23 马氏距离

Mahalanobis Distance

一种考虑数据分布的距离度量

> 耐心，坚持；今天的苦，就是明天的甜。
>
> ***Be patient and tough; someday this pain will be useful to you.***
>
> —— 奥维德 (Ovid) | 古罗马诗人 | 43 BC ~ 17/18 AD

- ◄ `numpy.linalg.eig()` 特征值分解
- ◄ `scipy.stats.distributions.chi2.cdf()` 卡方分布的 CDF
- ◄ `scipy.stats.distributions.chi2.ppf()` 卡方分布的百分点函数 PPF
- ◄ `seaborn.pairplot()` 成对散点图
- ◄ `seaborn.scatterplot()` 绘制散点图
- ◄ `sklearn.covariance.EmpiricalCovariance()` 估算协方差的对象，可以用来计算马氏距离

23.1 马氏距离：考虑数据分布的距离度量

本系列丛书的读者对马氏距离应该完全不陌生，本章将系统地讲解马氏距离及其应用。

定义

马氏距离 (Mahalanobis distance, Mahal distance)，也称马哈距离，具体定义如下：

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

其中，$\boldsymbol{\Sigma}$ 为样本数据 $\boldsymbol{X}$ 方差协方差矩阵，$\boldsymbol{\mu}$ 为 $\boldsymbol{X}$ 的质心。注意，马氏距离的单位为标准差。

从几何来讲，d 为定值时，(1) 为质心位于 $\boldsymbol{\mu}$ 的超椭圆 (hyper-ellipse)。

平移 → 旋转 → 缩放

对 $\boldsymbol{\Sigma}$ 特征值分解得到：

$$\boldsymbol{\Sigma}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

利用 (2) 获得 $\boldsymbol{\Sigma}^{-1}$ 的特征值分解：

$$\boldsymbol{\Sigma}^{-1}=\boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{\mathrm{T}} \tag{3}$$

将 (3) 代入 (1) 整理得到：

$$d=\left\|\underbrace{\boldsymbol{\Lambda}^{\frac{-1}{2}}}_{\text{Scale}}\underbrace{\boldsymbol{V}^{\mathrm{T}}}_{\text{Rotate}}\underbrace{\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}_{\text{Centralize}}\right\| \tag{4}$$

其中，$\boldsymbol{\mu}$ 完成中心化 (centralize)，$\boldsymbol{V}$ 矩阵完成旋转 (rotate)，$\boldsymbol{\Lambda}^{\frac{-1}{2}}$ 矩阵完成缩放 (scale)。整个几何变换过程如图 1 所示。对这部分内容感到陌生的读者，请参考本书第 11 章。大家如果忘记特征值分解相关内容，请回顾《矩阵力量》第 13、14 章。

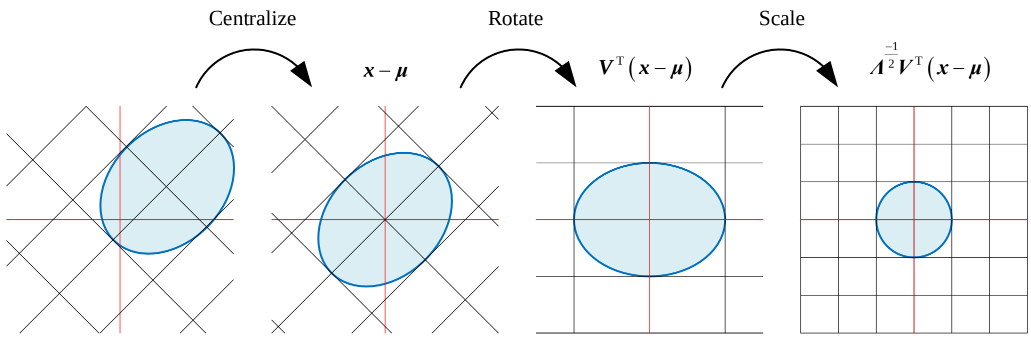

图 1. 几何变换：平移 → 旋转 → 缩放

马氏距离将协方差矩阵 $\boldsymbol{\Sigma}$ 纳入距离度量计算。马氏距离相当于对欧氏距离的一种修正，马氏距离完成数据正交化 (orthogonalization)，解决特征之间相关性问题。同时，马氏距离内含标准化 (standardization)，解决了特征之间尺度和单位不一致问题。

单特征

特别地，当特征数 $D = 1$ 时：

$$\boldsymbol{x} = [x], \quad \boldsymbol{\mu} = [\mu], \quad \boldsymbol{\Sigma} = [\sigma^2] \tag{5}$$

