

17

Probability Density Estimation

概率密度估计

若干概率密度函数加权叠合



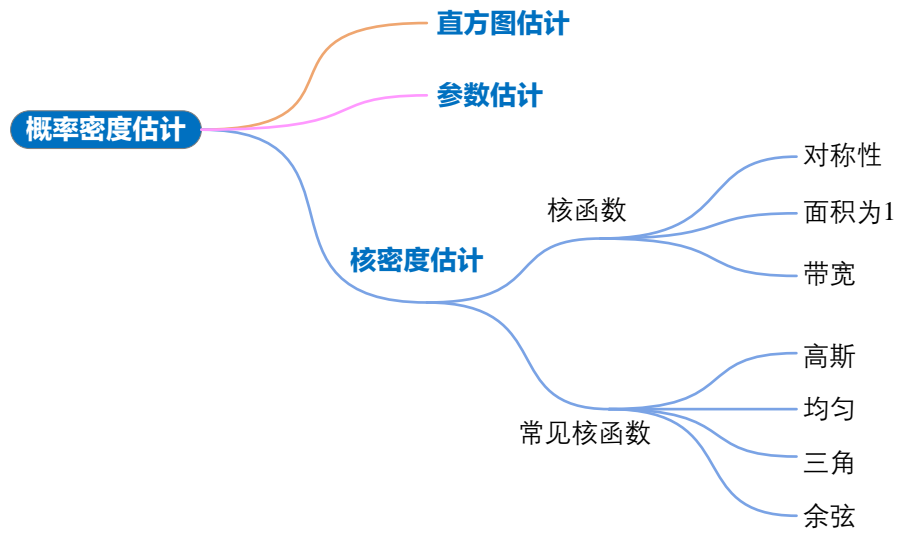
大自然是一个无限的球体，其中心无处不在，圆周无处可寻。

Nature is an infinite sphere of which the center is everywhere and the circumference nowhere.

—— 布莱兹·帕斯卡 (Blaise Pascal) | 法国哲学家、科学家 | 1623 ~ 1662



- ▶ `matplotlib.pyplot.fill_between()` 区域填充颜色
- ▶ `seaborn.kdeplot()` 绘制 KDE 概率密度估计曲线
- ▶ `sklearn.neighbors.KernelDensity()` 概率密度估计函数
- ▶ `statsmodels.api.nonparametric.KDEUnivariate()` 构造一元 KDE
- ▶ `statsmodels.nonparametric.kde.kernel_switch()` 更换核函数
- ▶ `statsmodels.nonparametric.kernel_density.KDEMultivariate()` 构造多元 KDE



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

17.1 概率密度估计：从直方图说起

简单来说，**概率密度估计** (probability density estimation) 就是寻找合适的随机变量概率密度函数，使其尽量贴合样本数据分布情况。

直方图

直方图实际上是最常用的一种概率密度估计方法。本书第 2 章介绍过，为了构造直方图，首先将样本数据的取值范围分为一系列左右相连等宽度的**组** (bin)，然后统计每个组内样本数据的频数。绘制直方图时，以组距为底边、以频数为高度，绘制一系列矩形图。

图 1 所示为鸢尾花四个特征上样本数据的频数直方图。合理地选择组距，让大家一眼能够通过直方图看出样本分布的大致情况。纵轴的频数，也可以替换成概率、概率密度。当纵轴为概率密度时，直方图这些矩形面积之和为 1，对应概率 1。

