Dive into Bayes' Theorem

20 深入贝叶斯定理

计算后验概率，利用花萼长度和宽度分类鸢尾花

> 杀不死你的，会让你更强大。
>
> ***What doesn't kill you, makes you stronger.***
>
> —— 弗里德里希·尼采 (Friedrich Nietzsche) | 德国哲学家 | 1844 ~ 1900

- `matplotlib.pyplot.contour3D()` 绘制三维等高线图
- `matplotlib.pyplot.contourf ()`绘制平面填充等高线
- `matplotlib.pyplot.fill_between()` 区域填充颜色
- `matplotlib.pyplot.plot_wireframe()` 绘制线框图
- `matplotlib.pyplot.scatter()` 绘制散点图
- `numpy.ones_like()` 用来生成和输入矩阵形状相同的全 1 矩阵
- `numpy.outer()` 计算外积，张量积
- `numpy.vstack()` 返回竖直堆叠后的数组
- `scipy.stats.gaussian_kde()` 高斯核密度估计
- `statsmodels.api.nonparametric.KDEUnivariate()` 构造一元 KDE

20.1 似然概率：给定分类条件下的概率密度

本章也是采用鸢尾花数据对鸢尾花分类进行预测；不同的是，本章采用花萼长度、花萼宽度两个特征。本章和上一章的编排类似，请大家对照阅读。

为了估算 $f_{X1,X2|Y}(x_1,x_2|C_1)$，首先提取标签为 C_1 (Setosa) 的 50 个样本，根据样本所在具体位置利用高斯 KDE 估计 $f_{X1,X2|Y}(x_1,x_2|C_1)$。

图 1 所示为通过高斯 KDE 方法估算得到的似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_1)$。$f_{X1,X2|Y}(x_1,x_2|C_1)$ 和水平面包裹的几何体的体积为 1。标签为 C_1 的鸢尾花数据，花萼长度主要集中在 4.5 ~ 5.5 cm 区域，花萼宽度则集中在 3 ~ 4 cm 区域。这个区域的 $f_{X1,X2|Y}(x_1,x_2|C_1)$ 曲面高度最高，也就是可能性最大。

本书第 6 章还给出过条件概率 $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_1)$ 平面等高线和条件边缘概率密度曲线，请大家回顾。

⚠ 注意，要计算概率，需要对 $f_{X1,X2|Y}(x_1,x_2|C_1)$ 进行二重积分。对 $f_{X1,X2|Y}(x_1,x_2|C_1)$ “偏积分”的结果为条件边缘概率密度 $f_{X1|Y}(x_1|C_1)$或 $f_{X2|Y}(x_2|C_1)$。

图 2 所示为似然概率 $f_{X1,X2|Y}(x_1,x_2|C_2)$ 曲面。图 3 为似然概率 $f_{X1,X2|Y}(x_1,x_2|C_3)$ 曲面。

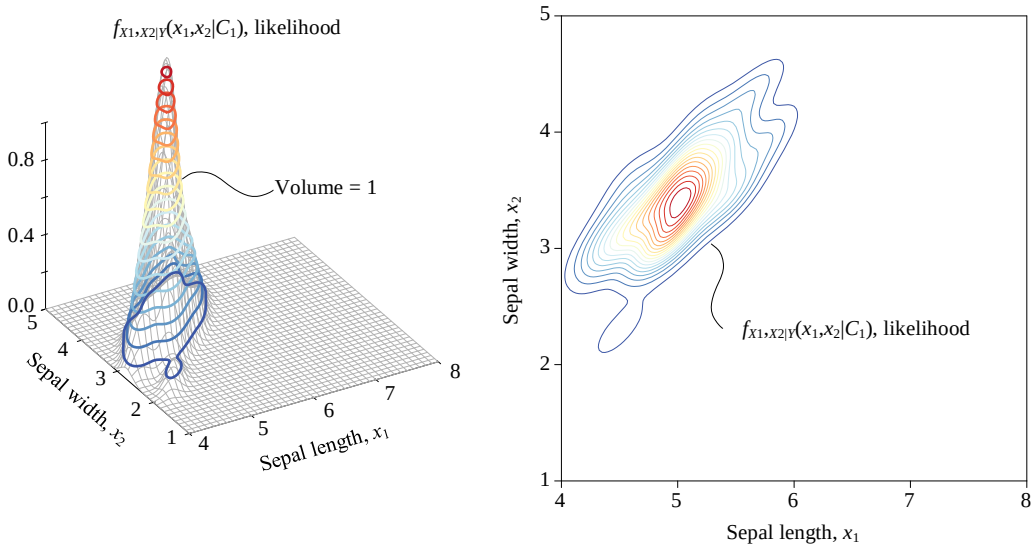

图 1. 似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_1)$

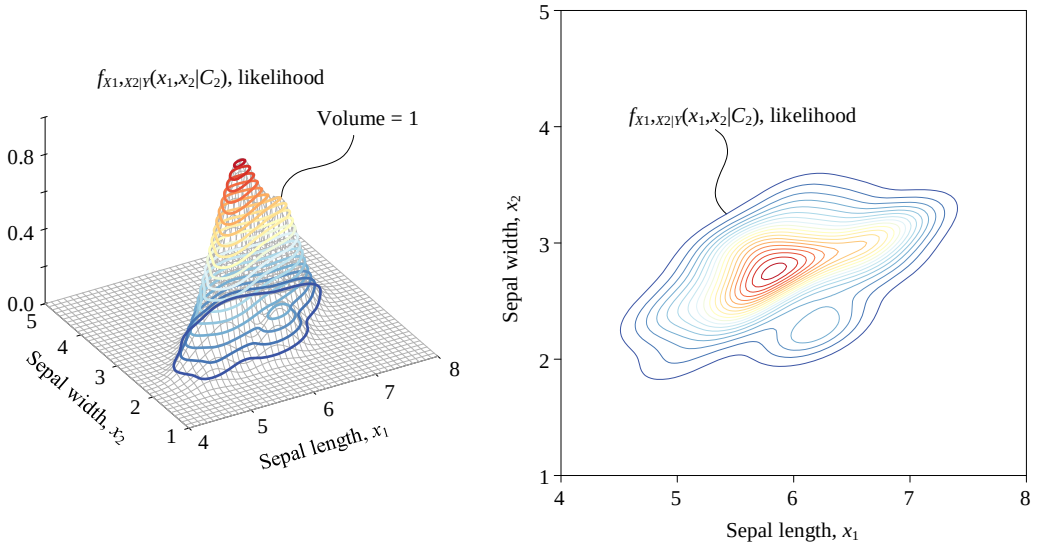

图 2. 似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_2)$

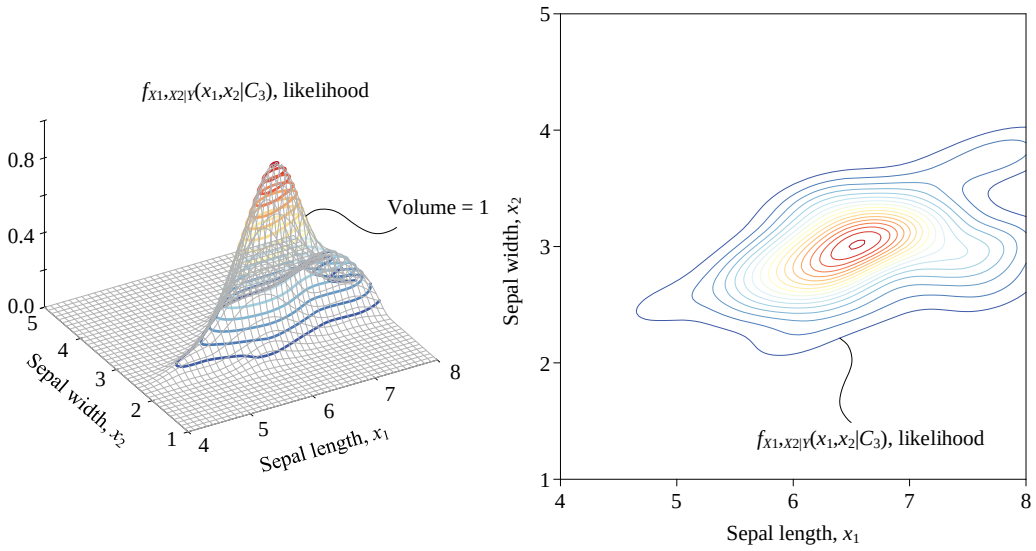

图 3. 似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_3)$

比较

图 4 比较 $f_{X1,X2|Y}(x_1,x_2|C_1)$、$f_{X1,X2|Y}(x_1,x_2|C_2)$、$f_{X1,X2|Y}(x_1,x_2|C_3)$ 三个似然概率平面等高线。