代入 (1) 得到：

$$d = \sqrt{(x-\mu)\frac{1}{\sigma^2}(x-\mu)} = \left|\frac{x-\mu}{\sigma}\right| \tag{6}$$

大家是不是觉得眼前一亮，这正是 z 分数的绝对值，d 的单位正是标准差。如图 2 (a) 所示，比如 $d = 3$，意味着马氏距离为“3 个标准差”。

当特征数 $D = 2$ 时，如图 2 (b) 所示，马氏距离的几何形态是同心椭圆。当特征数 $D = 3$ 时，如图 2 (c) 所示，马氏距离的几何形态是同心椭球。

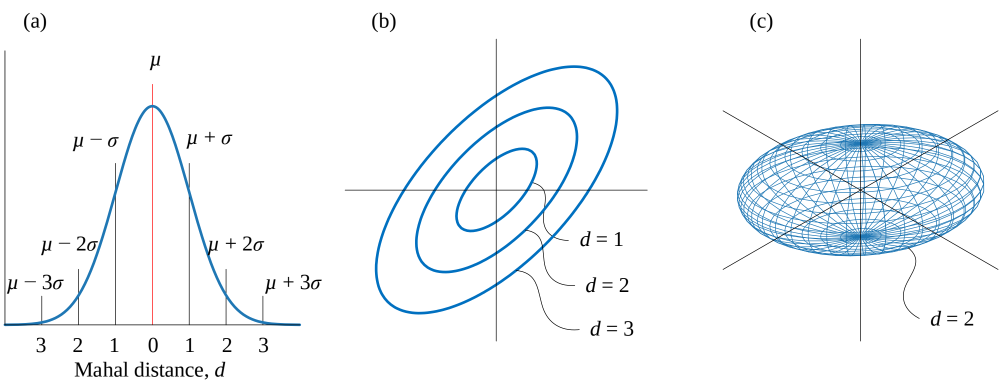

图 2. 马氏距离的几何形态

本章后文先比较三种常见距离：1) 欧氏距离；2) 标准化欧氏距离；3) 欧氏距离。

23.2 欧氏距离：最基本的距离

欧几里得距离 (Euclidean distance)，也称欧氏距离，是最“自然”的距离，是多维空间中两个点之间的绝对距离度量。

欧氏距离

$\boldsymbol{x}$ 和质心 $\boldsymbol{\mu}$ 的欧氏距离定义为：

$$d = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})} = \|\boldsymbol{x}-\boldsymbol{\mu}\| \tag{7}$$

欧氏距离本质上是 L^2 范数。

以鸢尾花花萼长度和花瓣长度两个特征数据为例，数据质心所在位置为：

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \tag{8}$$

注意，上式的两个特征单位为厘米。

如图 3 所示，平面上任意一点 $\boldsymbol{x}$ 到质心 $\boldsymbol{\mu}$ 的欧氏距离的解析式为：

$$\begin{aligned} d &= \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})} = \sqrt{\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix}\right)^{\mathrm{T}} \left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix}\right)} \\ &= \sqrt{(x_1 - 5.843)^2 + (x_3 - 3.758)^2} \end{aligned} \tag{9}$$

图 3 所示的三个同心圆距离质心 $\boldsymbol{\mu}$ 距离为 1 cm、2 cm、3 cm。此外，请大家注意图 4 中的网格，这个网格每个格子“方方正正”，边长都是 1 cm。

