Naive Bayes Classifier

4 朴素贝叶斯

假设特征之间条件独立，最大化后验概率

大家使用朴素贝叶斯分类器时，假设特征 (条件) 独立。之所以称之“朴素”，是因为那真是个“天真”的假设。

A learner that uses Bayes' theorem and assumes the effects are independent given the cause is called a Naïve Bayes classifier. That's because, well, that's such a naïve assumption.

—— 佩德罗·多明戈斯 (Pedro Domingos) | 《终极算法》作者，华盛顿大学教授 | 1965 ~

- `matplotlib.axes.Axes.contour()` 绘制平面和空间等高线图
- `matplotlib.Axes3D.plot_wireframe()` 绘制三维单色网格图
- `matplotlib.pyplot.bar()` 绘制直方图
- `seaborn.barplot()` 绘制直方图
- `seaborn.displot()` 绘制一元和二元条件边际分布
- `seaborn.jointplot()` 同时绘制分类数据散点图、分布图和边际分布图

4.1 重逢贝叶斯

贝叶斯是我们的老朋友，《概率统计》一册用了很大篇幅介绍了贝叶斯定理 (Bayes' theorem) 和应用。本章和下一章，贝叶斯定理将专门用来解决数据分类问题。这种分类方法叫做朴素贝叶斯分类 (Naive Bayes classification)。

托马斯·贝叶斯 (Thomas Bayes) | 英国数学家 | 1702 ~ 1761
贝叶斯统计的开山鼻祖，以贝叶斯定理闻名于世。关键词：● 贝叶斯定理 ● 朴素贝叶斯分类 ● 贝叶斯回归 ● 贝叶斯派

分类原理

简单来说，朴素贝叶斯分类核心思想是比较后验概率大小。比如，对于二分类问题 ($K = 2$)，就是比较某点 $\boldsymbol{x}$ 处，后验概率 (posterior) $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 的大小。

后验概率 $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 本质上是条件概率 (conditional probability)。白话说，$f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 代表"给定 $\boldsymbol{x}$ 被分类为 C_1 的概率"，$f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 代表"给定 $\boldsymbol{x}$ 被分类为 C_2 的概率"。

如果 $f_{Y|\chi}(C_1 \mid \boldsymbol{x}) > f_{Y|\chi}(C_2 \mid \boldsymbol{x})$，$\boldsymbol{x}$ 被预测分类为 C_1；反之，$f_{Y|\chi}(C_1 \mid \boldsymbol{x}) < f_{Y|\chi}(C_2 \mid \boldsymbol{x})$，$\boldsymbol{x}$ 就被预测分类为 C_2。倘若 $f_{Y|\chi}(C_1 \mid \boldsymbol{x}) = f_{Y|\chi}(C_2 \mid \boldsymbol{x})$，该点便在决策边界 (decision boundary) 上。

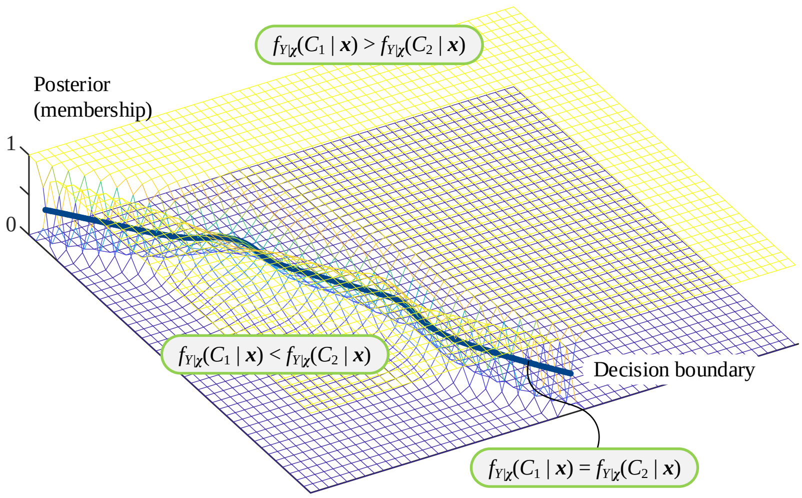

图 1. 二分类，比较后验概率大小，基于 KDE

比较图 1 所示 $f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 两个曲面。大家肯定已经发现，$f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 的取值在 [0, 1] 之间。实际上，$f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 并不是概率密度，它们本身就是概率。《概率统计》一册几次强调过这一点。