本章计算先验概率的方式和上一章完全一致，请大家回顾。然后利用贝叶斯定理，根据似然概率和先验概率就可以计算联合概率和证据因子。

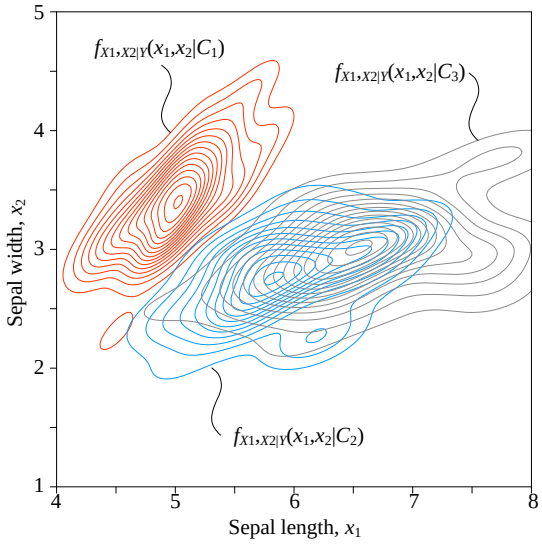

图 4. 比较三个似然概率曲面，平面等高线

20.2 联合概率：可以作为分类标准

联合概率 $f_{X1,X2,Y}(x_1,x_2,C_k)$ 描述三个事件 $X_1 = x_1$、$X_2 = x_2$、$Y = C_k$ 同时发生的可能性。

根据贝叶斯定理，联合概率 $f_{X1,X2,Y}(x_1,x_2,C_k)$ 可以通过似然概率 $f_{X1,X2|Y}(x_1,x_2|C_k)$ 和先验概率 $p_Y(C_k)$ 相乘得到：

$$\overbrace{f_{X1,X2,Y}\left(x_1,x_2,C_k\right)}^{\text{Joint}}=\overbrace{f_{X1,X2|Y}\left(x_1,x_2|C_k\right)}^{\text{Likelihood}}\overbrace{p_Y\left(C_k\right)}^{\text{Prior}} \tag{1}$$

对于鸢尾花分类问题，Y 为离散随机变量，而先验概率 $p_Y(C_k)$ 本身为概率质量函数，$p_Y(C_k)$ 在 (1) 中仅仅起到缩放作用。

图 5 所示为联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1,x_2,C_1)$，$f_{X1,X2,Y}(x_1,x_2,C_1)$ 和水平面包裹的几何体的体积为 $p_Y(C_1)$。图 6 和图 7 所示为 $f_{X1,X2,Y}(x_1,x_2,C_2)$ 和 $f_{X1,X2,Y}(x_1,x_2,C_3)$ 两个联合概率曲面。