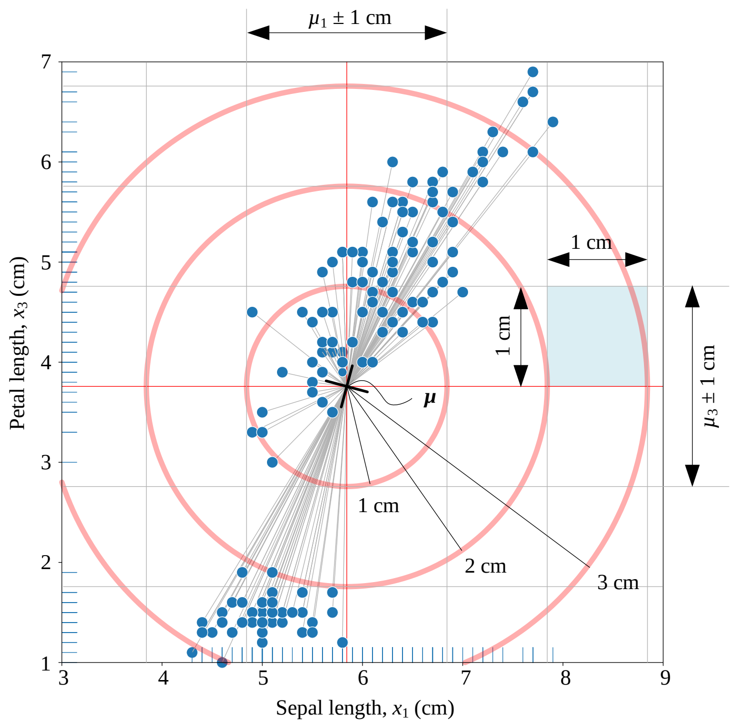

图 3. 花萼长度、花瓣长度平面上的欧氏距离等高线和网格

23.3 标准化欧氏距离：两个视角

第一视角：正椭圆

标准化欧氏距离 (standardized Euclidean distance) 定义如下：

$$d = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{D}^{-1}\boldsymbol{D}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{10}$$

其中，$\boldsymbol{D}$ 为对角方阵，对角线元素为标准差，运算如下：

$$\boldsymbol{D} = \operatorname{diag}\left(\operatorname{diag}(\boldsymbol{\Sigma})\right)^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_D \end{bmatrix} \tag{11}$$

特别地，当 $D = 2$ 时，标准化欧氏距离为：

$$d = \sqrt{\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}} = \sqrt{z_1^2 + z_2^2} \tag{12}$$

其中，z_1 和 z_2 是两个特征的 z 分数。可以说，z_1 的单位是 σ_1，z_2 的单位是 σ_2。

如图 3 所示，x_1x_3 平面上任意一点 $\boldsymbol{x}$ 到质心 $\boldsymbol{\mu}$ 的标准化欧氏距离为：

$$d = \sqrt{\frac{(x_1-5.843)^2}{0.685} + \frac{(x_3-3.758)^2}{3.116}} \tag{13}$$

上式中，鸢尾花花萼长度数据的方差为 0.685 cm^2，标准差 σ_1 为 0.827 cm。花瓣长度数据的方差为 3.116 cm^2，标准差 σ_3 为 1.765 cm。

图 4 所示为在花萼长度、花瓣长度平面上标准化欧氏距离为 1、2、3 的三个正椭圆。1、2、3 的单位可以理解为标准差。

大家注意图 4 中网格，网格的格子为矩形。矩形的宽度为 σ_1 = 0.827 cm，矩形的长度为 σ_3 = 1.765 cm。

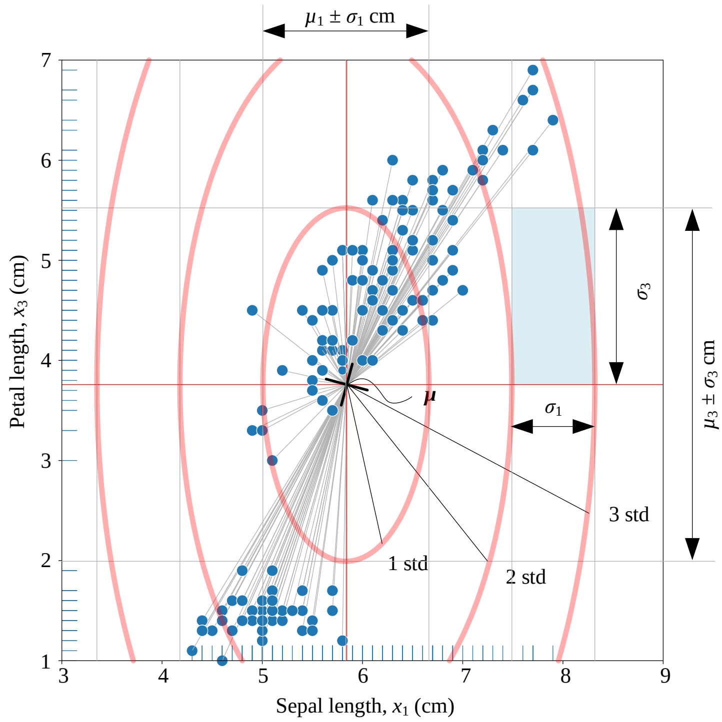

图 4. 花萼长度、花瓣长度平面上的标准化欧氏距离和网格

第二视角：正圆

先计算花萼长度、花瓣长度的 z 分数 z_1、z_3：

$$z_1 = \frac{x_1 - 5.843}{0.827}, \quad z_3 = \frac{x_3 - 3.758}{1.765} \tag{14}$$

