3 离群值

Detecting Outliers

利用统计方法和机器学习算法发现、处理离群值

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

3.1 离群值小传

离群值 (outlier)，又称逸出值、离群值，是指数据集中与其他数据点有显著差异的数据点，也就是说明显地偏大或偏小。离群值可能是由于异常情况、错误测量、数据录入错误或意外事件等原因而产生。离群值可能会对数据分析和建模造成问题，因为它们可能导致误差或偏差，并降低模型的准确性。因此，数据分析师通常会对数据集中的离群值进行检测和处理。

常见的离群值检测方法包括基于统计学的方法、基于距离的方法、基于密度的方法和基于模型的方法。处理离群值的方法包括删除、替换、调整或利用异常值建立新的模型等。

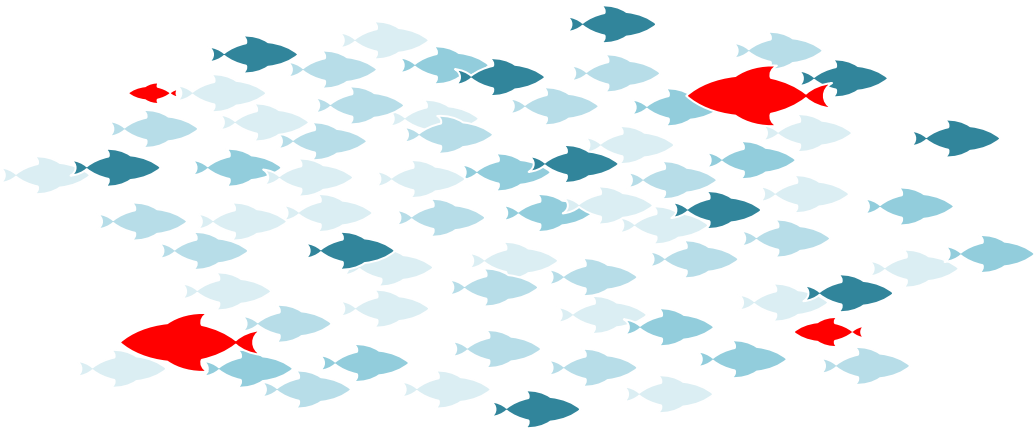

图 1. 离群点

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、极差、平均值、方差、标准差、线性相关性系数、分位等统计量计算带来偏差。

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系数计算破坏力更大。这一章专门介绍各种发现离群值的工具。

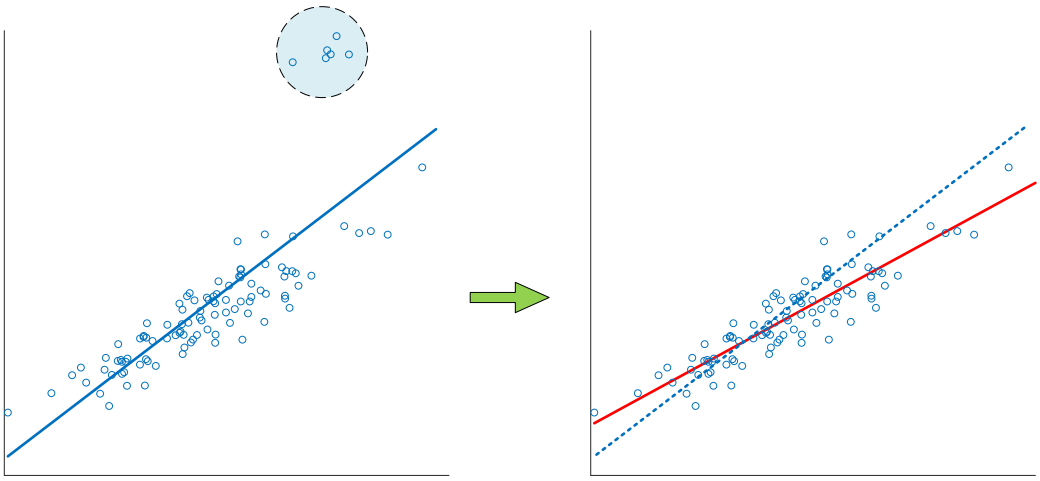

图 2. 离群点对回归分析的影响

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。这幅图也相当于是本章的思维导图。

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。

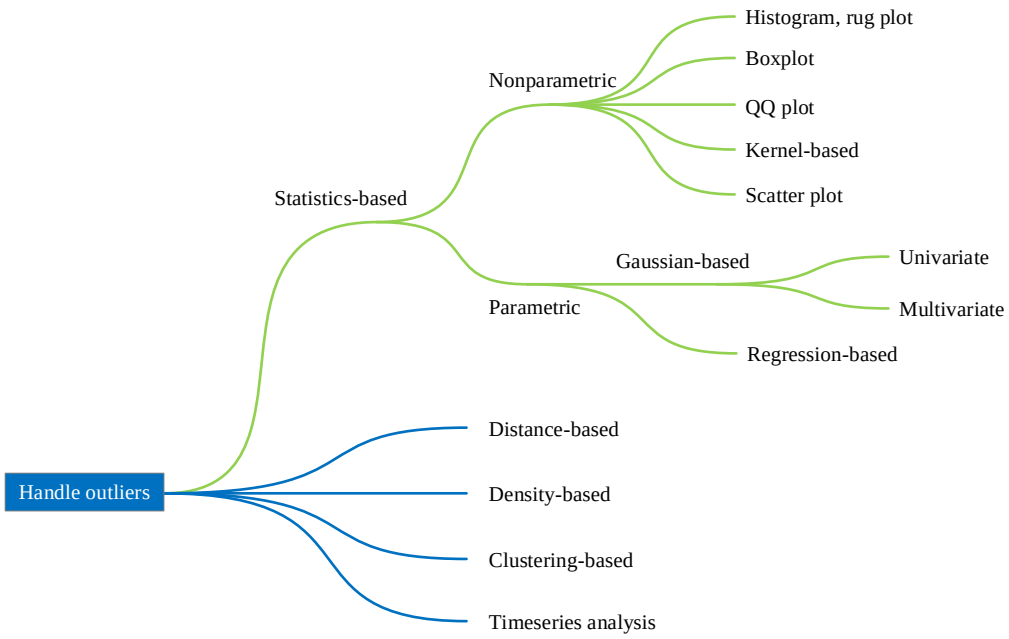

图 3. 处理离群点的常见方法

3.2 直方图：单一特征分布

鸢尾花书《统计至简》第 2 章专门介绍过**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 1%和 99%百分位用来界定数据分布的“左尾”和“右尾”。