但是，直方图的缺点也很明显，概率密度估计结果呈现阶梯状，不“平滑”。很多数据科学、机器学习应用场合，我们需要得到连续平滑的密度估计曲线。

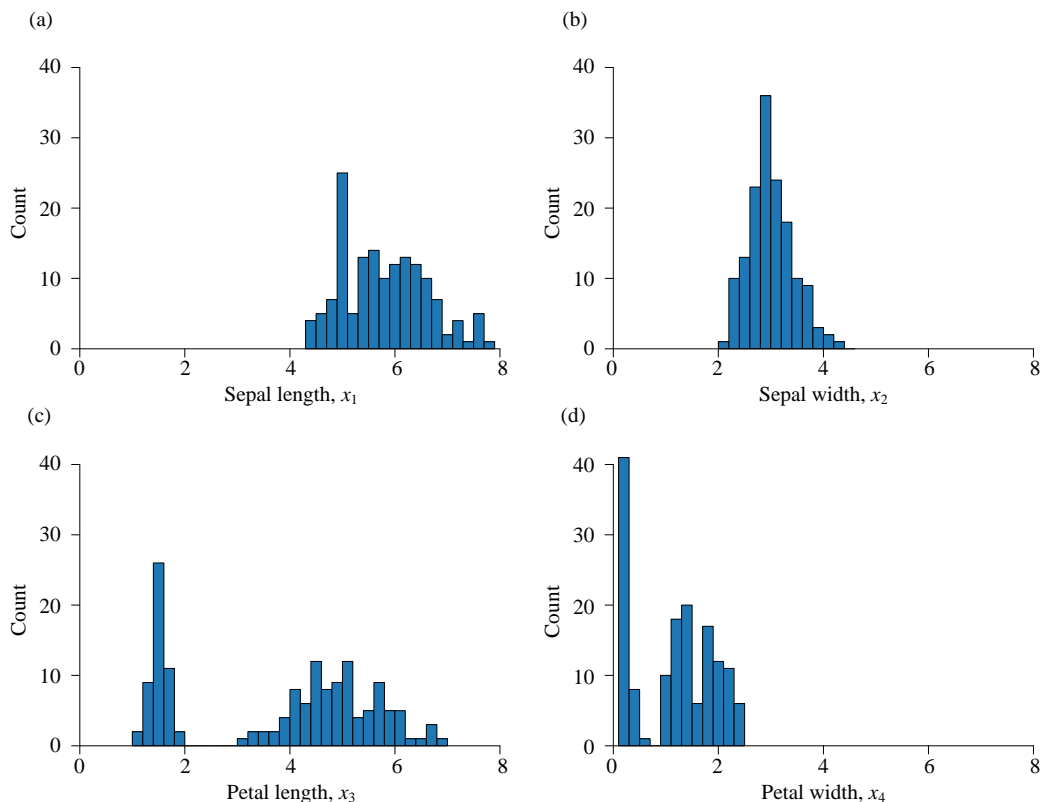


图 1. 鸢尾花四个特征的直方图，纵轴为频数

参数估计

本书前文介绍一些常见的概率分布函数，但是它们的形状远远不够描述现实世界采集的分布情况较为复杂的样本数据。

以高斯分布为例，我们可以很容易计算得到样本数据的均值 μ 和均方差 σ ，这样可以直接用正态分布来估计样本数据在某个单一特征上的分布情况：

$$\hat{f}_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (1)$$

估计概率密度时，直接利用均值 μ 和均方差 σ 这两个参数，因此这种方法也被称作参数估计。如图 2 所示，高斯分布显然比图 1 的直方图“平滑”的多。

这种方法的缺陷是显而易见的，对比图 1 和图 2，容易发现样本分布细节被忽略，最明显的是鸢尾花花瓣长度（比较图 1 (c)、图 2 (c)）、花瓣宽度（比较图 1 (d)、图 2 (d)）这两个特征上样本数据的分布。多数情况，样本数据分布不够“正态”，仅仅使用均值 μ 和均方差 σ 描述数据不合适。

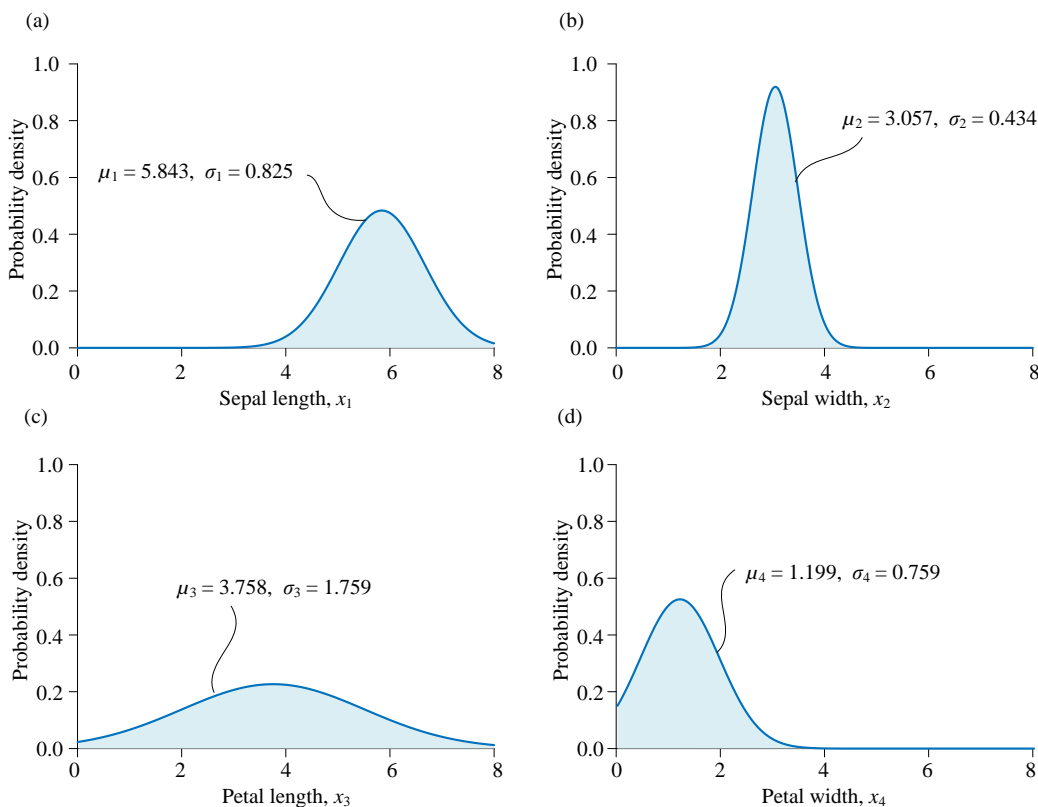


图 2. 用一元高斯分布估计鸢尾花四个特征的概率密度曲线

核密度估计

下面介绍本章的主角——**核密度估计** (Kernel Density Estimation, KDE)。本书前文很多场合已经用过核密度估计，比如第 2、5 章中都用高斯核密度估计过鸢尾花单一特征概率密度，以及联合概率密度。

