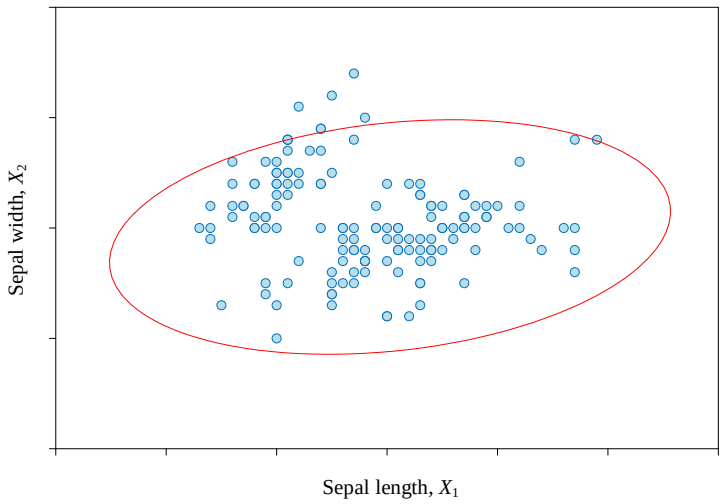


图 21. 鸢尾花数据前两个特征构造的协方差椭圆，马氏距离为 2

代码 Bk6_Ch07_01.ipynb 绘制本章前文主要图片。

概率密度估计检测离群值

马氏距离实际上假设数据服从多元正态分布。当多特征数据分布情况较大偏离多元正态分布，马氏距离就会失效。这时我们可以用概率密度估计来检测离群值。如图 22 所示，KDE 概率密度估计没有预设数据分布假设。有关 KDE 概率密度估计，大家可以回顾《统计至简》第 18 章。

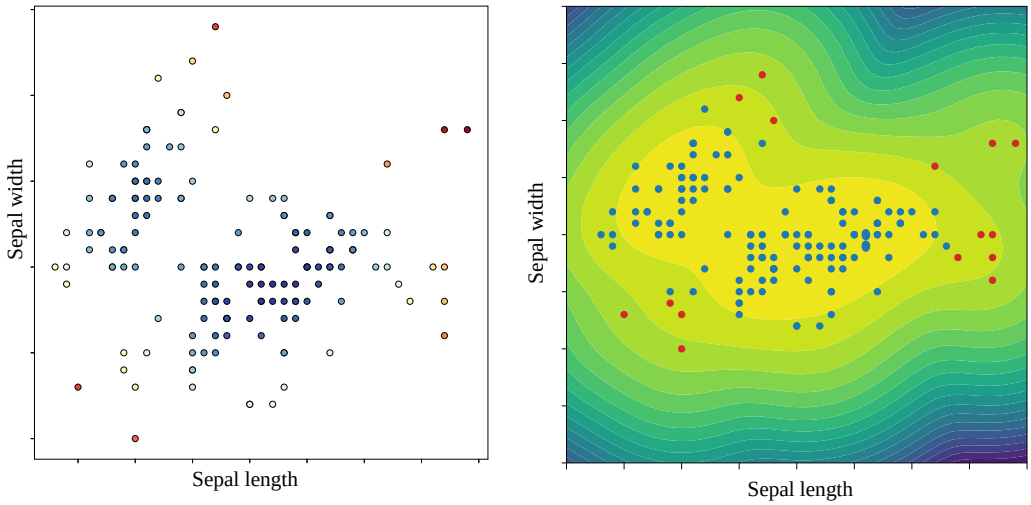


图 22. 概率密度估计判断离群值，左图散点颜色对应数据 KDE 概率密度估算值

机器学习方法

机器学习中很多算法都可以用来判断离群值。图 23 所示为用支持向量机和孤立森林算法判断鸢尾花数据中可能存在的离群值。更多机器学习算法，请大家参考《机器学习》一书。

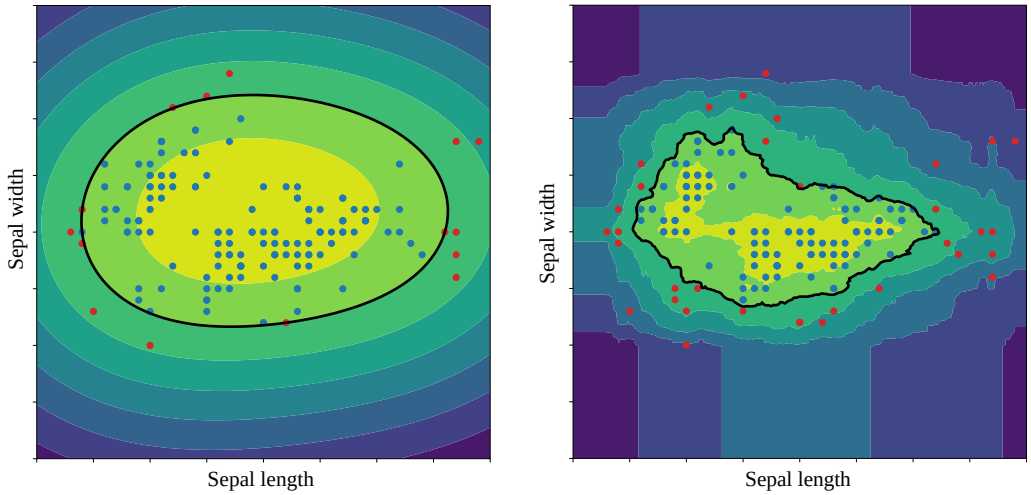
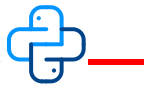
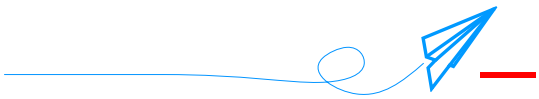


图 23. 支持向量机和孤立森林算法判定离群值



Bk6_Ch07_02.ipynb 绘制图 22 和图 23。



总结来说，离群值是指在数据集中与大多数观测值显著不同的那些观测值。它们可能是由于测量错误、异常情况或者真实但罕见的事件引起的。在机器学习中，离群值可能对模型产生负面影响，离群值的影响包括可能导致模型的偏离、降低模型的准确性，并影响对模型的解释性。

本章介绍了发现离群值的几种常用方法，比如直方图、散点图、QQ 图、箱型图、Z 分数、马氏距离、机器学习方法等等。

处理离群值时，最简单的办法就是直接删除离群值。但要小心不要过度删除，以免损失重要信息。我们也可以将离群值截断为某个特定的阈值，使其不超过该阈值。此外，我们还可以使用中位数、均值或其他统计量替换离群值。还有，对数据进行转换，如取对数，可以减缓离群值对模型的影响。

Scikit-learn 中有更多利用机器学习方法检测离群值的方法，请参考下例。

https://scikit-learn.org/stable/modules/outlier_detection.html

建议大家学完丛书《机器学习》一册内容，再回过头来自学这几个例子。