回顾一下，百分位 (percentile) 是指一个数值在一组数据中的排名位置，表示该数值小于等于百分位数的观测值所占的百分比。例如，50%百分位数是中位数，表示一半的数据小于等于中位数，另一半的数据大于等于中位数。

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。

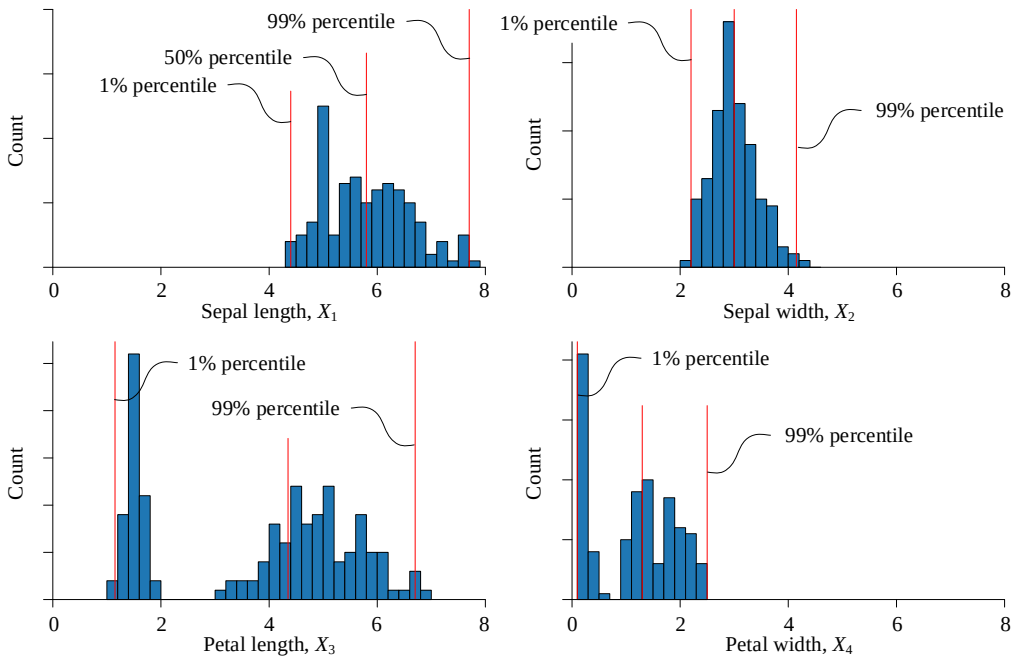

图 4. 鸢尾花数据直方图，以及 1%和 99%百分位

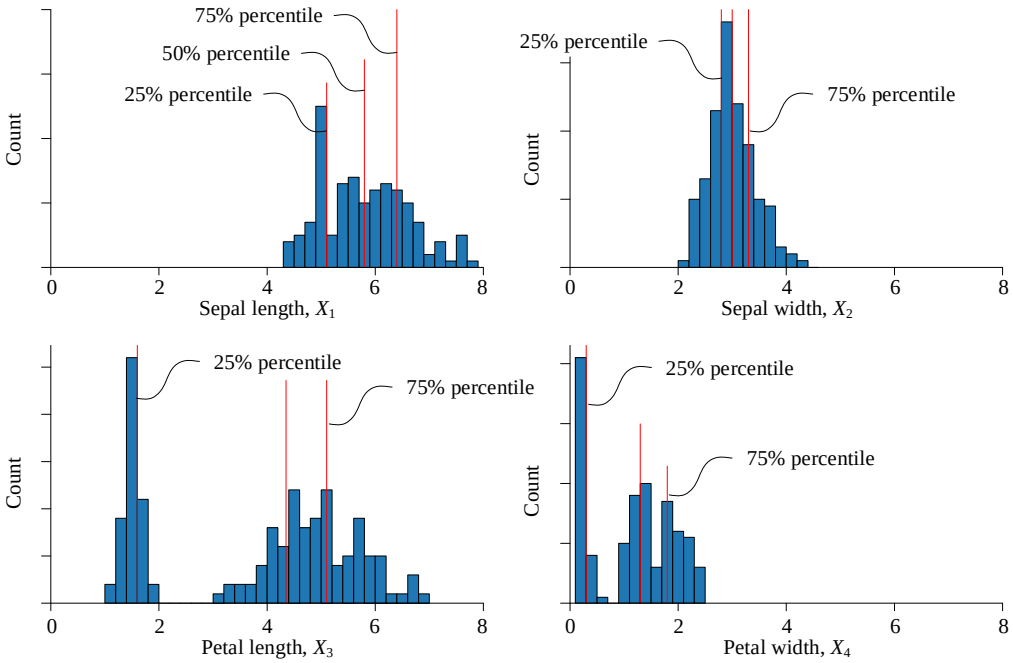

图 5. 鸢尾花数据直方图，以及 25%、50%和 75%百分位

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。

《可视之美》曾专门介绍过山脊图。

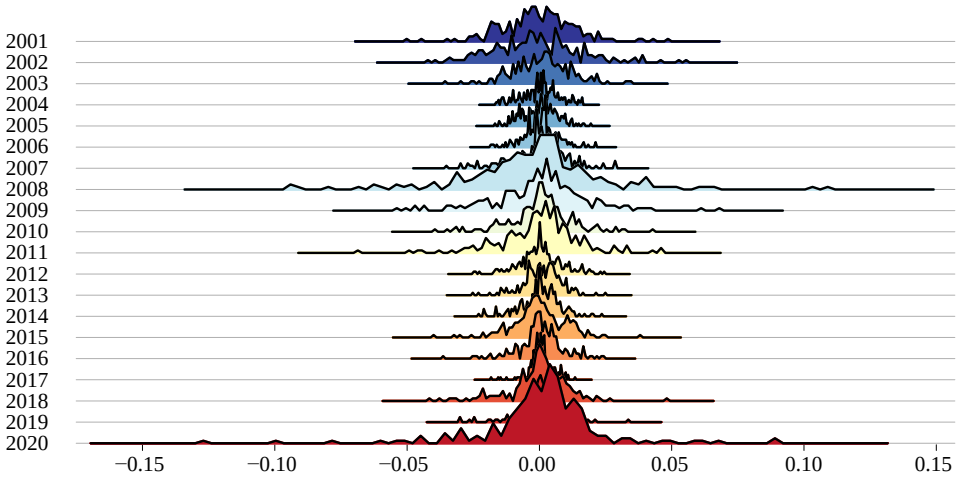

图 6. 标普 500 日收益率数据

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。概率密度估计 (Probability Density Estimation) 是指根据有限样本数据推断出未知概率密度函数的过程，常用于探索性数据分析和模型构建中。通过估计概率密度函数，可以更好地理解数据的分布特征、模型参数和模型拟合度。

高斯核密度估计 (Gaussian Kernel Density Estimation)，或高斯 KDE，是一种常用的概率密度估计方法，基于高斯核函数对数据进行平滑处理，估计未知的概率密度函数。该方法对连续变量的数据有较好的适用性，可以用于探索数据分布、识别离群值和构建概率模型等任务。