核密度估计需要指定一个核函数来描述每一个数据点，最常见的核函数是高斯核函数，本章还会介绍并比较其他核函数。

图 3 所示为通过高斯核函数核密度估计得到的平滑曲线，下面我们聊一聊核密度估计原理。

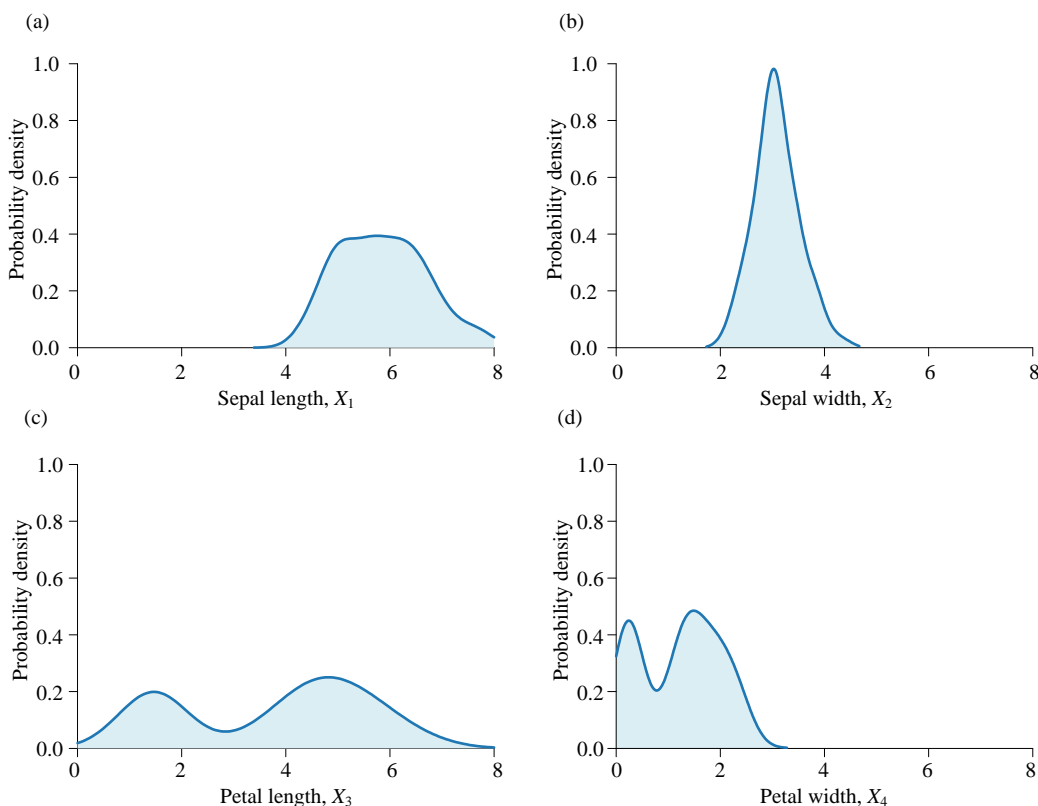


图 3. 鸢尾花四个特征的高斯 KDE 曲线



Bk5_Ch17_01.py 代码绘制图 3。代码使用 `seaborn.kdeplot()` 绘制 KDE 曲线。本章后续分别介绍几种不同的办法绘制 KDE 曲线。

17.2 核密度估计：若干核函数加权叠合

核密度估计其实是对直方图的一个自然拓展。直方图不够平滑，我们引入合适的核函数得到更加平滑的概率密度估计曲线。前文说到，核函数种类很多，本节以高斯核函数为例介绍核密度估计原理。

原理

任意一个数据点 $x^{(i)}$ ，都可以用一个函数来描述，这个函数就是核函数。如图 4 所示，一共有 7 个样本点，每一个样本点都用一个高斯核函数描述。白话说，图 4 中这 7 条曲线等权重叠加便得到核密度估计概率密度曲线。

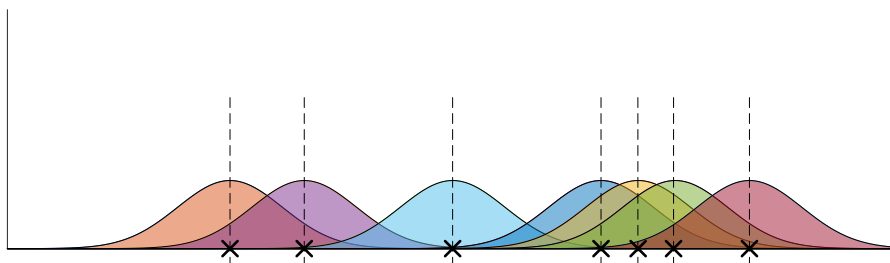


图 4. 用多个核函数描述样本数据

叠加 → 平均

而对于 n 个样本数据点 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ，我们可以用 n 个核函数分别代表每个数据点：

$$\underbrace{\frac{1}{h} K \left(\frac{x - x^{(i)}}{h} \right)}_{\text{Area} = 1}, \quad -\infty < x < +\infty \quad (2)$$

Shift
Scale

其中， h ($h > 0$) 是核函数本身的缩放系数，又叫带宽。每个核函数和水平面构成图形的面积为 1。

这 n 个核函数先叠加，然后再平均，便得到概率密度估计函数：

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x^{(i)}) = \underbrace{\frac{1}{n} \frac{1}{h} \sum_{i=1}^n K \left(\frac{x - x^{(i)}}{h} \right)}_{\text{Weight} \quad \text{Area} = n}, \quad -\infty < x < +\infty \quad (3)$$

上式中， $1/n$ 让 n 个面积为 1 的函数面积归一化。也就是说，每个核函数贡献的面积为 $1/n$ 。

高斯核函数

下面我们以高斯核函数为例，聊聊如何理解 (2)。

高斯核函数 $K(x)$ 的定义：

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (4)$$

上述高斯核函数显然和横轴围成的面积为 1。

对称性

核函数要求具有对称性，即：

$$K(x) = K(-x) \quad (5)$$

显然，(4) 定义的高斯核函数满足对称性。

而 (2) 中 $x - x^{(i)}$ 代表曲线在水平方向平移。由于核函数 $K(x)$ 关于纵轴对称，因此 $K(x - x^{(i)})$ 关于 $x = x^{(i)}$ 对称。

缩放

(2) 中的带宽 h 则代表图像在水平方向的缩放。大家是否还记得图 5？这两幅图来自《数学要素》第 12 章。我们在讲解函数图像变换时提过，原函数 $f(x)$ 和 $cf(cx)$ 面积相同，其中 $c > 0$ 。