根据 $f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 两个曲面高度值，即概率值，我们可以确定决策边界 (图 1 中深蓝色实线)。

此外，对于二分类问题，$f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 之和为 1，下面简单证明一下。

全概率定理、贝叶斯定理

对于二分类问题，根据全概率定理 (law of total probability) 和贝叶斯定理 (Bayes' theorem)，$f_{\chi}(\boldsymbol{x})$ 可以通过下式计算得到：

$$
\begin{aligned}
f_{\chi}(\boldsymbol{x}) &= f_{\chi,Y}(\boldsymbol{x},C_1) + f_{\chi,Y}(\boldsymbol{x},C_2) \\
&= f_{Y|\chi}(C_1|\boldsymbol{x})\, f_{\chi}(\boldsymbol{x}) + f_{Y|\chi}(C_2|\boldsymbol{x})\, f_{\chi}(\boldsymbol{x})
\end{aligned}
\tag{1}
$$

$f_{\chi}(\boldsymbol{x})$ 不为 0 时，(1) 左右消去 $f_{\chi}(\boldsymbol{x})$，得到：

$$
1 = f_{Y|\chi}(C_1|\boldsymbol{x}) + f_{Y|\chi}(C_2|\boldsymbol{x}) \tag{2}
$$

白话解释，对于二分类问题，某点 $\boldsymbol{x}$ 要么属于 C_1，要么属于 C_2。

成员值：比较大小

后验概率值 $f_{Y|\chi}(C_1\,|\,\boldsymbol{x})$ 和 $f_{Y|\chi}(C_2\,|\,\boldsymbol{x})$ 取值在 [0, 1] 之间，且满足 (2)；因此，后验概率也常被称作成员值 (membership score)。

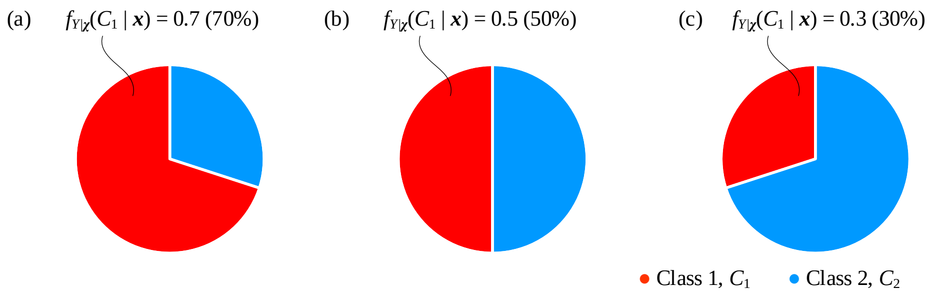

图 2. 二分类成员值

如图 2 (a) 所示，$f_{Y|\chi}(C_1\,|\,\boldsymbol{x}) = 0.7$ (70%)，也就是说 $\boldsymbol{x}$ 属于 C_1 的可能性为 70%，即成员值为 0.7。这种情况，$\boldsymbol{x}$ 预测分类为 C_1。

$f_{Y|\chi}(C_1\,|\,\boldsymbol{x}) = 0.5$ (50%) 时，对于二分类问题，$\boldsymbol{x}$ 应该位于决策边界上，如图 2 (b) 所示。

若 $f_{Y|\chi}(C_1\,|\,\boldsymbol{x}) = 0.3$ (30%)，$\boldsymbol{x}$ 属于 C_1 成员值为 0.3。显然，$\boldsymbol{x}$ 应该被预测分类为 C_2，如图 2 (c) 所示。

仅对于二分类问题，如果 $f_{Y|\chi}(C_1 \mid \boldsymbol{x}) > 0.5$，可以预测 $\boldsymbol{x}$ 分类为 C_1。

联合概率：比较大小

根据贝叶斯定理，对于二分类问题，证据因子 $f_\chi(\boldsymbol{x})$ 不为 0 时，后验概率 $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 为：

$$\begin{cases} \underbrace{f_{Y|\chi}\left(C_1 \mid \boldsymbol{x}\right)}_{\text{Posterior}} = \dfrac{\overbrace{f_{\chi,Y}\left(\boldsymbol{x}, C_1\right)}^{\text{Joint}}}{\underbrace{f_\chi\left(\boldsymbol{x}\right)}_{\text{Evidence}}} \\ \underbrace{f_{Y|\chi}\left(C_2 \mid \boldsymbol{x}\right)}_{\text{Posterior}} = \dfrac{\overbrace{f_{\chi,Y}\left(\boldsymbol{x}, C_2\right)}^{\text{Joint}}}{\underbrace{f_\chi\left(\boldsymbol{x}\right)}_{\text{Evidence}}} \end{cases} \tag{3}$$