图 7 所示为高斯 KDE 图像，叠加 rug 图。图上同样标出 1%和 99%百分位点位置。rug 图是一种数据可视化方法，用于展示数据分布和密度。它将每个数据点在 x 轴上表示为一条短线，形成了数据点的密度分布图。rug 图通常与直方图或核密度图结合使用，可以更直观地显示数据集的分布情况。

《统计至简》专门讲解概率密度估计，请大家回顾高斯核密度估计。

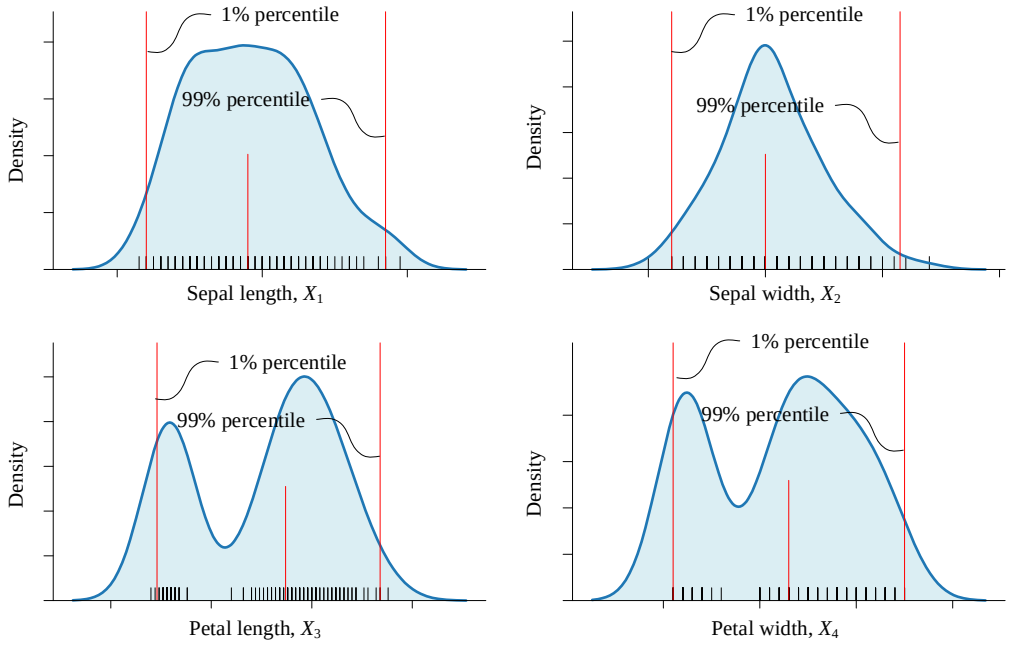

图 7. KDE 密度估计，叠加 rug 图

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。缩尾调整通过截断分布的长尾部分来减少异常值对估计结果的影响。在实际应用中，我们可以根据领域知识或经验选择合适的截断点，并将超出截断点的异常值设置为固定的截断值。缩尾调整

可以改善分布拟合和参数估计的稳定性和精度，但也可能引入信息损失和偏差。在选择截断点时需要谨慎，并在分析前后进行敏感性分析。

请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

3.3 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。鸢尾花书读者对散点图肯定很熟悉。**散点图** (scatter plot) 是一种常用的数据可视化方法，用于展示两个变量之间的关系。散点图将每个数据点表示为一个点，在二维坐标系上绘制，其中一个变量在横轴上表示，另一个变量在纵轴上表示。

散点图可以帮助我们直观地观察变量之间的相关性、趋势和异常值，是探索性数据分析和建模中不可或缺的工具。散点图还可以用于比较不同组之间的变化和趋势，或者用不同的颜色或形状表示不同的组或类别。