上一章介绍过，比较三个联合概率曲面高度可以用作鸢尾花分类预测的依据。

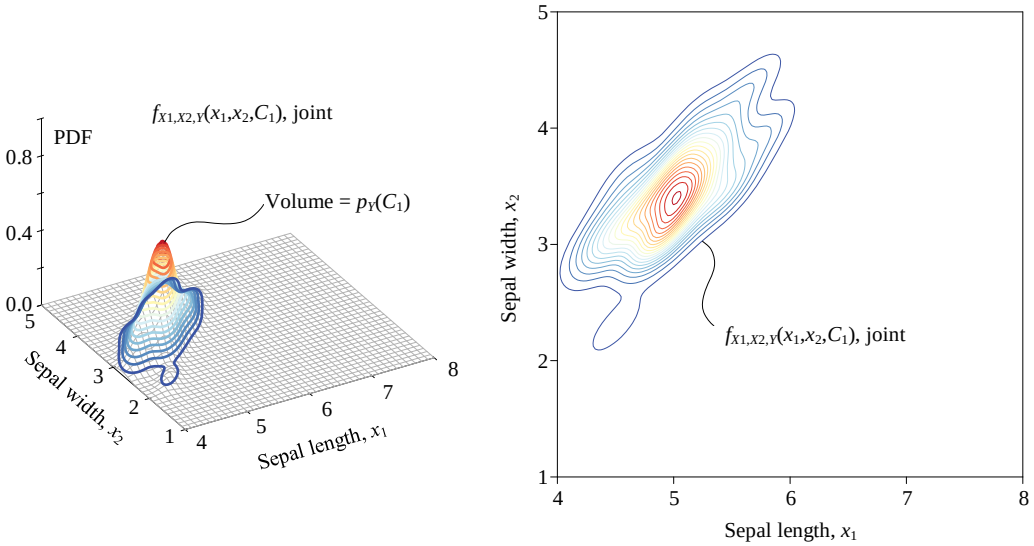

图 5. 联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1,x_2,C_1)$

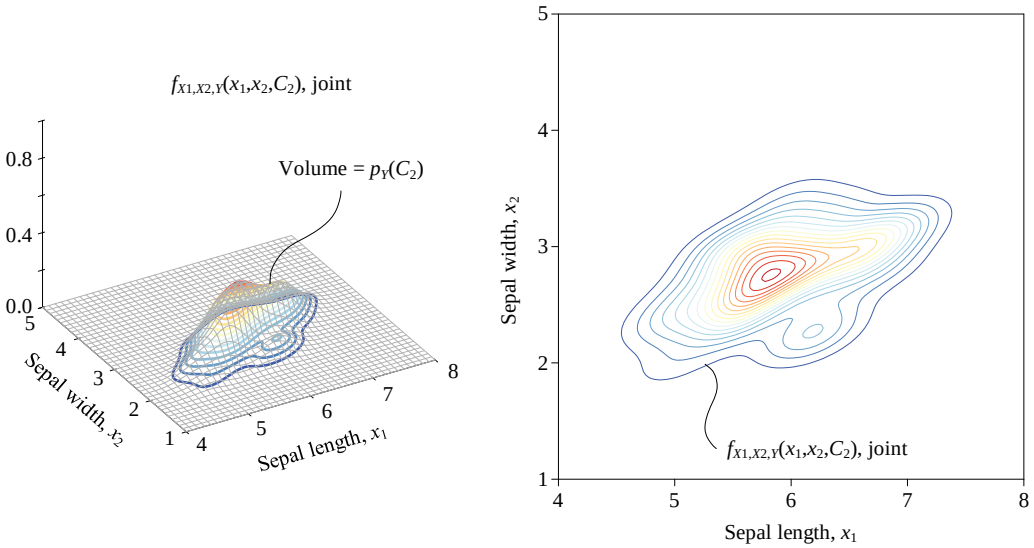

图 6. 联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1,x_2,C_2)$

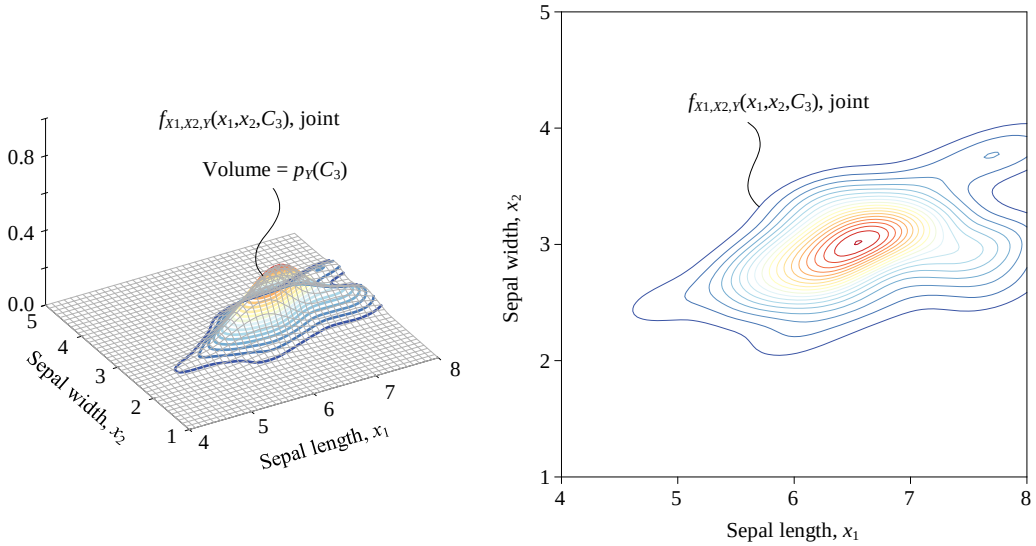

图 7. 联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1,x_2,C_3)$

20.3 证据因子：和分类无关

证据因子 $f_{X1,X2}(x_1,x_2)$ 描述样本数据的分布情况，和分类无关。

C_1、C_2、C_3 为一组不相容分类，对鸢尾花数据样本空间 Ω 形成分割。根据全概率定理，下式成立：

$$\overbrace{f_{X1,X2}(x_1,x_2)}^{\text{Evidence}}=\sum_{k=1}^{3}\overbrace{f_{X1,X2,Y}(x_1,x_2,C_k)}^{\text{Joint}}=\sum_{k=1}^{3}\overbrace{f_{X1,X2|Y}(x_1,x_2|C_k)}^{\text{Likelihood}}\overbrace{p_Y(C_k)}^{\text{Prior}} \tag{2}$$

上式可以用来估算 $f_{X1,X2}(x_1,x_2)$。

把 (2) 展开来写，证据因子 $f_{X1,X2}(x_1,x_2)$ 可以通过下式计算得到：

$$\begin{aligned} f_{X1,X2}(x_1,x_2) &= f_{X1,X2,Y}(x_1,x_2,C_1)+f_{X1,X2,Y}(x_1,x_2,C_2)+f_{X1,X2,Y}(x_1,x_2,C_3) \\ &= f_{X1,X2|Y}(x_1,x_2|C_1)\,p_Y(C_1)+f_{X1,X2|Y}(x_1,x_2|C_2)\,p_Y(C_2)+f_{X1,X2|Y}(x_1,x_2|C_3)\,p_Y(C_3) \end{aligned} \tag{3}$$

图 8 所示为叠加联合概率 PDF 曲面，计算证据因子 PDF 的过程。图 8 左侧三个几何体的体积分别为 $p_Y(C_1)$、$p_Y(C_2)$、$p_Y(C_3)$。显然 $p_Y(C_1)$、$p_Y(C_2)$、$p_Y(C_3)$ 三者之和为 1。