观察 (3)，发现分母上均为证据因子 $f_\chi(\boldsymbol{x})$。这说明，后验概率 $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 正比于联合概率 (joint probability, joint) $f_{Y,\chi}(C_1, \boldsymbol{x})$ 和 $f_{Y,\chi}(C_2, \boldsymbol{x})$，即：

$$\begin{cases} \underbrace{f_{Y|\chi}\left(C_1 \mid \boldsymbol{x}\right)}_{\text{Posterior}} \propto \underbrace{f_{\chi,Y}\left(\boldsymbol{x}, C_1\right)}_{\text{Joint}} \\ \underbrace{f_{Y|\chi}\left(C_2 \mid \boldsymbol{x}\right)}_{\text{Posterior}} \propto \underbrace{f_{\chi,Y}\left(\boldsymbol{x}, C_2\right)}_{\text{Joint}} \end{cases} \tag{4}$$

也就是说，对于二分类问题，比较联合概率 $f_{Y,\chi}(C_1, \boldsymbol{x})$ 和 $f_{Y,\chi}(C_2, \boldsymbol{x})$ 大小，便可以预测分类！

图 3 给出的是某个二分类问题中，联合概率 $f_{Y,\chi}(C_1, \boldsymbol{x})$ 和 $f_{Y,\chi}(C_2, \boldsymbol{x})$ 两个曲面。通过比较 $f_{Y,\chi}(C_1, \boldsymbol{x})$ 和 $f_{Y,\chi}(C_2, \boldsymbol{x})$ 两个曲面高度，我们可以得出和图 1 一样的分类结论。

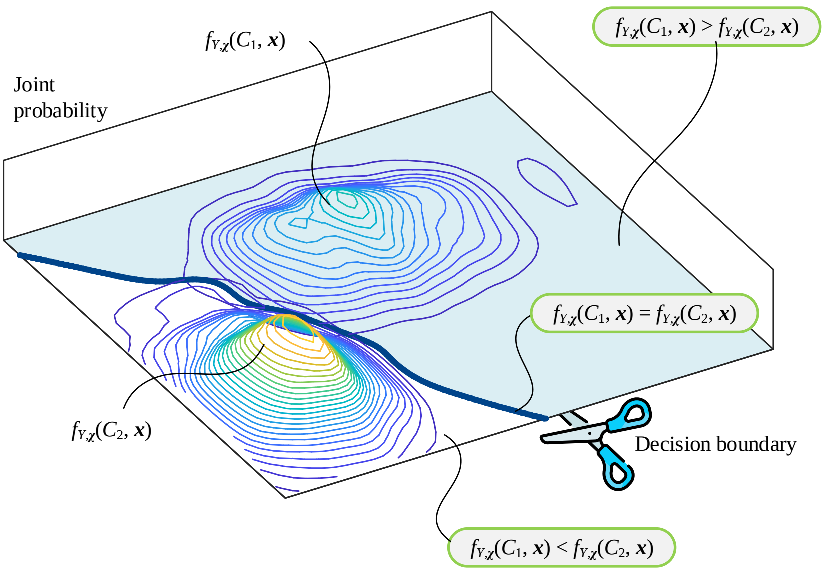

图 3. 二分类，比较联合概率大小，基于 KDE

推广：从二分类到多分类

根据前文分析，我们可以总结得到朴素贝叶斯分类优化问题——最大化后验概率：

$$\hat{y} = \arg\max_{C_k} f_{Y|\chi}\left(C_k \middle| \boldsymbol{x}\right) \tag{5}$$

其中，$k = 1,2,\dots K$。

证据因子 $f_\chi(\boldsymbol{x})$ 不为 0 时，后验概率正比于联合概率，即：

$$\underbrace{f_{Y|\chi}\left(C_k \middle| \boldsymbol{x}\right)}_{\text{Posterior}} \propto \underbrace{f_{\chi,Y}\left(\boldsymbol{x}, C_k\right)}_{\text{Joint}} \tag{6}$$