几何视角，上式经过了中心化、缩放两步。

然后再计算标准化欧氏距离：

$$d = \sqrt{z_1^2 + z_3^2} \tag{15}$$

图 5 所示花萼长度 z 分数、花瓣长度 z 分数平面上的标准化欧氏距离等高线。不难发现，在这个平面上，等高线为正圆，圆心位于原点。

图 5 中网格为正方形，这是因为数据已经标准化。

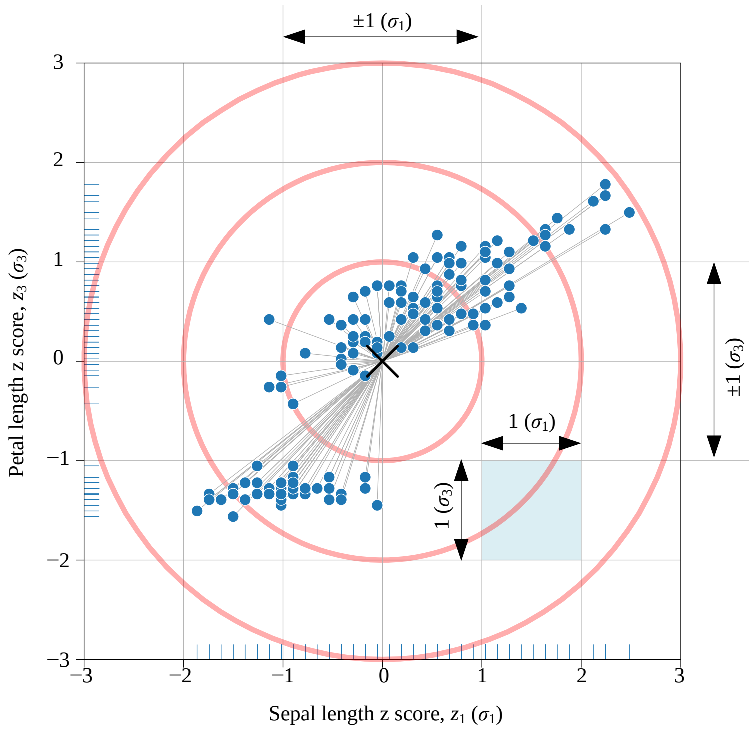

图 5. 花萼长度 z 分数、花瓣长度 z 分数平面上的标准化欧氏距离

23.2 马氏距离：两个视角

旋转椭圆

鸢尾花花萼长度、花瓣长度协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma}=\begin{bmatrix}0.685 & 1.274\\ 1.274 & 3.116\end{bmatrix} \tag{16}$$

协方差 $\boldsymbol{\Sigma}$ 的逆为：

$$\boldsymbol{\Sigma}^{-1}=\begin{bmatrix}6.075 & -2.484\\ -2.484 & 1.336\end{bmatrix} \tag{17}$$

代入 (1)，得到马氏距离的解析式：

$$\begin{aligned} d &= \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\begin{bmatrix}6.075 & -2.484\\ -2.484 & 1.336\end{bmatrix}(\boldsymbol{x}-\boldsymbol{\mu})} \\ &= \sqrt{\left(\begin{bmatrix}x_1\\ x_3\end{bmatrix}-\begin{bmatrix}5.843\\ 3.758\end{bmatrix}\right)^{\mathrm{T}}\begin{bmatrix}6.075 & -2.484\\ -2.484 & 1.336\end{bmatrix}\left(\begin{bmatrix}x_1\\ x_3\end{bmatrix}-\begin{bmatrix}5.843\\ 3.758\end{bmatrix}\right)} \\ &= \sqrt{6.08x_1^2-4.97x_1x_3+1.34x_3^2-52.32x_1+18.99x_3+117.21} \end{aligned} \tag{18}$$

图 6 中三个椭圆分别代表马氏距离为 1、2、3。这个旋转椭圆的长轴就是第 25 章要介绍的第一主成分 (first principal component) 方向，而旋转椭圆的短轴就是第二主成分 (second principal component) 方向。