图 9 所示为 $f_{X1,X2}(x_1,x_2)$ 的曲面和平面等高线图。可以发现 $f_{X1,X2}(x_1,x_2)$ 较好地描述了样本数据分布。

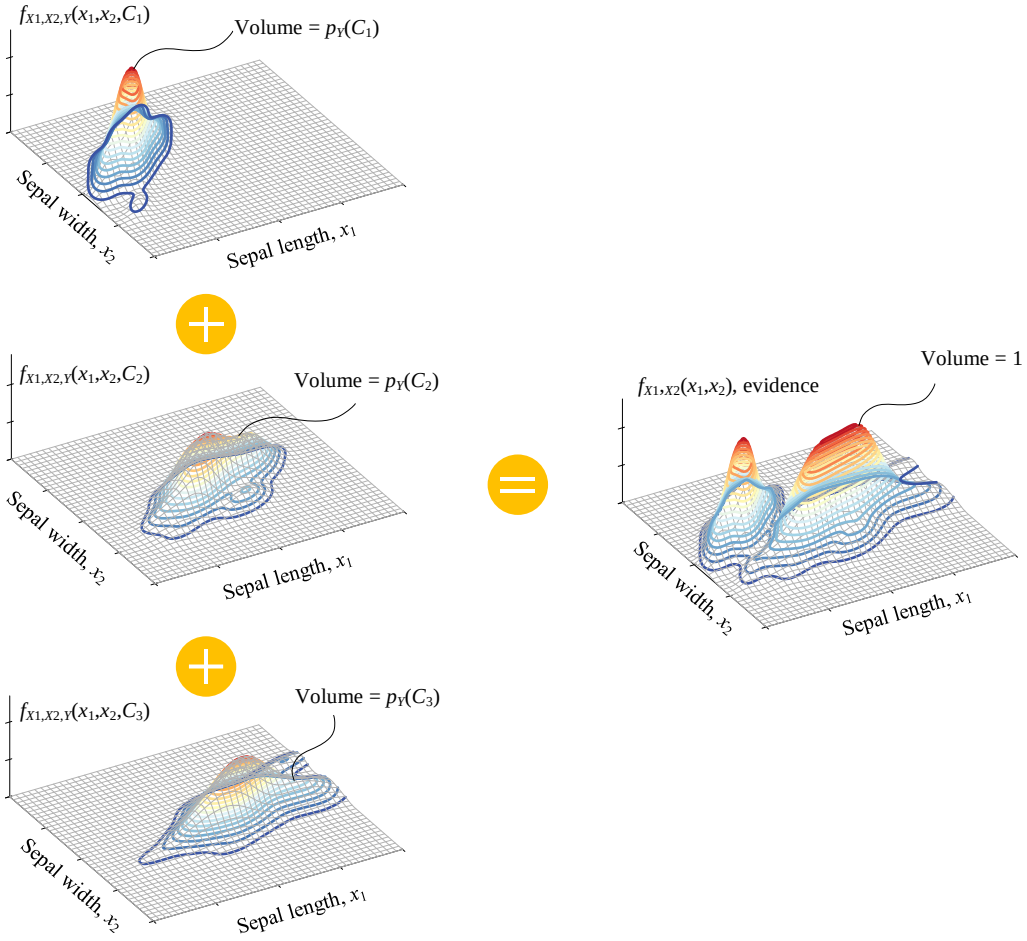

图 8. 叠加联合概率曲面，估算证据因子概率密度函数 $f_{X1,X2}(x_1,x_2)$

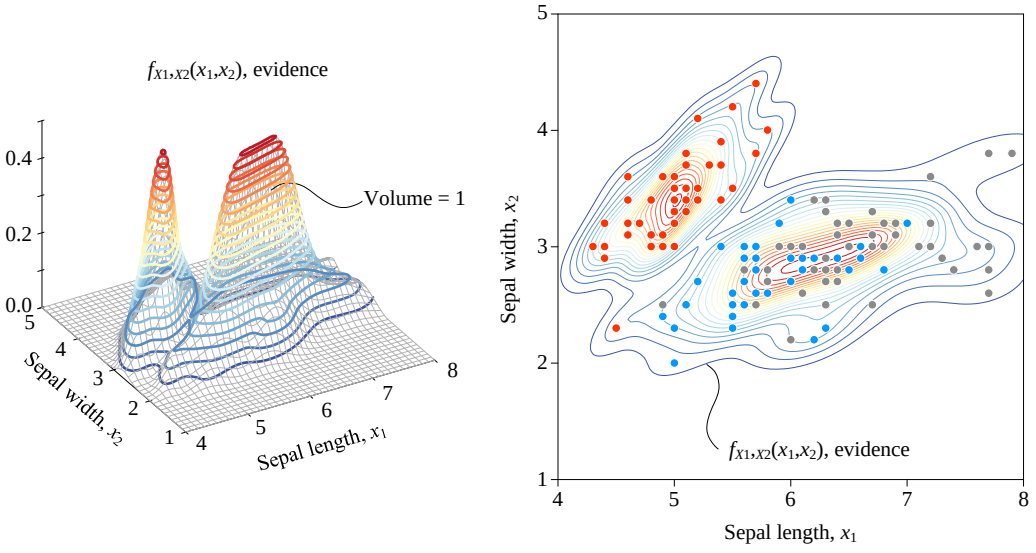

图 9. $f_{X1,X2}(x_1,x_2)$ 曲面及平面等高线

20.4 后验概率：也是分类的依据

$f_{Y|X1,X2}(C_k \mid x_1,x_2)$ 作为条件概率，指的是在 $X_1 = x_1$ 和 $X_2 = x_2$ 发生条件下，事件 $Y = C_k$ 发生的概率。上一章提到，$f_{Y|X1,X2}(C_k \mid x_1,x_2)$ 本身为概率，也就是说 $f_{Y|X1,X2}(C_k \mid x_1,x_2)$ 的取值范围为 [0, 1]；因此，后验概率 $f_{Y|X1,X2}(C_k \mid x_1,x_2)$ 又叫成员值。

根据贝叶斯定理，当 $f_{X1,X2}(x_1,x_2) > 0$ 时，后验概率 PDF $f_{Y|X1,X2}(C_k \mid x_1,x_2)$ 可以根据下式计算得到：

$$\overbrace{f_{Y|X1,X2}\left(C_k \middle| x_1,x_2\right)}^{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1,x_2,C_k\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1,x_2\right)}_{\text{Evidence}}} \tag{4}$$

图 10、图 11、图 12 所示分别为后验概率 $f_{Y|X1,X2}(C_1 \mid x_1,x_2)$、$f_{Y|X1,X2}(C_2 \mid x_1,x_2)$、$f_{Y|X1,X2}(C_3 \mid x_1,x_2)$ 对应曲面和平面等高线。

上一章提到，后验概率 (成员值) 存在以下关系：

$$\sum_{k=1}^{3} \underbrace{f_{Y|X1,X2}\left(C_k \middle| x_1,x_2\right)}_{\text{Posterior}} = 1 \tag{5}$$