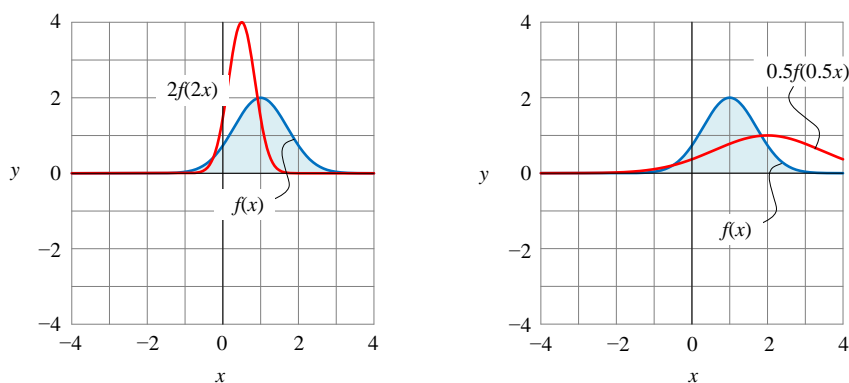


图 5. 原函数 $y = f(x)$ 水平方向、竖直方向伸缩，图片来自《数学要素》第 12 章

面积为 1

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$K(x)$ 的重要性质之一是面积为 1，也就是 $K(x)$ 对 x 在 $(-\infty, +\infty)$ 积分为 1：

$$\int_{-\infty}^{+\infty} K(x) dx = 1 \quad (6)$$

(4) 中高斯核函数显然满足这一条件。

利用换元积分，很容易得到如下等式：

$$\int_{-\infty}^{+\infty} K(x) dx = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \quad (7)$$

上式解释了为什么 $f(x)$ 和 $cf(cx)$ 面积相同。

可视化“叠加”

以图 4 为例，假设 7 个样本数据构成的集合为 $\{-3, -2, 0, 2, 2.5, 3, 4\}$ 。

如果 $h = 1$ ，参考 (3)，可用高斯核函数构造概率密度估计函数：

$$\hat{f}_x(x) = \frac{1}{7} \left(\frac{e^{-\frac{(x+3)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{(x+2)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{(x-2)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{(x-2.5)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{(x-3)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{-\frac{(x-4)^2}{2}}}{\sqrt{2\pi}} \right) \quad (8)$$

如图 6 所示，每个数据点给总的概率密度曲线估计贡献一条曲线。每一条曲线和横轴的面积均为 $1/7$ 。叠加得到的曲面和横轴围成图形的面积为 1。

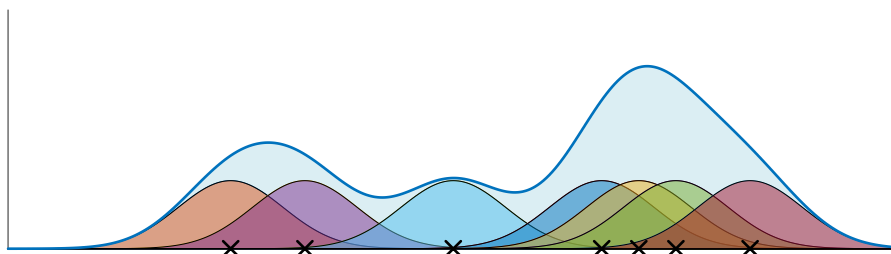


图 6. 用 7 个高斯核函数构造得到的概率密度估计曲线

以鸢尾花数据为例

图 7 所示为利用 `statsmodels.api.nonparametric.KDEUnivariate()` 对象得到的概率密度估计曲线。也可以通过它获得如图 8 所示累积概率密度估计曲线。