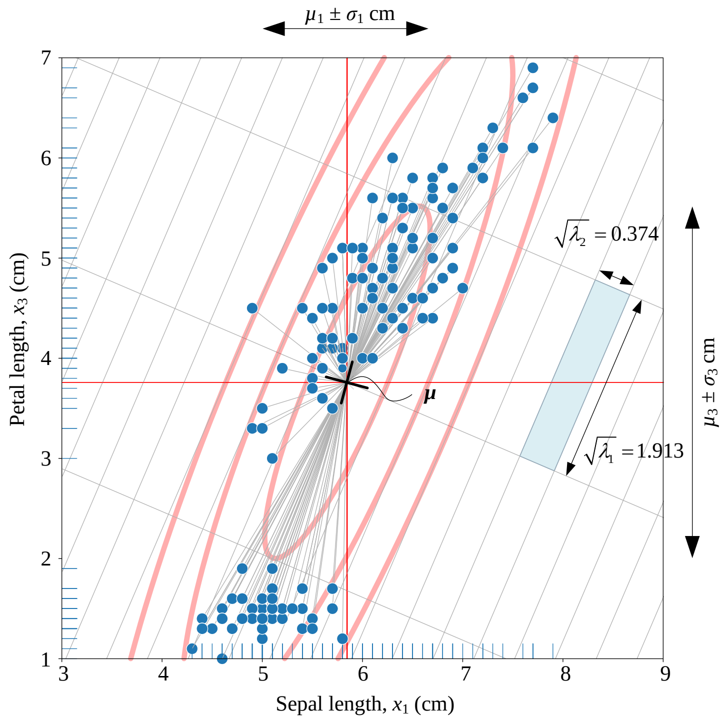

图 6. 花萼长度、花瓣长度平面上的马氏距离等高线和网格

对协方差矩阵特征值分解得到的特征值方阵为：

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} = \begin{bmatrix} 3.661 & \\ & 0.140 \end{bmatrix} \tag{19}$$

两个特征值实际上就是数据投影在第一、第二主成分方向上的结果的方差，也叫主成分方差。上式的单位也都是平方厘米 cm^2。

而这两个特征值的平方根就是主成分标准差：

$$\sqrt{\lambda_1} = 1.913 \text{ cm}, \quad \sqrt{\lambda_2} = 0.374 \text{ cm} \tag{20}$$

它俩分别是旋转椭圆的半长轴、半短轴长度。

如图 6 所示，图中的网格就是度量马氏距离的坐标系。网格矩形倾斜角度和主成分方向相同。矩形的长度为 $\sqrt{\lambda_1}$，宽度为 $\sqrt{\lambda_2}$。

第二视角：正圆

令：

$$z = \underbrace{\Lambda^{\frac{-1}{2}}}_{\text{Scale}} \underbrace{V^{\mathrm{T}}}_{\text{Rotate}} \underbrace{\left(x - \mu\right)}_{\text{Centralize}} \tag{21}$$

将上式代入 (1)，得到马氏距离为 z 的 L^2 范数：

$$d = \sqrt{z^{\mathrm{T}} z} = \|z\| \tag{22}$$

如图 7 所示，在第一、第二主成分平面上，马氏距离为正圆。

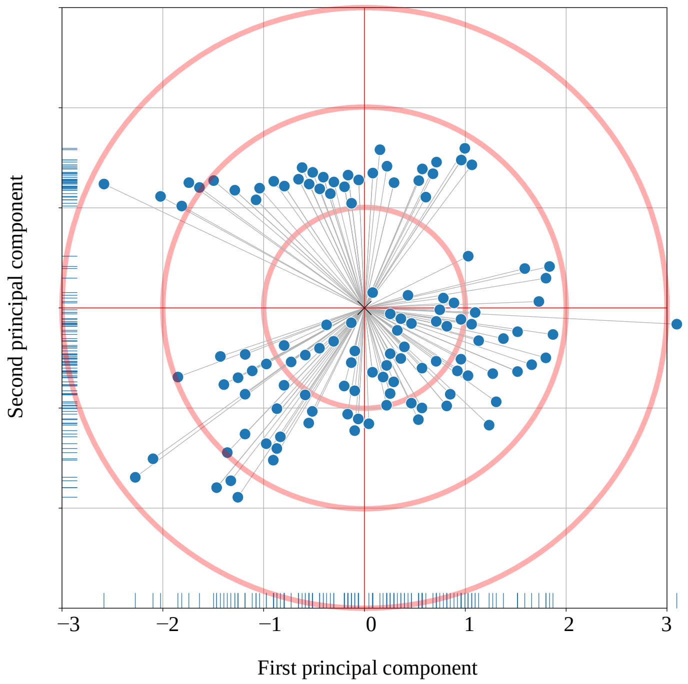

图 7. 第一、第二主成分平面上马氏距离等高线和网格

Bk5_Ch23_01.py 绘制图 3、图 4、图 6。

成对特征图

马氏距离椭圆也可以画在成对特征图上。图 8 和图 9 分别展示考虑不考虑标签和考虑标签的马氏距离椭圆。这些图像可以帮助我们分析理解数据，比如解读相关性、发现离群值等。《数据有道》一册将专门讲解如何发现离群值。