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。

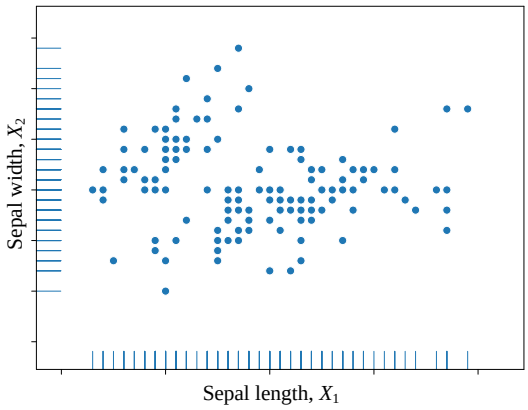

图 8. 散点图，横轴花萼长度，纵轴花萼宽度

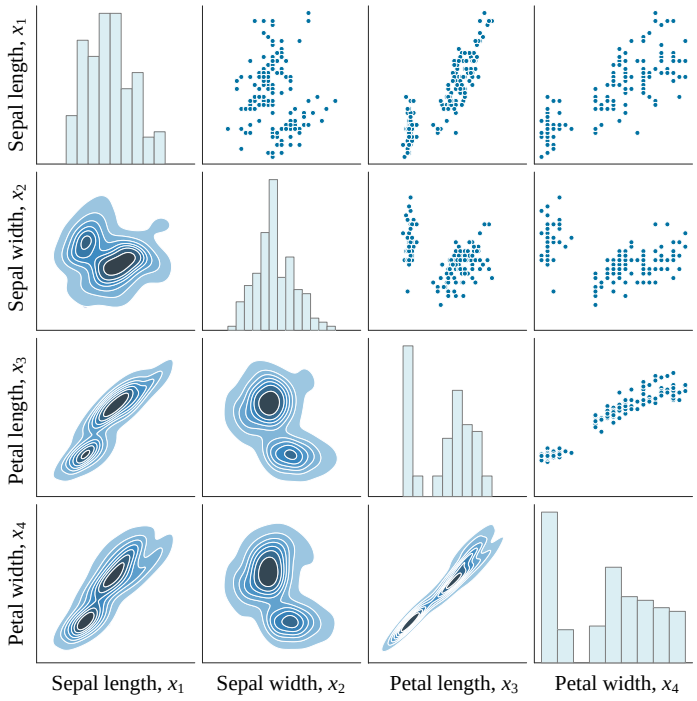

图 9. 鸢尾花数据成对特征分析图

3.4 QQ 图：分位数–分位数

《统计至简》第 9 章专门介绍过 QQ 图。

QQ 图 (Quantile-Quantile plot) 是一种用于检查数据是否符合某种理论分布的数据可视化方法。QQ 图将样本数据的分位数与理论分布的分位数进行比较，并将它们绘制在同一坐标系中。如果数据符合理论分布，则点将沿着一条直线分布。如果数据偏离理论分布，则点将偏离直线。通过观察点的分布情况，我们可以判断数据是否符合某种理论分布，或者是否存在偏差或离群值等问题。QQ 图常用于正态性检验、分布拟合和模型诊断等任务。

QQ 图的横坐标通常是理论分布的分位数，纵坐标通常是样本数据的分位数。在正态 QQ 图中，横坐标通常是标准正态分布的分位数，或 Z 分数；纵坐标是样本数据的分位数。在其他类型的 QQ 图中，横坐标和纵坐标的标尺将取决于所使用的理论分布和样本数据的类型。