下一节将讲解带宽 h 如何影响概率密度估计曲线。

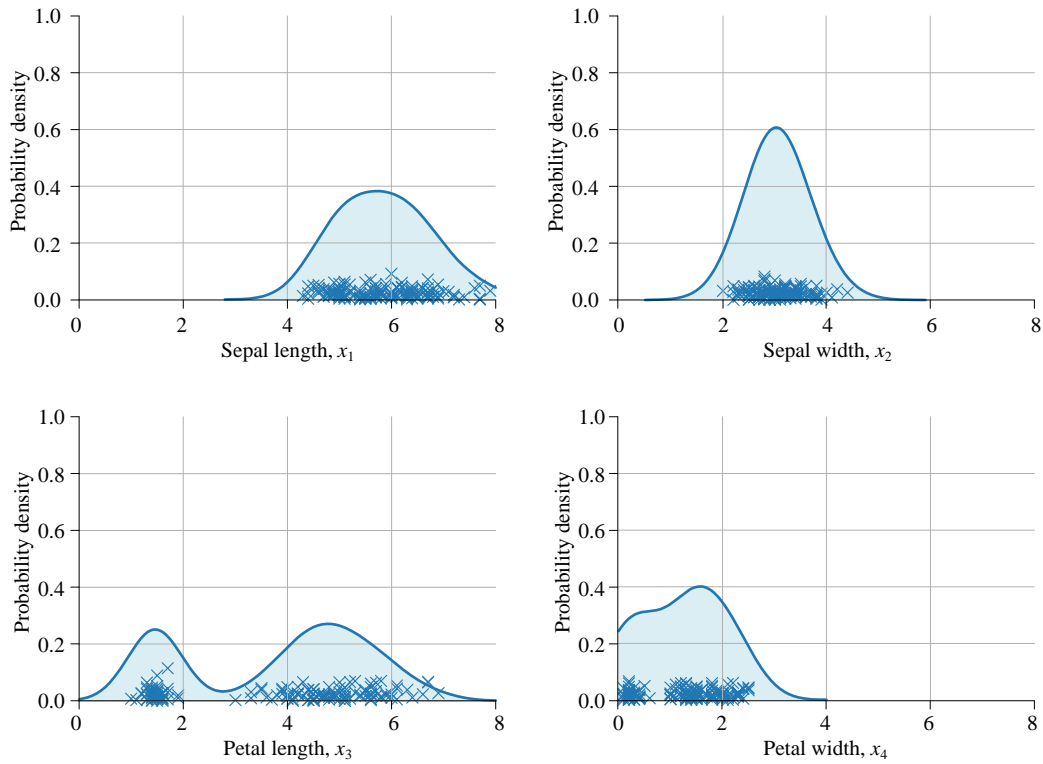


图 7. 鸢尾花四个特征数据的概率密度函数曲线

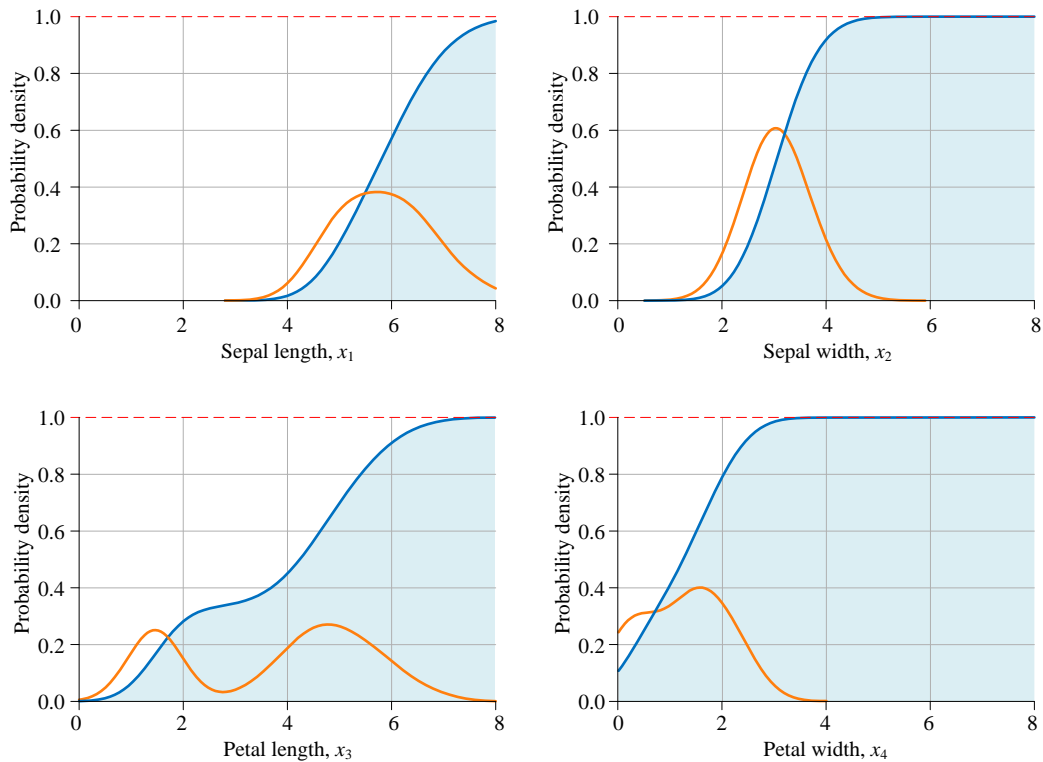
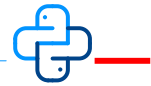


图 8. 鸢尾花四个特征数据的累积概率密度函数曲线



Bk5_Ch17_02.py 代码绘制图 7 和图 8。大家可以自行改变代码中带宽 h 。

17.3 带宽：决定核函数高矮胖瘦

带宽 h 选取对概率密度估计函数至关重要。 h 决定了每一个核函数的高矮胖瘦。图 9 所示为带宽 h 对高斯核函数形状影响。简单来说， h 小，核函数细高； h 大，核函数矮胖。

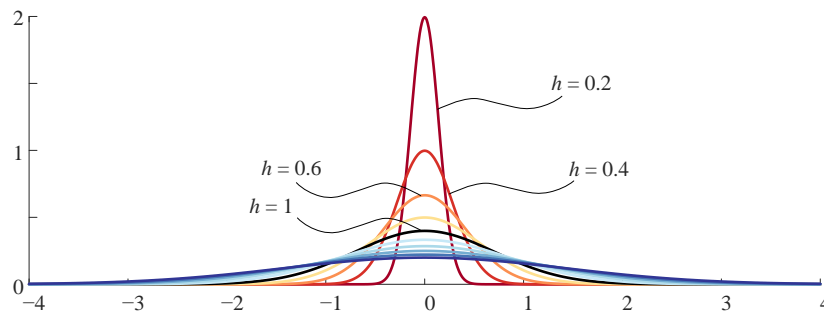


图 9. 带宽 h 对高斯核函数形状的影响

如图 10 所示，过小的 h ，会让概率密度估计曲线不够平滑；而太大的 h ，会让概率密度曲线过于平滑，大量有用信息被忽略。注意，不管 h 的大小，合成得到的概率密度曲线横轴包裹区域的面积始终保持为 1。图 11 和图 12 分别展示 $h = 0.1$ 、1 时鸢尾花概率密度估计曲线。