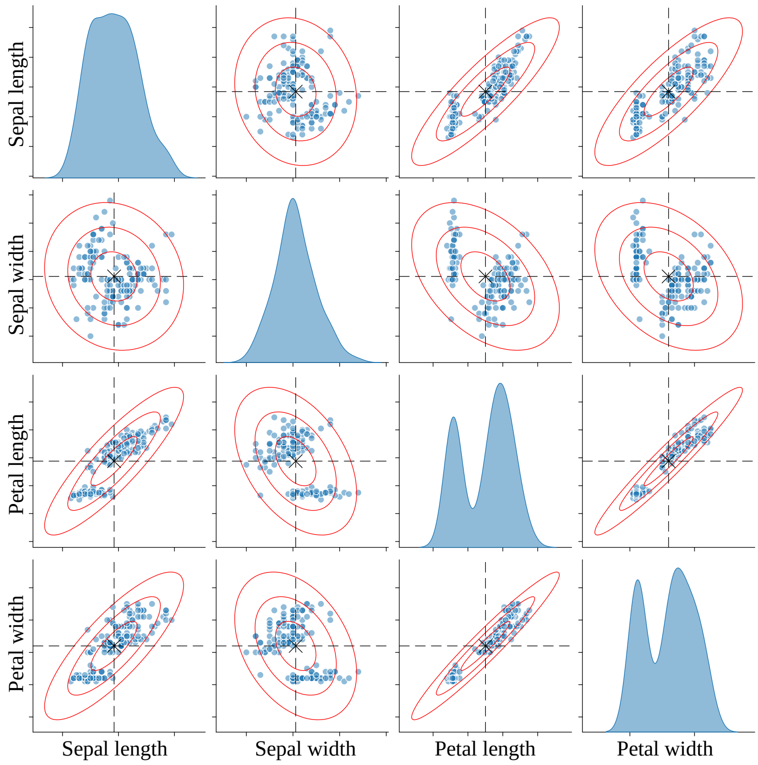

图 8. 成对特征图上绘制马氏距离等高线，不考虑标签

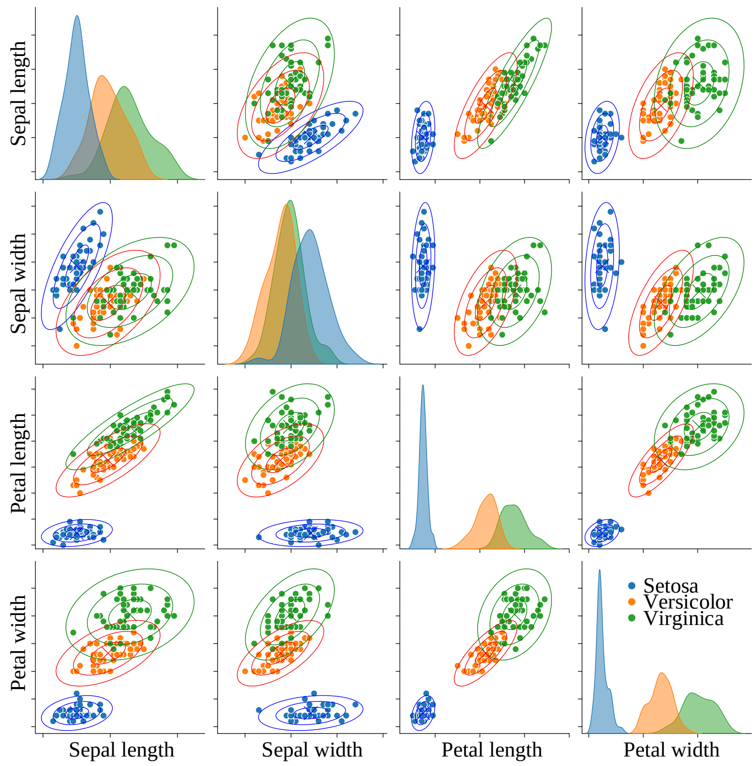

图 9. 成对特征图上绘制马氏距离等高线，考虑标签

Bk5_Ch23_02.py 绘制图 8 和图 9。

23.3 马氏距离和“68-95-99.7 法则”

本书第 9 章介绍过一元高斯分布的“68-95-99.7 法则”。这个法则具体是指，如果数据近似服从一元高斯分布 $N(\mu, \sigma)$，则约 68.3%、95.4%和 99.7%的数据分布在距均值 (μ) 1 个 ($\mu \pm \sigma$)、2 个 ($\mu \pm 2\sigma$) 和 3 个 ($\mu \pm 3\sigma$) 正负标准差范围之内。