图 10 所示为 QQ 图原理，图中横轴为正态分布的分位数。

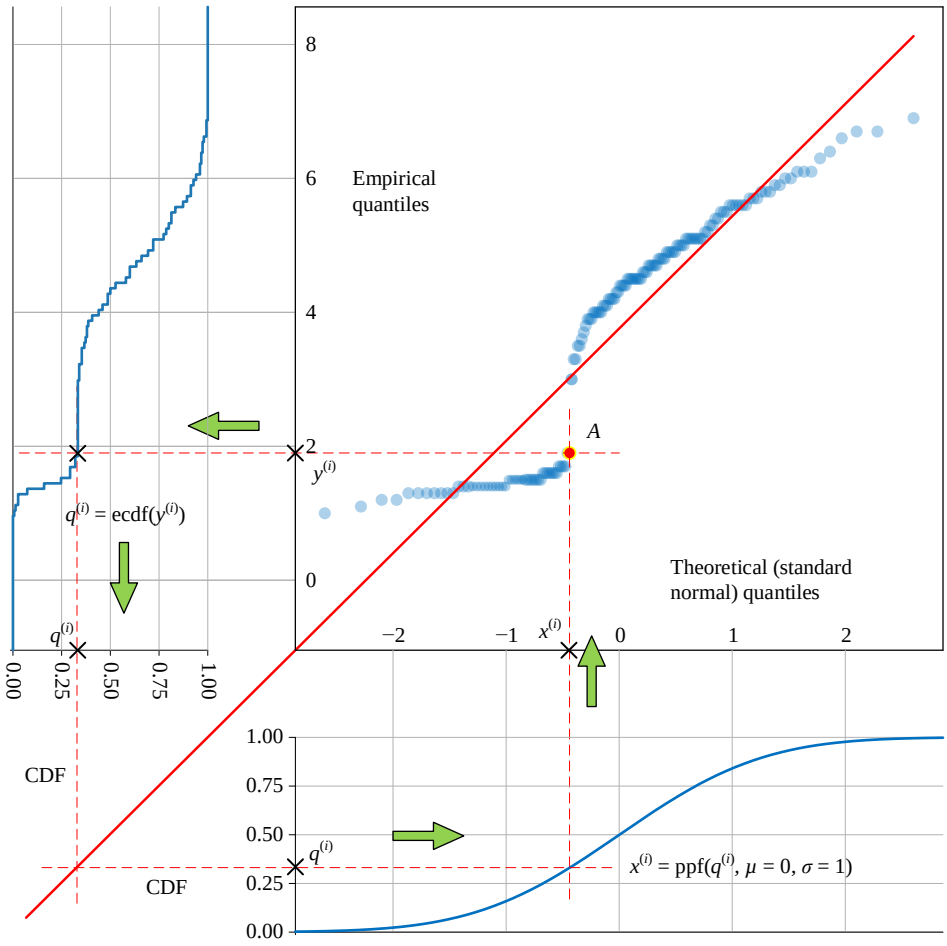

图 10. QQ 图原理，横轴为正态分布，图片来自《统计至简》第 9 章

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。

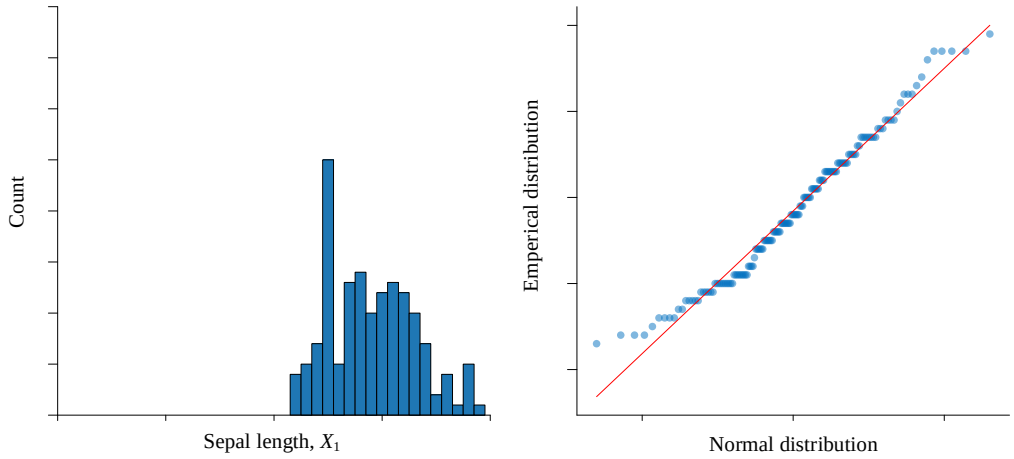

图 11. 花萼长度直方图和 QQ 图

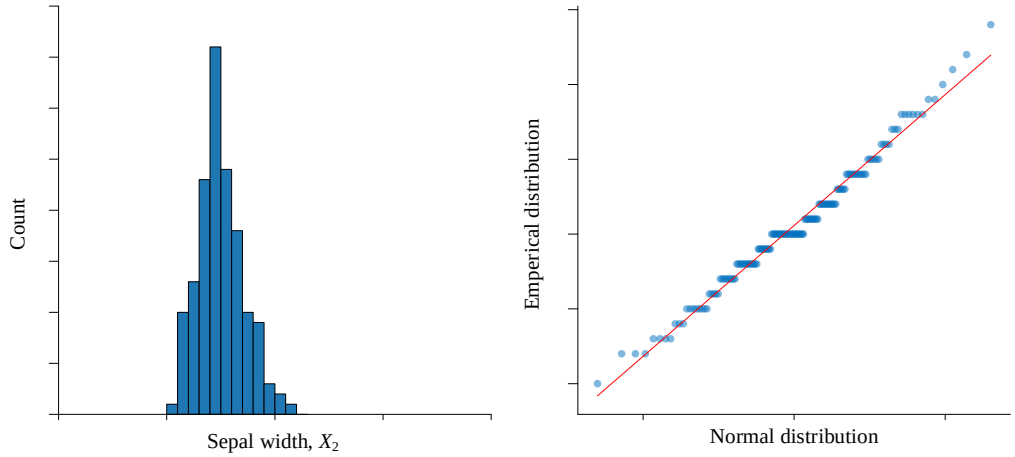

图 12. 花萼宽度直方图和 QQ 图

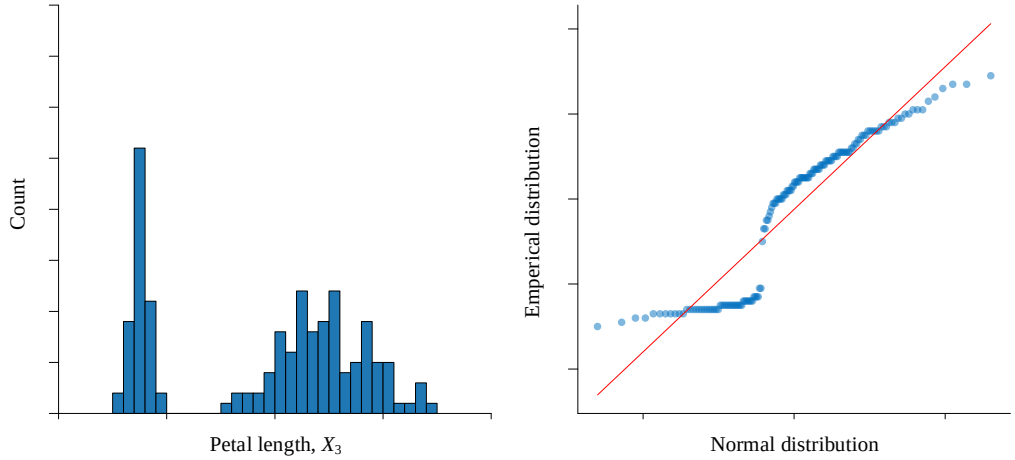

图 13. 花瓣长度直方图和 QQ 图

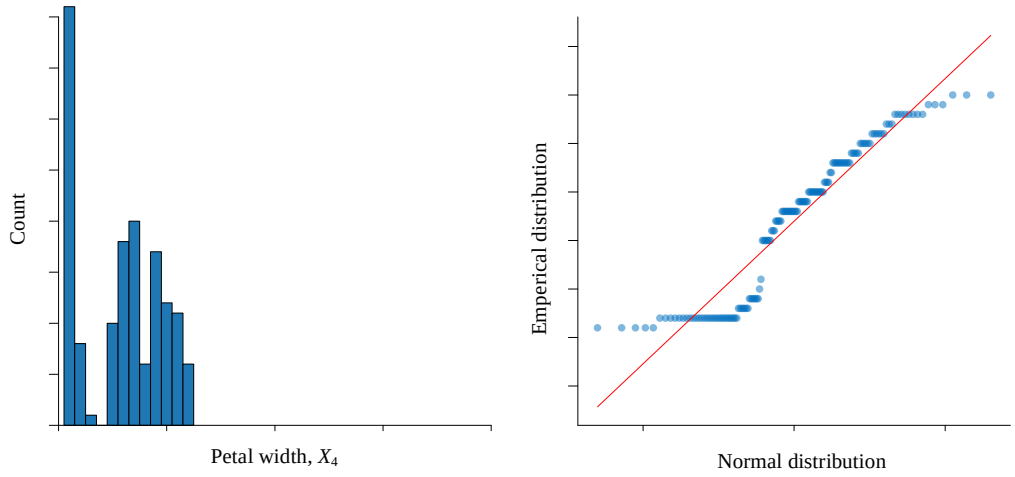

图 14. 花瓣宽度直方图和 QQ 图

3.5 箱型图：上界、下界之外样本

《统计至简》第 2 章专门介绍箱型图。

箱型图 (box plot) 是一种展示数据分布和离群值的方法。箱型图通过绘制数据的四分位数 (Q_1、Q_2、Q_3) 和可能的离群值来呈现数据的位置和离散程度。箱型图常用于探索性数据分析和统计推断，可用于比较不同组之间的数据分布和趋势。

图 15 所示为箱型图原理。Q_1 也叫下四分位，Q_2 也叫中位数，Q_3 也称上四分位。

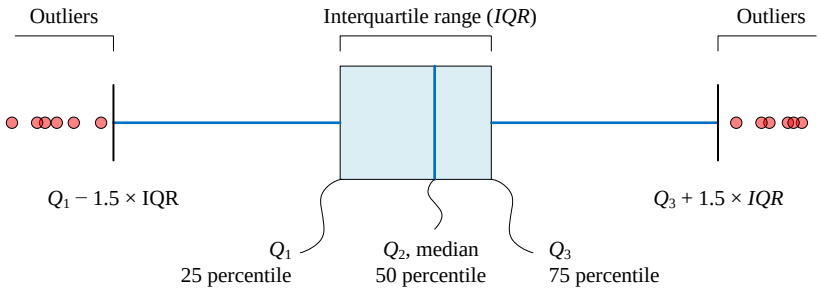

图 15. 箱型图原理