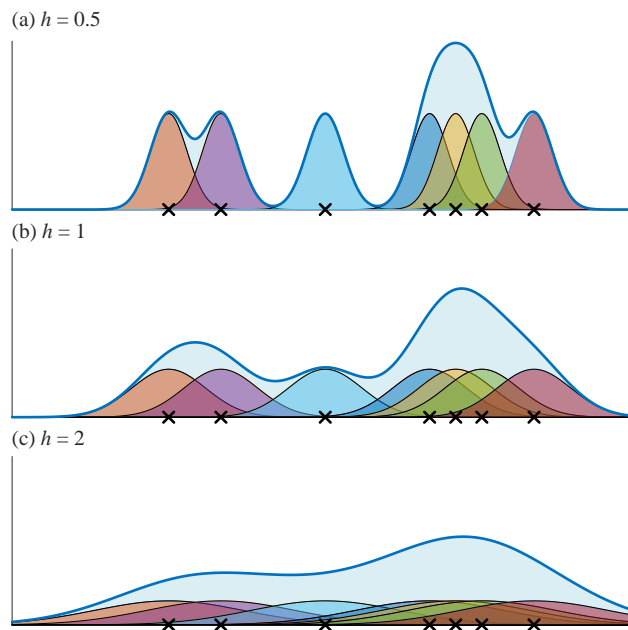


图 10. 核函数带宽对概率密度估计曲线影响

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

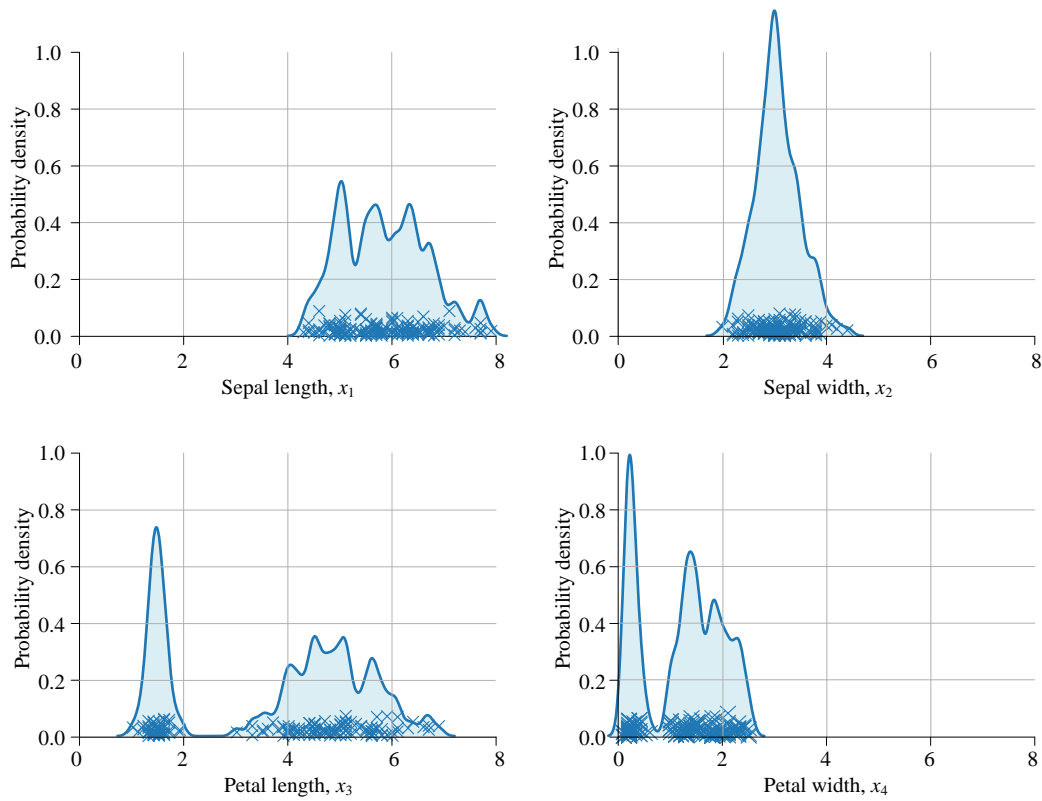


图 11. 鸢尾花四个特征数据的概率密度函数曲线, $h = 0.1$

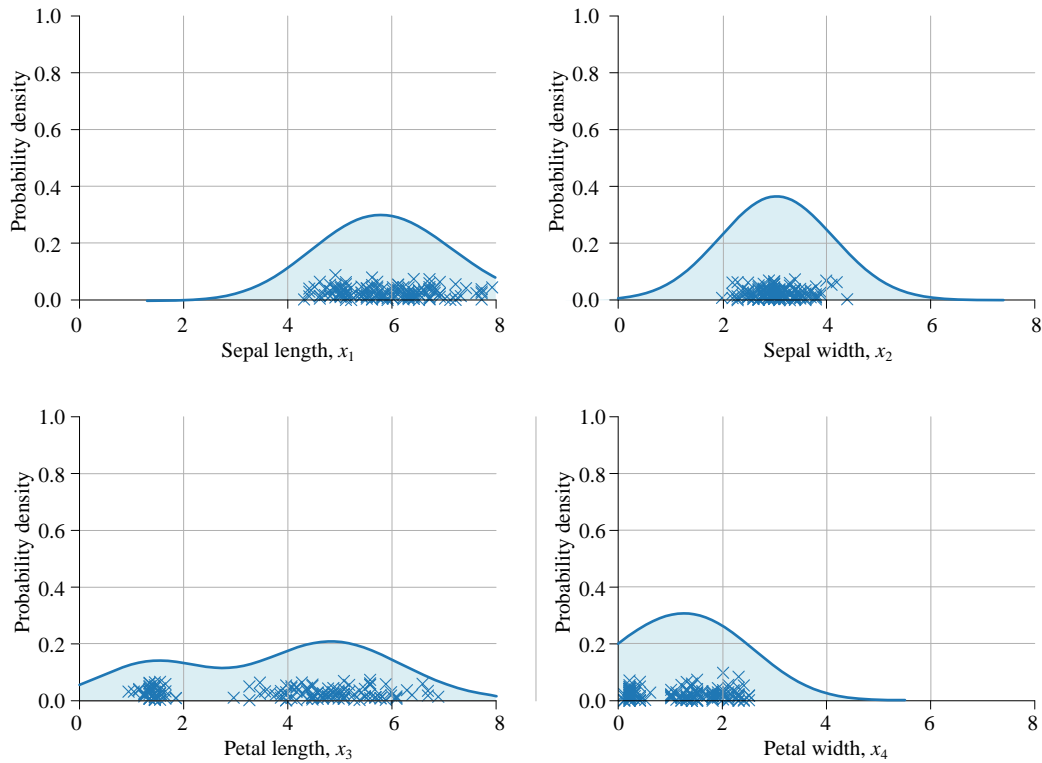


图 12. 鸢尾花四个特征数据的概率密度函数曲线, $h = 1$

对于高斯核函数，合理的 h 可以通过下式估算：

$$h \approx 1.06 \cdot n^{-\frac{1}{5}} \sigma \quad (9)$$

其中， σ 为样本数据的标准差， n 为样本数量。

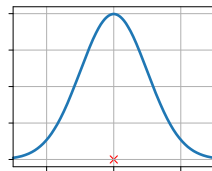
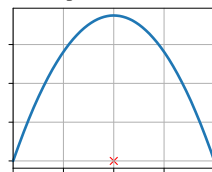

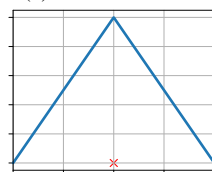
17.4 核函数：8 种常见核函数

总结来说，核函数需要满足两个重要条件：(1) 对称性；(2) 面积为 1。用公式表达：

$$\begin{aligned} K(x) &= K(-x) \\ \int_{-\infty}^{+\infty} K(x) dx &= \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \end{aligned} \quad (10)$$

表 1 总结 8 种满足以上两个条件的常用核函数。图 13 所示为这 8 种不同核函数估计得到的鸢尾花萼长度概率密度曲线。

表 1.8 种常见核函数

核函数	函数	函数图像
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$	(a) 'gau' 
Epanechnikov	$K(x) = \frac{3}{4}(1-x^2), x \leq 1$	(b) 'epa' 
Uniform	$K(x) = \frac{1}{2}, x \leq 1$	(c) 'uni' 