而 68.3%、95.4%和 99.7%这三个数实际上卡方分布直接相关。当 $D = 1$ 时，X_1 服从正态分布 $N(\mu_1, \sigma_1)$，经过标准化得到的随机变量 Z_1 则服从标准正态分布：

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \sim N(0,1) \tag{23}$$

也就是说，Z_1 的平方服从自由度为 1 的卡方分布：

$$Z_1^2 \sim \chi^2_{(\mathrm{df}=1)} \tag{24}$$

注意，实际上 Z_1 的平方再开方，即 Z_1 的绝对值就是马氏距离。

$D = 2$ 时，马氏距离平方 d^2 服从 df = 2 的卡方分布：

$$d^2 \sim \chi^2_{(\mathrm{df}=2)} \tag{25}$$

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

$$d^2 = (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_{(\mathrm{df}=D)} \tag{26}$$

也就是说，距离为 d 的马氏距离超椭圆围成的几何图形内部的概率 α 可以用卡方分布 CDF 查表获得。比如，Scipy 中卡方分布的对象为 scipy.stats.distributions.chi2，计算 $D = 2$，马氏距离 d = 3 条件下，马氏距离椭圆围成的图形的概率 α 为 scipy.stats.distributions.chi2.cdf(d^2 = 9, df = 2)。

相反，如果给定概率值 α 和自由度，可以用卡方分布的百分点函数 PPF，即 CDF 的逆函数 (inverse CDF)，反求马氏距离的平方 d^2。这个值开方就是马氏距离 d。

比如，给定概率值 0.9，自由度为 2，利用 scipy.stats.distributions.chi2.ppf(0.9, df=2) 可以求得马氏距离的平方值 d^2，开方就是马氏距离 d。

如图 10 (a) 所示，自由度为 2，给定一系列概率值 (0.90 ~ 0.99)，利用卡方分布的百分点函数 PPF，我们便获得一系列马氏距离椭圆。图 10 (b) 对照马氏距离取值为 1 ~ 5。