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。

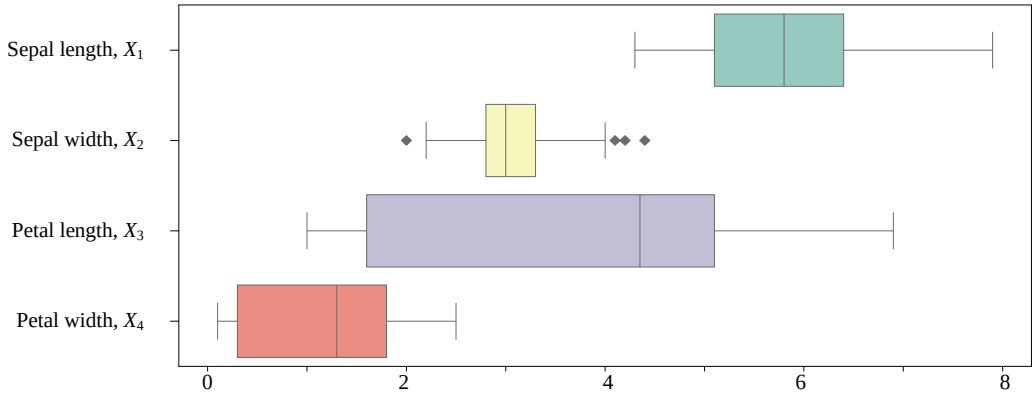

图 16. 鸢尾花箱型图

3.6 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：

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

⚠ 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±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±2s，这相当于 95%置信度，即 $\alpha = 0.05$。

➡ 图 17 展示了《统计至简》第 9 章介绍的 68–95–99.7 法则，请大家回顾。

⚠ 注意，图 17 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。

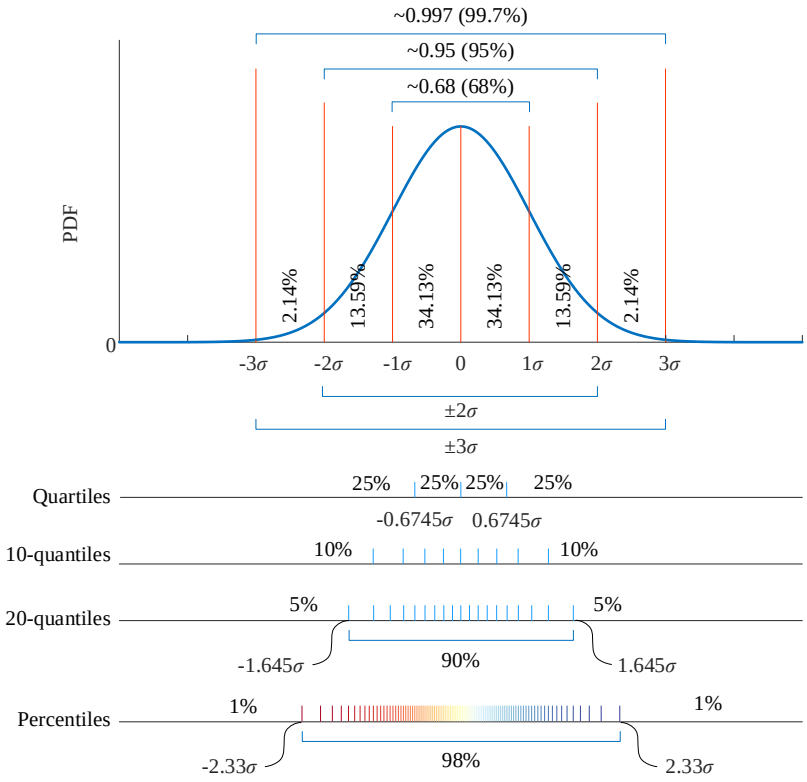

图 17. 标准差，注意图中并不区分总体标准差 σ 和样本标准差 s

Z 分数

Z 分数 (Z score) 是一种用于标准化数据的方法。Z 分数表示一个数据点距离均值的标准差数目，通常用于将不同尺度和分布的数据标准化为标准正态分布。Z 分数可以帮助我们比较不同数据点之间的相对位置和大小，判断数据是否偏离均值，并进行异常值检测和离群值分析。在实际应用中，Z 分数也经常用于构建模型、计算概率和决策阈值等任务。