这意味着，图 10、图 11、图 12 三幅图曲面叠加在一起得到高度为 1 的“平台”。

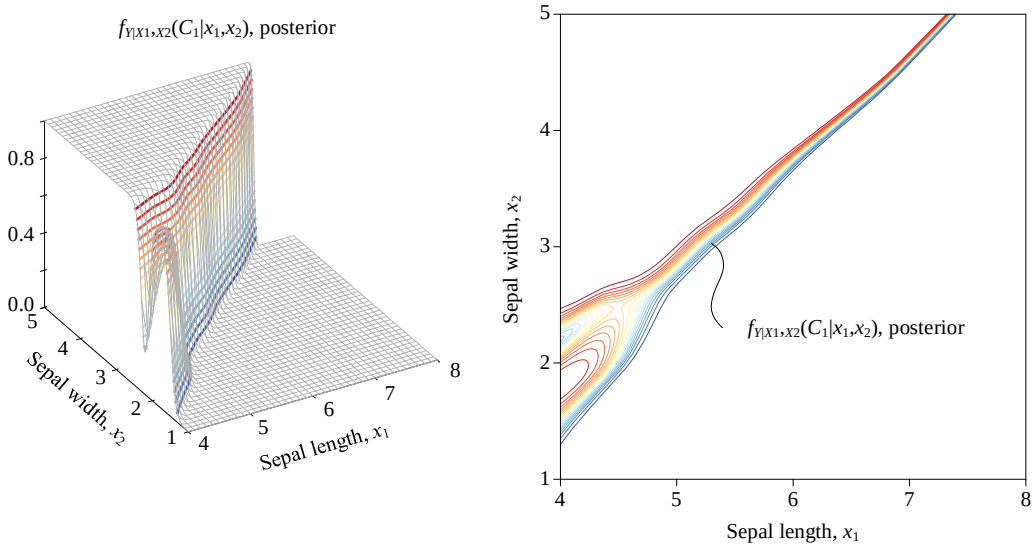

图 10. 后验概率 $f_{Y|X1,X2}(C_1 \mid x_1,x_2)$ 对应曲面和平面等高线

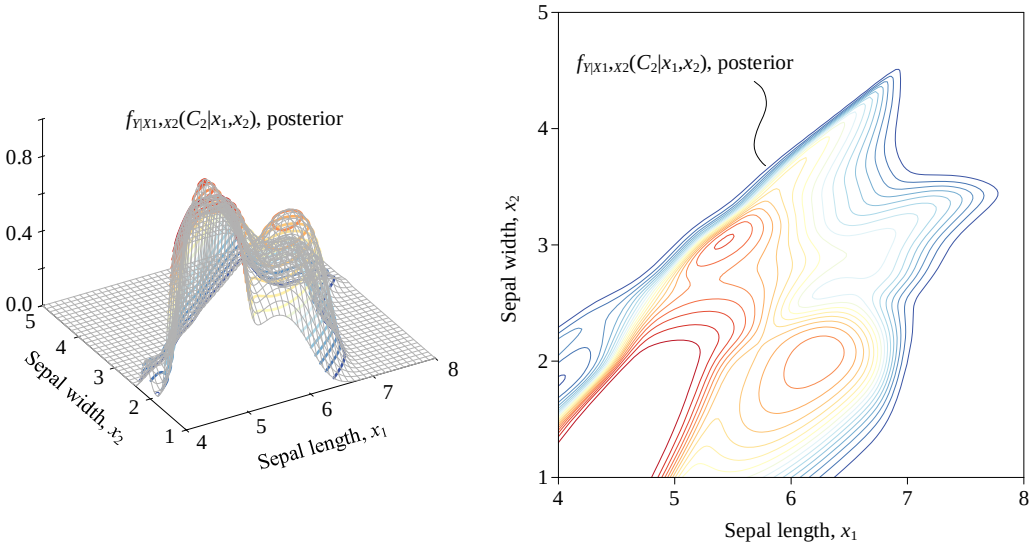

图 11. 后验概率 $f_{Y|X1,X2}(C_2 \mid x_1,x_2)$ 对应曲面和平面等高线

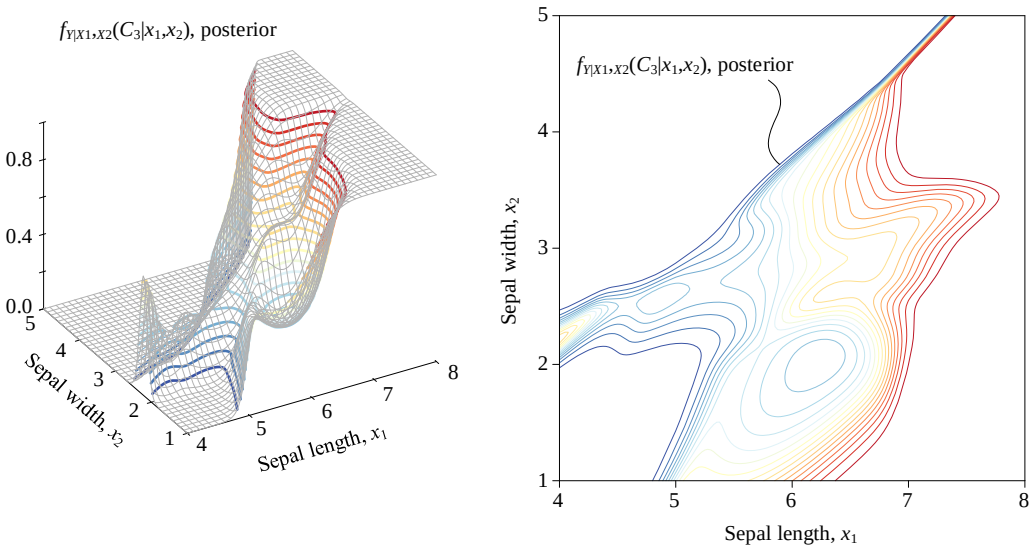

图 12. 后验概率 $f_{Y|X1,X2}(C_3 \mid x_1,x_2)$ 对应曲面和平面等高线

分类依据

在给定任意花萼长度 x_1 和花萼宽度 x_2 的条件下，比较图 13 所示三个后验概率 $f_{Y|X1,X2}(C_1 \mid x_1,x_2)$、$f_{Y|X1,X2}(C_2 \mid x_1,x_2)$、$f_{Y|X1,X2}(C_3 \mid x_1,x_2)$大小，最大后验概率对应的标签就可以作为鸢尾花分类依据。

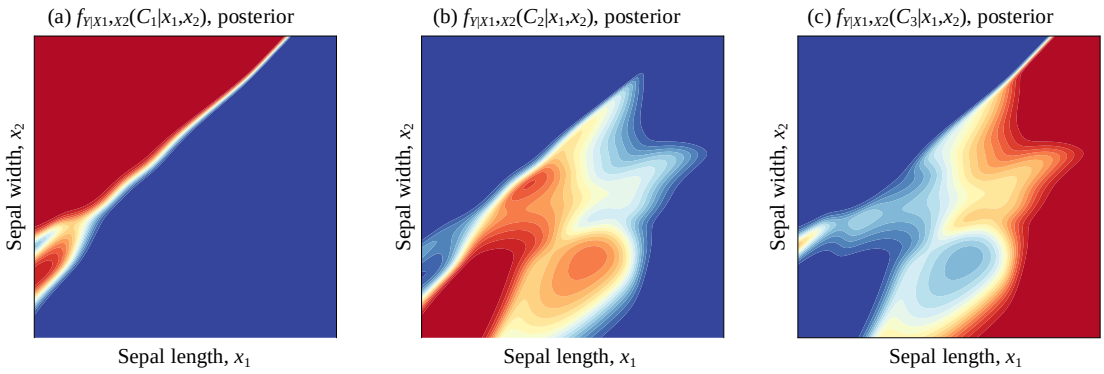

图 13. 比较三个后验概率曲面平面填充等高线

也就是说，这个分类问题对应的优化目标为最大化后验概率，即：

$$\hat{y} = \arg\max_{C_k} f_{Y|X1,X2}\left(C_k \middle| x_1, x_2\right) \tag{6}$$

其中，$k = 1,2,\ldots K$。对于鸢尾花三分类问题，$K = 3$。

图 14 这幅图中曲线就是所谓决策边界 (decision boundary)，决策边界将平面划分成三个区域，每个区域对应一类鸢尾花标签。

《机器学习》一册将探讨更多分类算法。

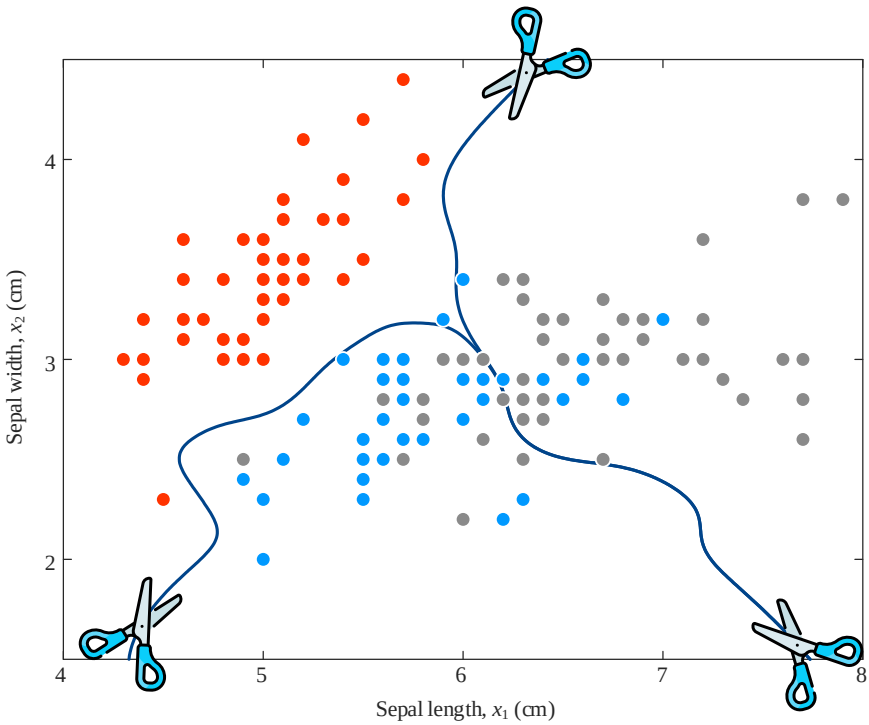

图 14. 朴素贝叶斯决策边界，基于核密度估计 KDE

20.5 独立：不代表条件独立

本章最后以鸢尾花数据为例区分“独立”和“条件独立”这两个概念。

如果假设鸢尾花花萼长度 X_1 和花萼宽度 X_2 两个随机变量独立，联合概率 $f_{X1,X2}(x_1,x_2)$ 可以通过下式计算得到：

$$\underbrace{f_{X1,X2}(x_1,x_2)}_{\text{Joint}} = \underbrace{f_{X1}(x_1)}_{\text{Marginal}} \cdot \underbrace{f_{X2}(x_2)}_{\text{Marginal}} \tag{7}$$

图 15 所示为假设 X_1 和 X_2 独立时，估算得到的联合概率 $f_{X1,X2}(x_1,x_2)$ 曲面和平面等高线。观察图 15 等高线，容易发现假设 X_1 和 X_2 独立估算得到的联合概率 $f_{X1,X2}(x_1,x_2)$ 并没有很好地描述鸢尾花数据分布。