这些椭圆中，马氏距离 3 几乎对应 99%这个概率值。也就是说，如果二元随机数近似服从二元高斯分布，约有 99%的随机数落在马氏距离为 3 的椭圆内。

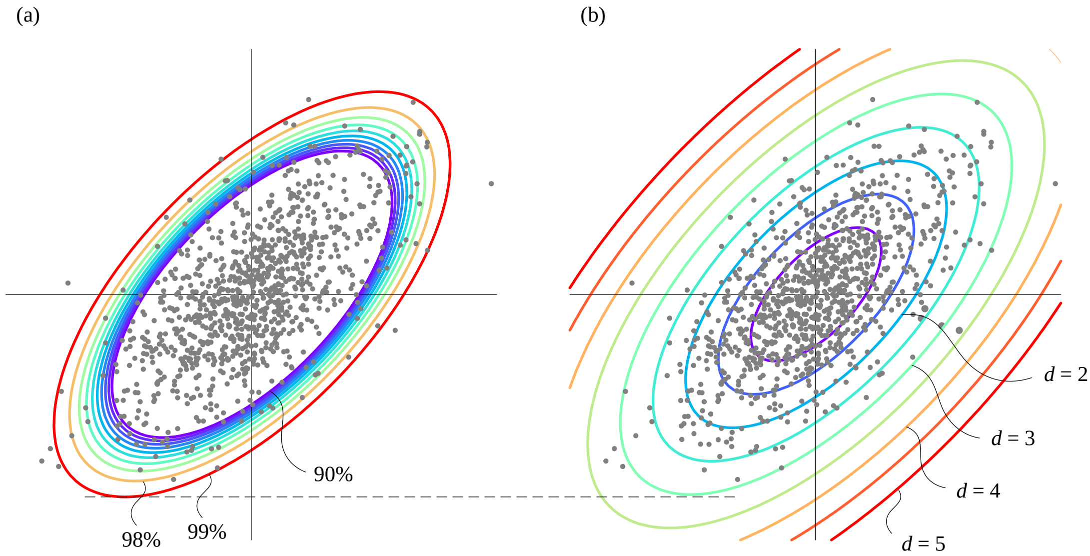

图 10 特征数 $D = 2$ 时，概率值 α 和马氏距离椭圆位置

Bk5_Ch23_03.py 绘制图 10。

图 11 所示为马氏距离 d、自由度 df、概率值 α 三者关系曲线。

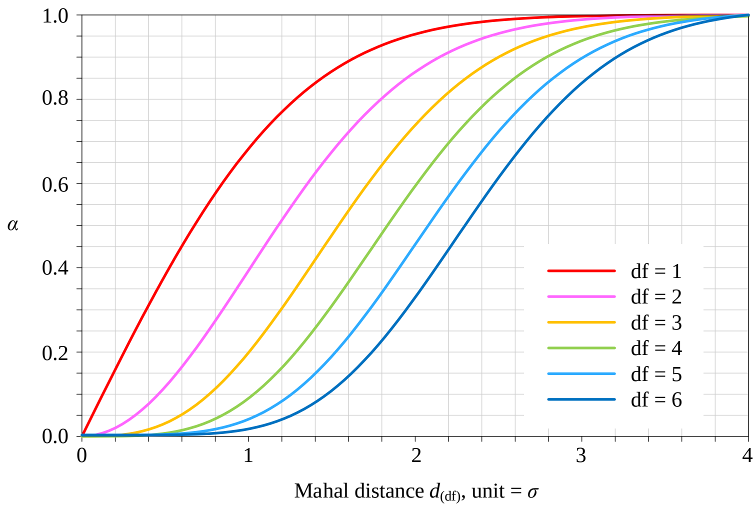

图 11 马氏距离 d、自由度 df、概率值 α 三者关系

为了方便查表，大家可以参考图 12 和图 13。图 12 中，给定马氏距离 d、自由度 df，查表得到 α。这张表中，我们可以看到一元高斯分布的 68-95-99.7 法则。

而自由度 df = 2 时，这个法则变为马氏距离为 1、2、3 的椭圆对应 39%、86%、98.9%，我们也可以管它叫 39-86-98.9 法则。

图 13 中，给定概率值 α、自由度 df，查表得到马氏距离 d。

Degree of freedom, df	Mahal distance, d: 1	1.25	1.5	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4
1	0.6827	0.7887	0.8664	0.9199	0.9545	0.9756	0.9876	0.9940	0.9973	0.9988	0.9995	0.9998	0.9999
2	0.3935	0.5422	0.6753	0.7837	0.8647	0.9204	0.9561	0.9772	0.9889	0.9949	0.9978	0.9991	0.9997
3	0.1987	0.3321	0.4778	0.6179	0.7385	0.8327	0.8999	0.9440	0.9707	0.9857	0.9934	0.9972	0.9989
4	0.0902	0.1845	0.3101	0.4526	0.5940	0.7191	0.8188	0.8910	0.9389	0.9681	0.9844	0.9929	0.9970
5	0.0374	0.0943	0.1864	0.3096	0.4506	0.5917	0.7174	0.8179	0.8909	0.9392	0.9685	0.9848	0.9932
6	0.0144	0.0448	0.1047	0.1990	0.3233	0.4642	0.6042	0.7281	0.8264	0.8971	0.9434	0.9711	0.9862

图 12. 给定马氏距离 d、自由度 df，查表得到概率值 α

Degree of freedom, df	Probability α that the random value will fall inside the ellipsoid: 0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.993	0.996	0.999
1	1.6449	1.6954	1.7507	1.8119	1.8808	1.9600	2.0537	2.1701	2.3263	2.5758	2.6968	2.8782	3.2905
2	2.1460	2.1945	2.2475	2.3062	2.3721	2.4477	2.5373	2.6482	2.7971	3.0349	3.1502	3.3231	3.7169
3	2.5003	2.5478	2.5997	2.6571	2.7216	2.7955	2.8829	2.9912	3.1365	3.3682	3.4806	3.6492	4.0331
4	2.7892	2.8361	2.8873	2.9439	3.0074	3.0802	3.1663	3.2729	3.4158	3.6437	3.7542	3.9199	4.2973
5	3.0391	3.0856	3.1363	3.1923	3.2552	3.3272	3.4124	3.5178	3.6590	3.8841	3.9932	4.1568	4.5293
6	3.2626	3.3088	3.3591	3.4147	3.4770	3.5485	3.6329	3.7373	3.8773	4.1002	4.2083	4.3702	4.7390

图 13. 给定概率值 α、自由度 df，查表得到马氏距离 d

Bk5_Ch23_04.py 绘制图 11。

用卡方分布将马氏距离转换为概率时，有些文献错误地将自由度给定为 $D-1$，即特征数 D 减 1。下面这篇文章详尽地解释如何正确使用马氏距离，建议大家参考。

https://peerj.com/articles/6678/