Triangular	$K(x) = 1- x , x \leq 1$	(d) 'tri' 

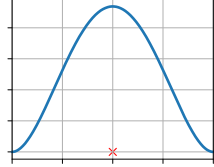
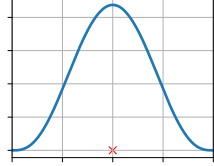
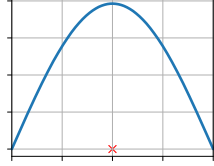
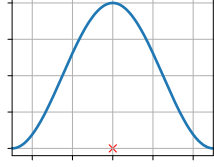
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

Biweight	$K(x) = \frac{15}{16}(1-x^2)^2, x \leq 1$	(e) 'biw' 
Triweight	$K(x) = \frac{35}{32}(1-x^2)^3, x \leq 1$	(f) 'triw' 
Cosine	$K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right), x \leq 1$	(g) 'cos' 
Cosine2	$K(x) = 1 + \cos(2\pi x), x \leq \frac{1}{2}$	(h) 'cos2' 

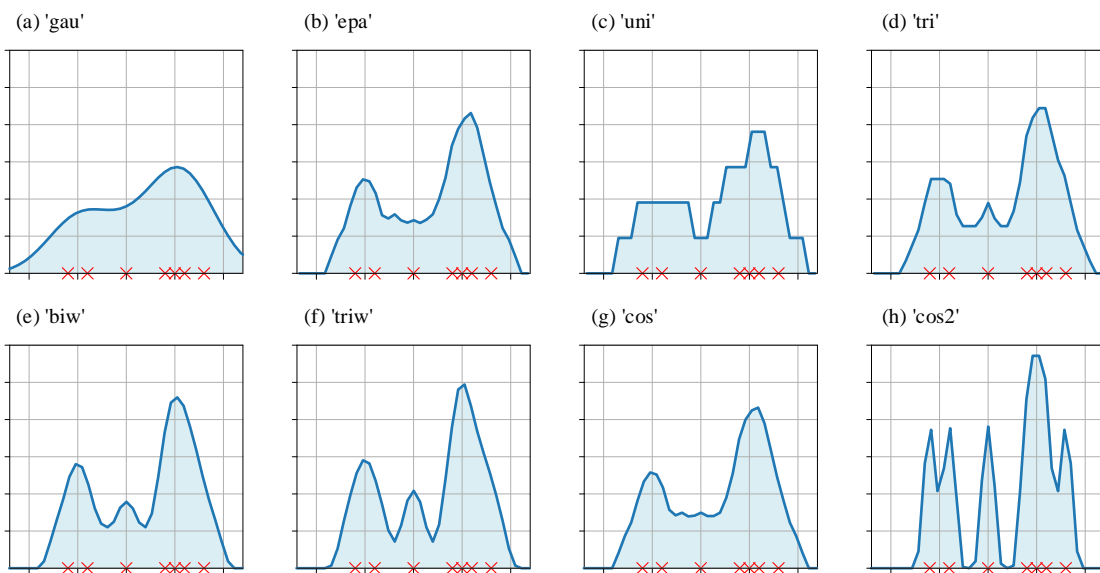


图 13. 八个不同核函数得到的不同的概率密度估计



Bk5_Ch17_03.py 代码绘制表 1 和图 13。也请大家学习使用 `sklearn.neighbors.KernelDensity()` 函数获得概率密度估计曲线。

17.5 二元 KDE：概率密度曲面

二元，乃至多元 KDE 的原理和前文所述的一元 KDE 完全相同。对于 n 个多维样本数据点 $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$ ，如下多个核函数叠加、再平均便得到概率密度估计：

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{x} - \mathbf{x}^{(i)}) \quad (11)$$