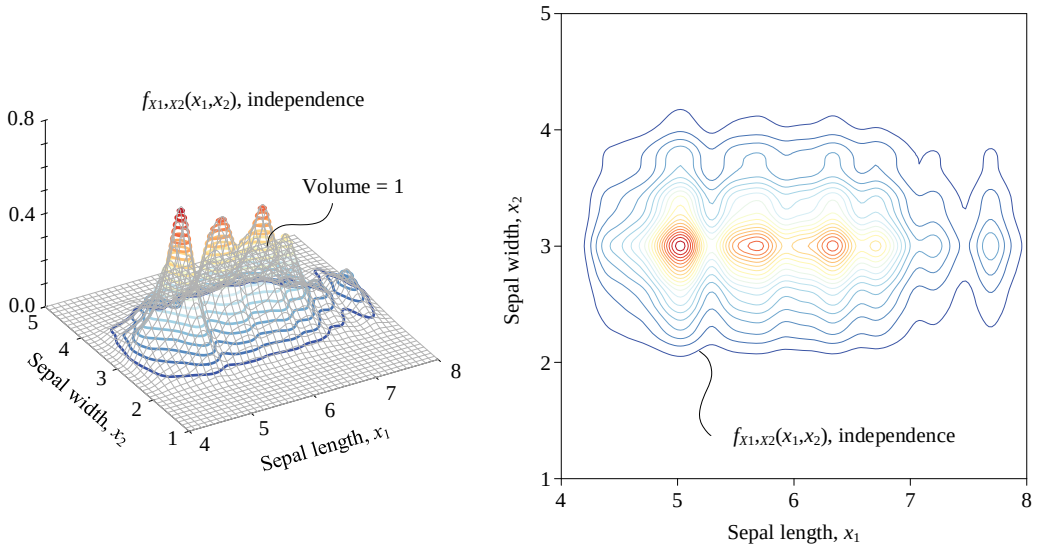

图 15. X_1 和 X_2 独立时，估算得到的联合概率 $f_{X1,X2}(x_1,x_2)$ 曲面和曲面等高线

图 16 所示为将 $f_{X1,X2}(x_1,x_2)$ 曲面在两个不同平面的投影。可以发现在不同平面上的投影都相当于该方向上边缘分布的高度上缩放。

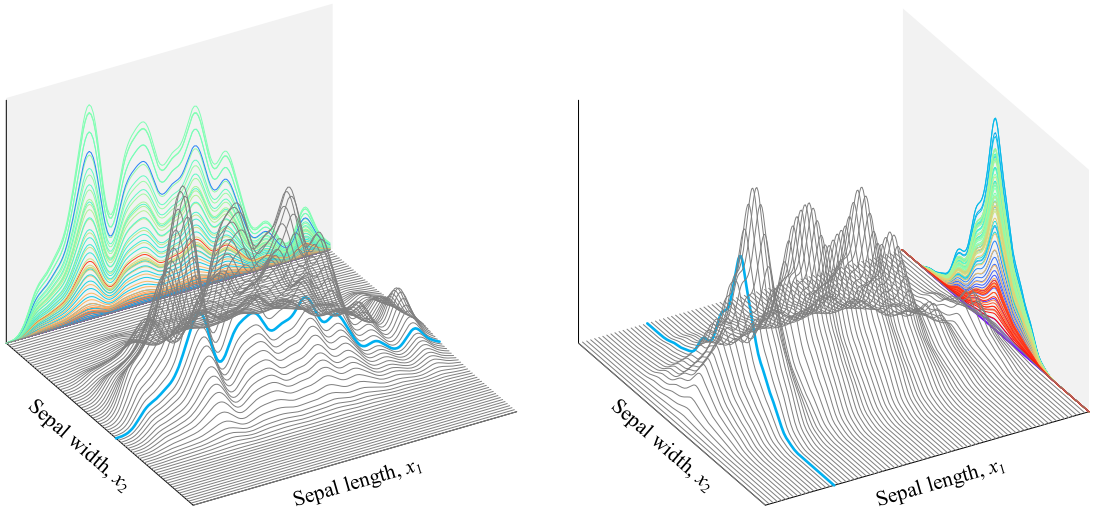

图 16. $f_{X1,X2}(x_1,x_2)$ 曲面在两个不同平面的投影，假设特征独立

20.6 条件独立：不代表独立

回顾本书前文讲过的条件独立。如果 $\Pr(X,Y|Z) = \Pr(X|Z) \cdot \Pr(Y|Z)$，则称事件 X、Y 对于给定事件 Z 是条件独立的。也就是说，当 Z 发生条件下，X 发生与否与 Y 发生与否无关。

对于鸢尾花样本数据，给定 $Y = C_k$ 的条件下，如果假设花萼长度 X_1、花萼宽度 X_2 条件独立，则下式成立：

$$f_{X1,X2|Y}\left(x_1,x_2|C_k\right) = f_{X1|Y}\left(x_1|C_k\right) \cdot f_{X2|Y}\left(x_2|C_k\right) \tag{8}$$

上式相当于，一个类别、一个类别地分析数据。

$Y = C_1$ 条件

给定 $Y = C_1$ 的条件下，如果假设 X_1、X_2 条件独立，则：

$$f_{X1,X2|Y}\left(x_1,x_2|C_1\right) = f_{X1|Y}\left(x_1|C_1\right) \cdot f_{X2|Y}\left(x_2|C_1\right) \tag{9}$$

图 17 所示为在 $Y = C_1$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_1)$。

本书第 6 章给出过假设条件独立情况下 $f_{X1,X2|Y}(x_1,x_2|C_1)$、边缘似然概率 $f_{X1|Y}(x_1|C_1)$、$f_{X2|Y}(x_2|C_1)$ 三者关系，请大家回顾。如果把 $f_{X1|Y}(x_1|C_1)$、$f_{X2|Y}(x_2|C_1)$ 看做两个向量的话，$f_{X1,X2|Y}(x_1,x_2|C_1)$ 就是两者的张量积。

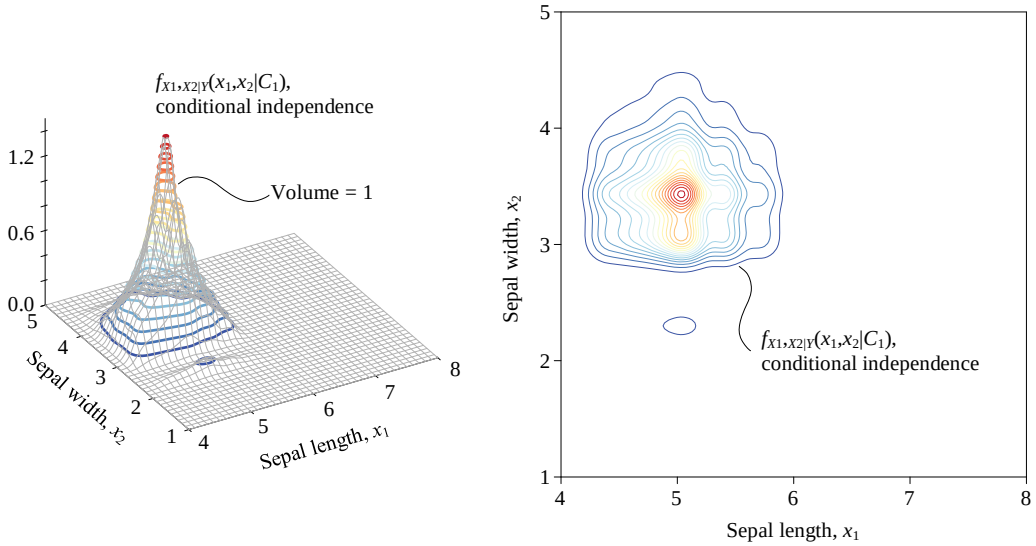

图 17. 在 $Y = C_1$ 的条件下，X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_1)$

$Y = C_2$ 条件

给定 $Y = C_2$ 的条件下，如果假设 X_1、X_2 条件独立，则：

$$f_{X1,X2|Y}\left(x_1,x_2|C_2\right) = f_{X1|Y}\left(x_1|C_2\right)\cdot f_{X2|Y}\left(x_2|C_2\right) \tag{10}$$