从 Z 分数角度，(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 18 所示为鸢尾花数据四个特征的 Z 分数。

《统计至简》第 9 章还介绍过 Z 分数。

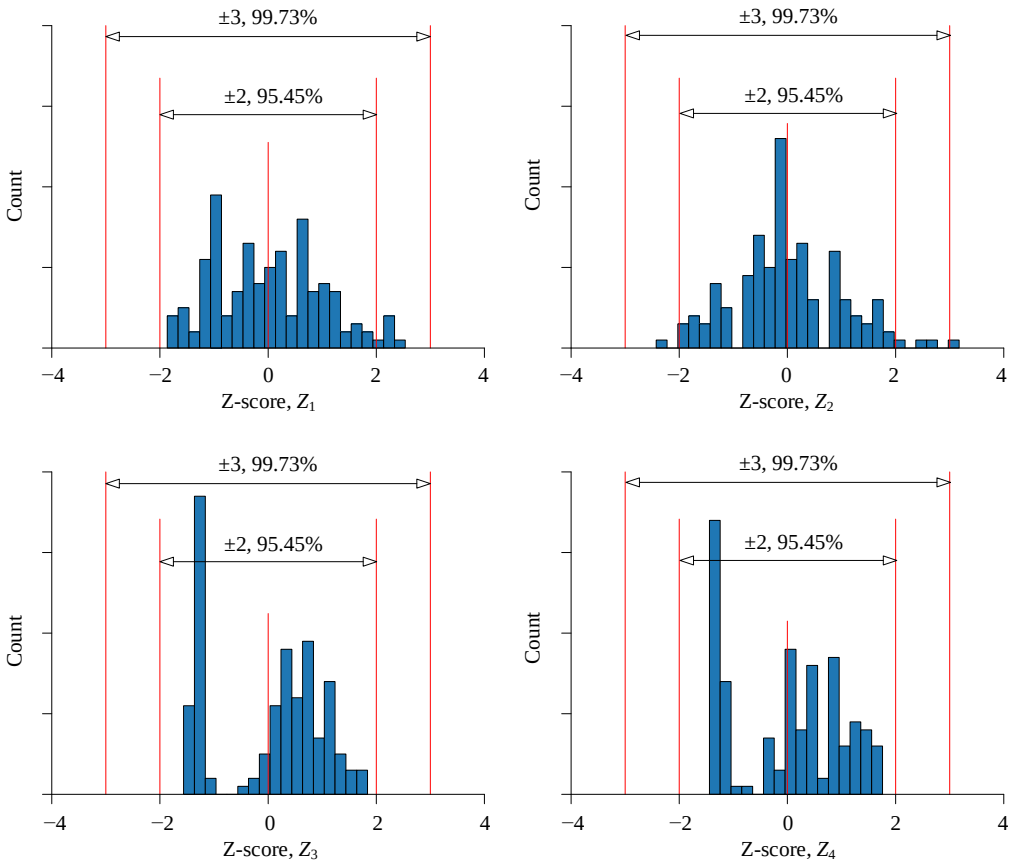

图 18. 鸢尾花 Z 分数

3.7 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)。马氏距离是一种考虑不同特征之间相关性的距离度量方法。马氏距离可以通过将样本点与数据集的均值向量进行比较，并考虑数据集的协方差矩阵来计算。与欧几里得距离不同，马氏距离可以捕捉不同特征之间的相关性和尺度差异，因此更适用于高维数据或特征相关的数据分析任务。马氏距离常用于聚类、分类、异常检测和模式识别等任务。

马氏距离定义如下：

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

其中，查询点 $\boldsymbol{q}$ 一般为数据质心，$\boldsymbol{\Sigma}$ 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 19 (a) 所示为，不同的马氏距离等高线对应不同的置信区间。图 19 (b) 而所示为$\pm\sigma \sim \pm 4\sigma$ 置信区间。

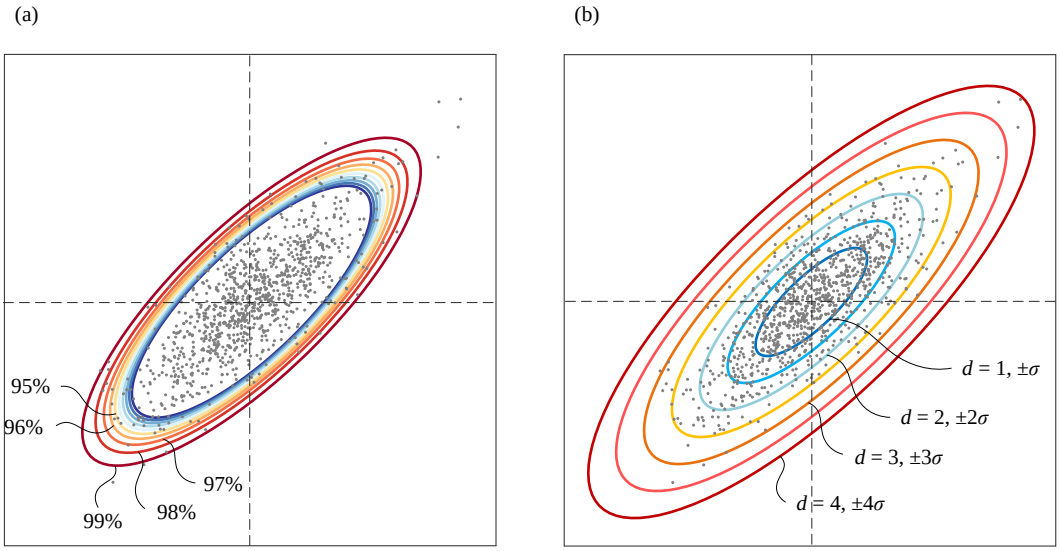

图 19. 协方差椭圆：(a) 95% ~ 99% 置信区间；(b) ±σ ~ ±4σ 置信区间

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 20 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。