▲ 注意，默认 \mathbf{x} 和 $\mathbf{x}^{(i)}$ 均为列向量。 $\mathbf{x}^{(i)}$ 起到平移作用。

高斯核函数

高斯核函数 $K_H(\mathbf{x})$ 的定义为：

$$K_H(\mathbf{x}) = \det(\mathbf{H})^{-\frac{1}{2}} K\left(\mathbf{H}^{\frac{1}{2}} \mathbf{x}\right) \quad (12)$$

带宽的形式为矩阵 \mathbf{H} ， \mathbf{H} 为正定矩阵。以二元高斯核函数为例， $K(\mathbf{x})$ 定义为：

$$K(\mathbf{x}) = \frac{1}{2\pi} \exp\left(\frac{-\mathbf{x}^T \mathbf{x}}{2}\right) \quad (13)$$

图 14 所示为高斯核二元 KDE 原理。图中，每个样本点都用一个 IID 二元高斯分布曲面描述。这些曲面先叠加、再平均便获得二元高斯核 KDE 估计得到概率密度曲面。

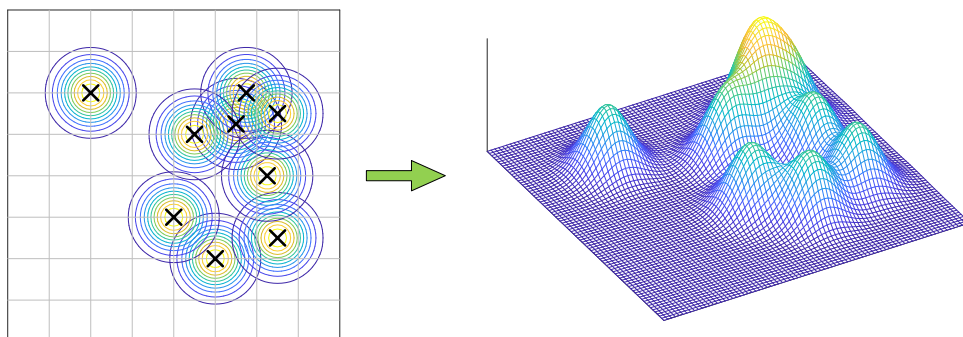


图 14. 二元高斯 KDE 原理

以鸢尾花数据为例

图 15 和图 16 所示为鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面。

`sklearn.neighbors.KernelDensity()` 函数也可以用来概率密度估计。注意，这个函数返回的是对数概率密度 $\ln(\text{PDF})$ 。

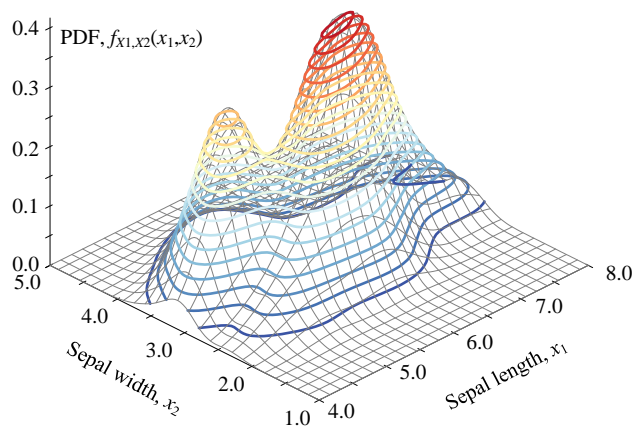


图 15. 鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面

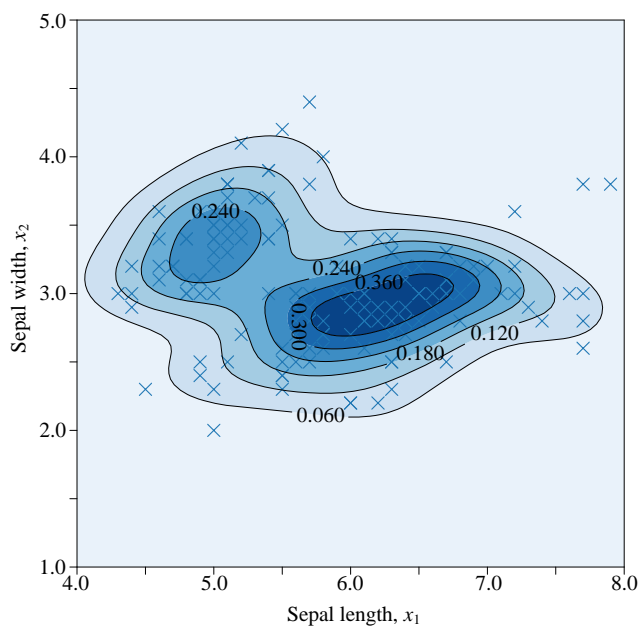
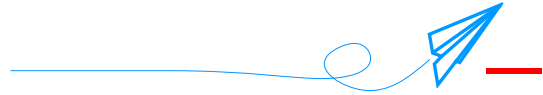


图 16. 鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面等高线图



Bk5_Ch17_04.py 代码绘制图 15 和图 16。Bk6_Ch17_05.py 用 Seaborn 绘制 KDE 曲面等高线。



在实际应用中，概率密度估计可以用来描述和模拟数据的分布特征，进行分类、聚类、异常检测等数据挖掘任务，也可以用于模型选择和参数估计。

常见的概率密度估计方法包括核密度估计、直方图估计、参数估计等。本章主要介绍的是核密度估计。核密度估计是一种非参数方法，可以用来估计连续随机变量的概率密度函数。请大家务必掌握高斯核密度估计，我们会在贝叶斯分类中看到这个工具的应用。

至此，本书有关频率学派的内容到此结束。前文反复提过，《统计至简》重贝叶斯学派，轻频率学派，下面连续五章我们将看到贝叶斯定理在分类、推断两类问题的应用。