图 18 所示为在 $Y = C_2$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_2)$。

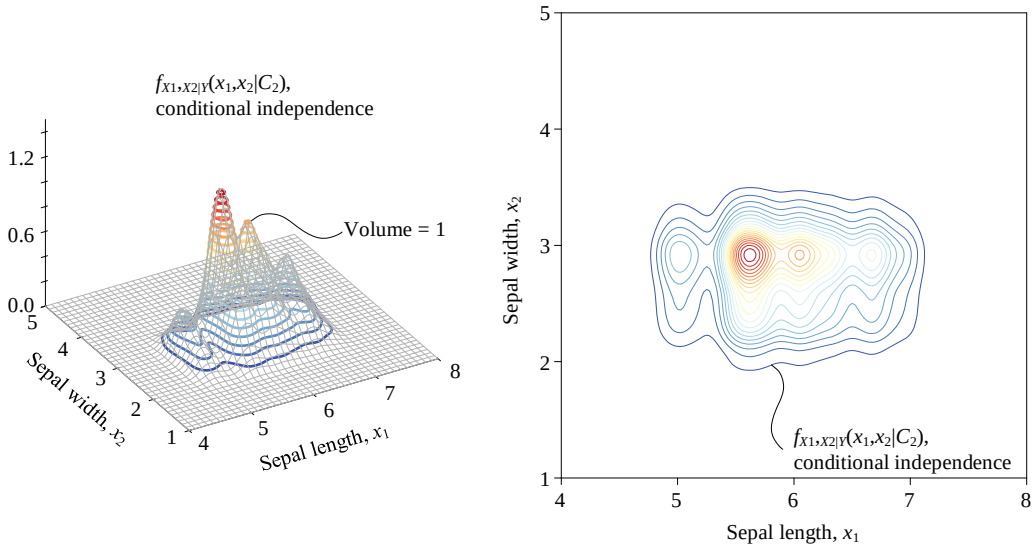

图 18. 在 $Y = C_2$ 的条件下，X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_2)$

$Y = C_3$ 条件

给定 $Y = C_3$ 的条件下，如果假设 X_1、X_2 条件独立，则：

$$f_{X1,X2|Y}\left(x_1,x_2|C_3\right)=f_{X1|Y}\left(x_1|C_3\right)\cdot f_{X2|Y}\left(x_2|C_3\right) \tag{11}$$

图 19 所示为在 $Y = C_3$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_3)$。

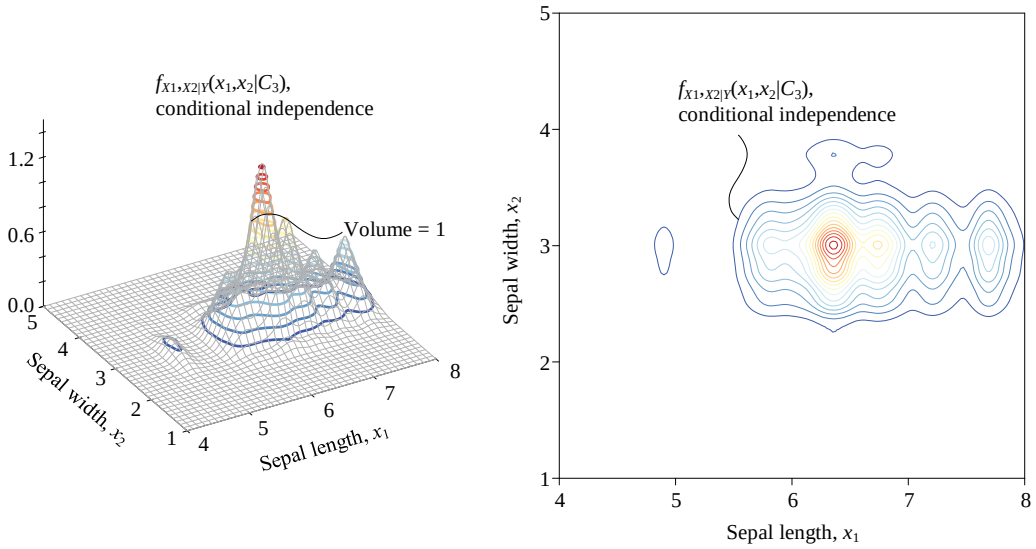

图 19. 在 $Y = C_3$ 的条件下，X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_3)$

估算证据因子

假设条件独立，证据因子 $f_{X1,X2}(x_1,x_2)$ 可以通过下式计算得到：

$$\begin{aligned}f_{X1,X2}\left(x_1,x_2\right)&=f_{X1,X2|Y}\left(x_1,x_2|C_1\right)\cdot p_Y\left(C_1\right)+f_{X1,X2|Y}\left(x_1,x_2|C_2\right)\cdot p_Y\left(C_2\right)+f_{X1,X2|Y}\left(x_1,x_2|C_3\right)\cdot p_Y\left(C_3\right)\\&=f_{X1|Y}\left(x_1|C_1\right)\cdot f_{X2|Y}\left(x_2|C_1\right)\cdot p_Y\left(C_1\right)+\\&\quad f_{X1|Y}\left(x_1|C_2\right)\cdot f_{X2|Y}\left(x_2|C_2\right)\cdot p_Y\left(C_2\right)+\\&\quad f_{X1|Y}\left(x_1|C_3\right)\cdot f_{X2|Y}\left(x_2|C_3\right)\cdot p_Y\left(C_3\right)\end{aligned} \tag{12}$$

上式代表一种多元概率密度估算方法。图 20 所示为假设条件独立，估算 $f_{X1,X2}(x_1,x_2)$ 概率密度的过程。图 21 所示为 $f_{X1,X2}(x_1,x_2)$ 曲面和平面等高线。