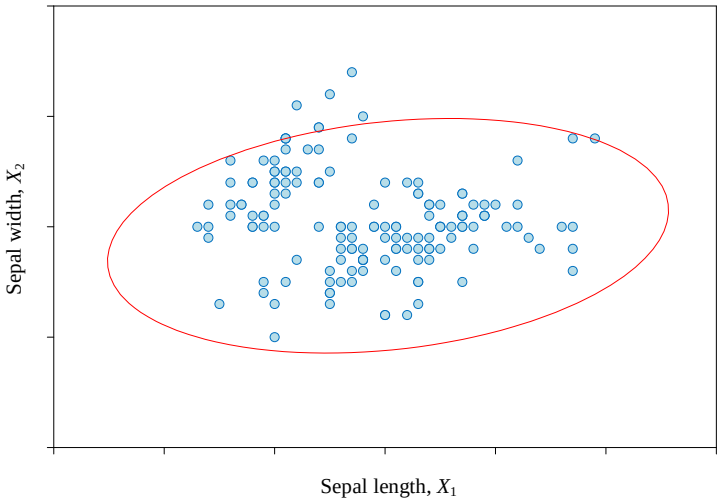

图 20. 鸢尾花数据前两个特征构造的协方差椭圆，马氏距离为 2

代码 Bk6_Ch03_01.py 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 21 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。

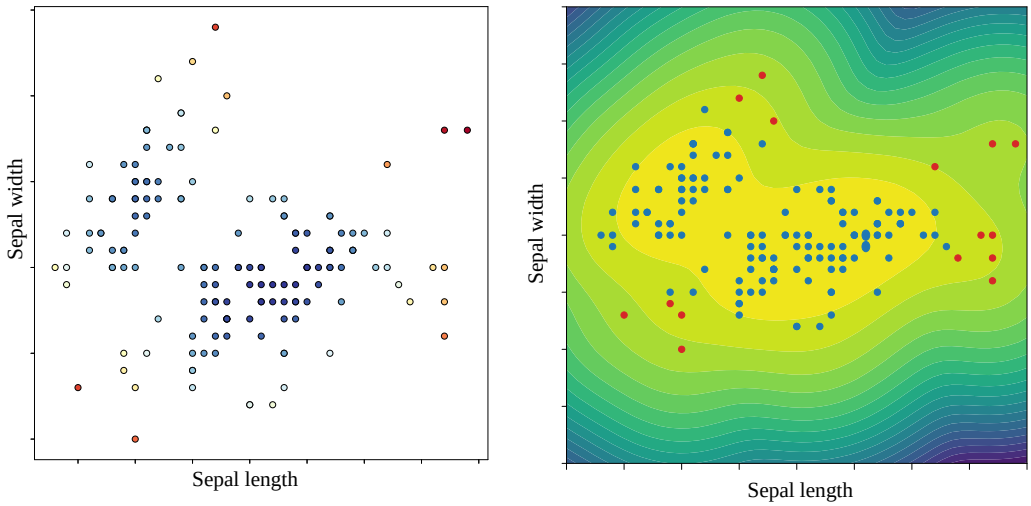

图 21. 概率密度估计判断离群值，左图散点颜色对应数据 KDE 概率密度估算值

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 22 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。

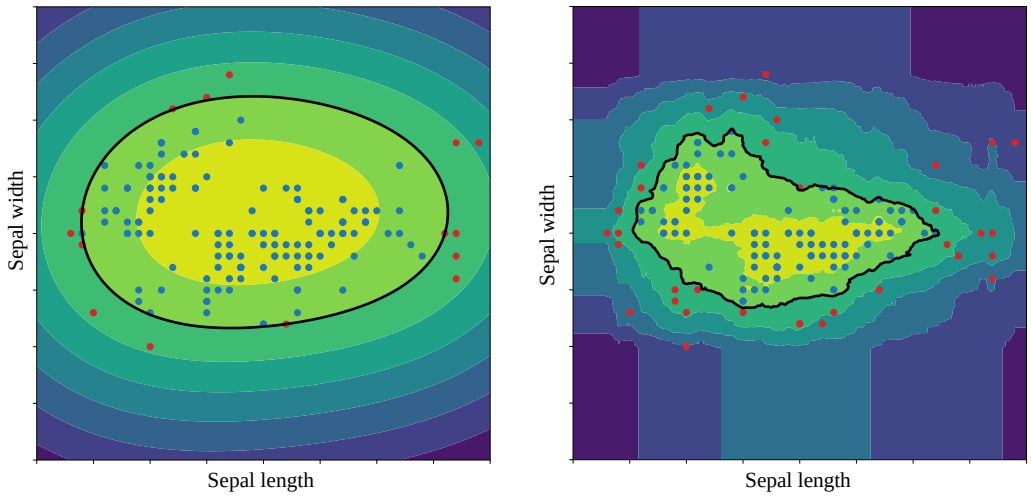

图 22. 支持向量机和孤立森林算法判定离群值

Bk6_Ch03_02.py 绘制图 21 和图 22。

离群值指的是数据集中与其他值相差较远的异常值。离群值可能会对数据分析结果产生较大的影响，导致模型不准确或偏差。离群值的产生原因包括测量误差、数据录入错误、采集异常、样本选择偏差等。

解决方法包括删除离群值、修正离群值、分别分析离群值等。注意事项包括要对数据进行探索性分析，了解数据分布和异常值的特点，合理处理离群值，避免对分析结果造成负面影响。同时，在进行离群值处理时需要谨慎，避免过度修正，影响数据的真实性和可靠性。

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。