条件独立这一假设对于朴素贝叶斯方法至关重要。《机器学习》一册将分别介绍朴素贝叶斯分类，和高斯朴素贝叶斯分类。

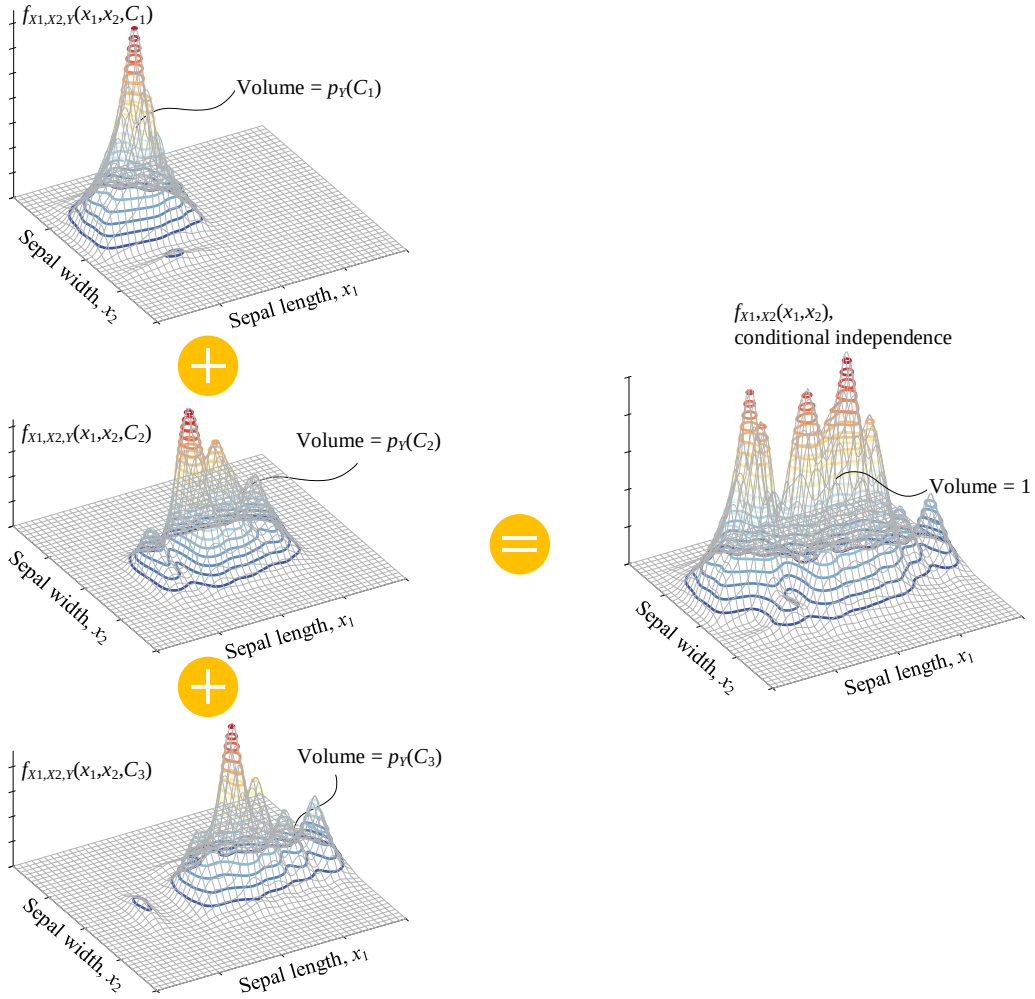

图 20. 假设条件独立，合成叠加得到证据因子 $f_{X1,X2}(x_1,x_2)$

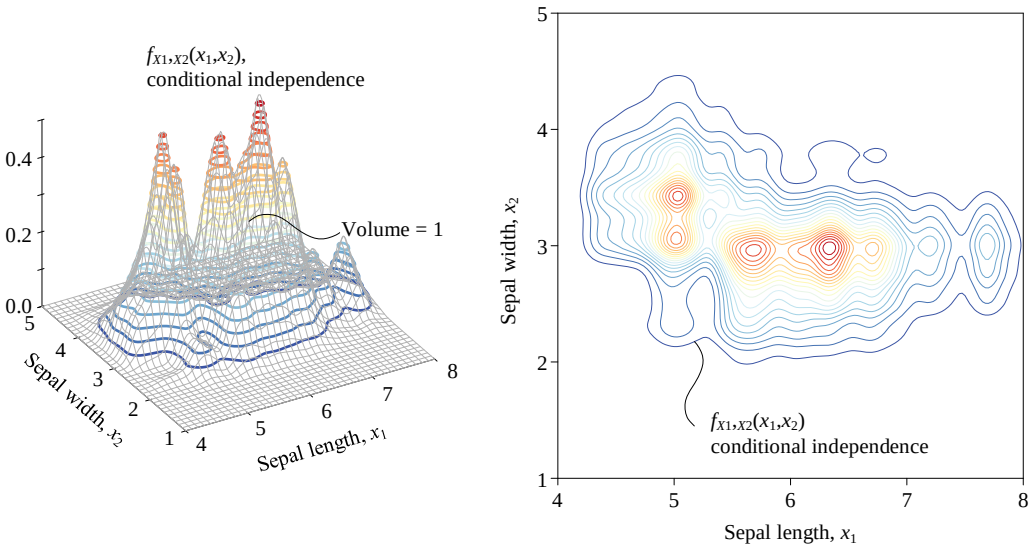

图 21. 假设条件独立，证据因子 $f_{X1,X2}(x_1,x_2)$ 曲面和平面等高线

如图 22 所示，显然采用条件独立假设估算得到的证据因子概率密度函数 $f_{X1,X2}(x_1,x_2)$ 对样本数据分布的贴合度更高。图 23 所示为 $f_{X1,X2}(x_1,x_2)$ 在两个竖直平面上的投影，请大家对比图 16 分析。

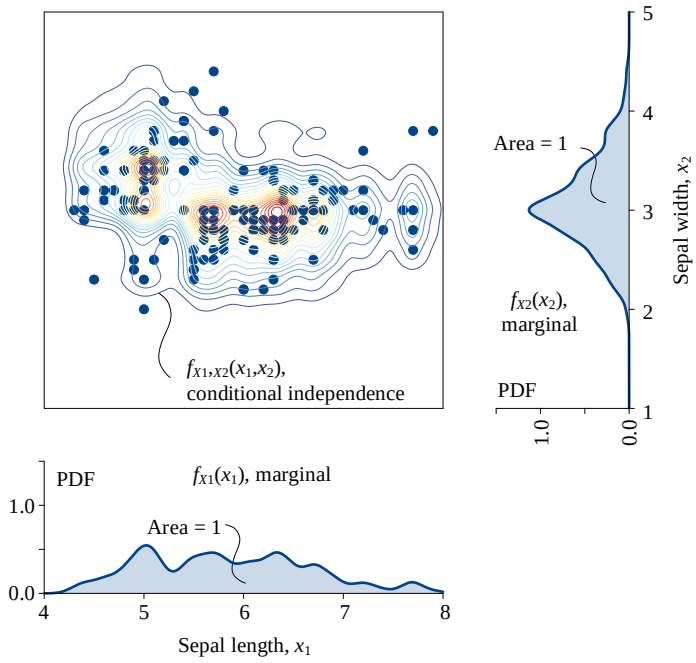

图 22. 假设条件独立，证据因子 $f_{X1,X2}(x_1,x_2)$ 等高线，和边缘概率密度 $f_{X1}(x_1)$、$f_{X2}(x_2)$ 曲线关系

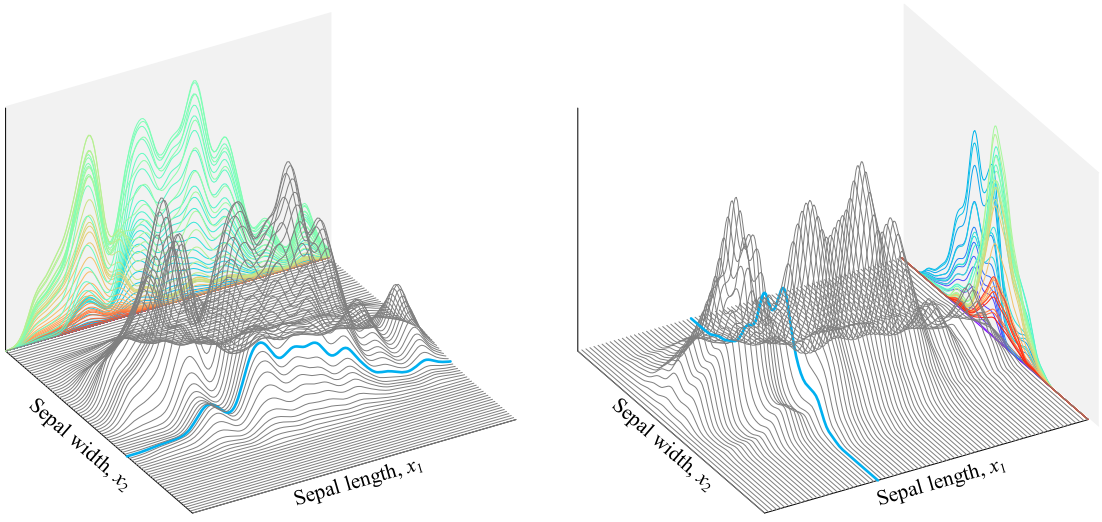

图 23. 假设条件独立，证据因子 $f_{X1,X2}(x_1,x_2)$ 曲面在两个平面投影曲线

Bk5_Ch020_01.py 代码绘制本书绝大部分图像。