因此，(5) 等价于：

$$\hat{y} = \arg\max_{C_k} f_{\chi,Y}\left(\boldsymbol{x}, C_k\right) \tag{7}$$

也就是，贝叶斯分类中，“最大化后验概率”等价于“最大化联合概率”。

至此，我们解决了朴素贝叶斯分类的“贝叶斯”部分，下一节讨论何谓“朴素”。

阅读这一节感到吃力的话，请大家回顾《概率统计》最后两章内容。

4.2 朴素贝叶斯的“朴素”之处

朴素贝叶斯分类，何以谓之“朴素”？

本章副标题已经给出答案——假设特征之间条件独立 (conditional independence)！

注意，“特征条件独立”不同于“特征独立”。

特征独立

对于 x_1 和 x_2 两特征情况，“特征独立”指的是：

$$f_\chi\left(\boldsymbol{x}\right) = f_{X1,X2}\left(x_1, x_2\right) = f_{X1}\left(x_1\right) f_{X2}\left(x_2\right) \tag{8}$$

$f_{X1}(x_1)$ 和 $f_{X2}(x_2)$ 为两个特征上的边际概率密度函数，如图 4 所示。

推广到 D 个特征情况，“特征独立”指的是：

$$f_{\chi}(\boldsymbol{x}) = f_{X1}(x_1)\, f_{X2}(x_2) \ldots f_{X_D}(x_D) = \prod_{j=1}^{D} f_{X_j}(x_j) \tag{9}$$

图 4 中等高线为“特征独立”条件下，证据因子 $f_{\chi}(\boldsymbol{x})$ 概率密度分布。不知道大家看到这幅图时，是否想到《矩阵力量》中讲过的向量张量积。

$f_{X1}(x_1)$ 和 $f_{X2}(x_2)$ 描述 X_1 和 X_2 两特征的分布还比较准确。但是，假设特征独立，用 (8) 估算证据因子概率密度 $f_{\chi}(\boldsymbol{x})$ 时，偏差很大。比较图 4 等高线和散点分布就可以看出来。

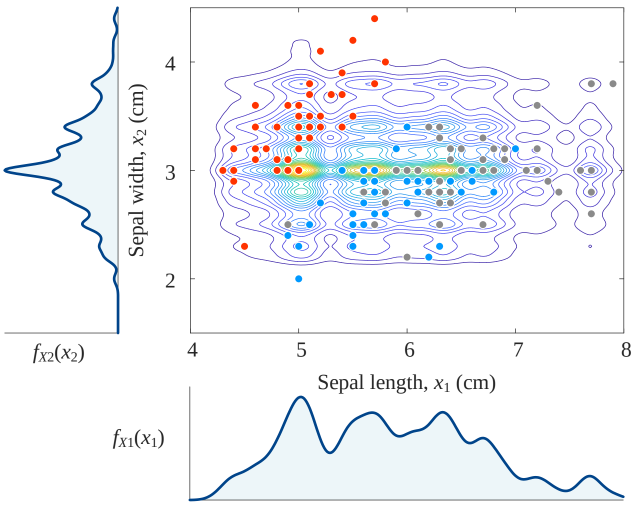

图 4. “特征独立”条件下，证据因子 $f_{\chi}(\mathbf{x})$ 概率密度，基于 KDE

特征条件独立

对于两特征 (D = 2)、两分类 (K = 2) 情况，“特征条件独立”指的是：

$$\begin{cases} \underbrace{f_{X1,X2|Y}(x_1,x_2|C_1)}_{\text{Likelihood}} = \underbrace{f_{X1|Y}(x_1|C_1)\, f_{X2|Y}(x_2|C_1)}_{\text{Conditional independence}} \\ \underbrace{f_{X1,X2|Y}(x_1,x_2|C_2)}_{\text{Likelihood}} = \underbrace{f_{X1|Y}(x_1|C_2)\, f_{X2|Y}(x_2|C_2)}_{\text{Conditional independence}} \end{cases} \tag{10}$$

推广到 D 个特征情况，“特征条件独立”假设下，似然概率为：

$$\underbrace{f_{\chi|Y}(\boldsymbol{x}|C_k)}_{\text{Likelihood}} = f_{X1|Y}(x_1|C_k)\, f_{X2|Y}(x_2|C_k) \ldots f_{X_D|Y}(x_D|C_k) = \prod_{j=1}^{D} f_{X_j|Y}(x_j|C_k) \tag{11}$$

⚠ 注意，“特征独立”，无法推导得到“特征条件独立”；同样，“特征条件独立”，无法推导得到“特征独立”。

特征条件独立 → 联合概率

根据贝叶斯定理，联合概率为：

$$\underbrace{f_{\chi,Y}\left(\boldsymbol{x},C_k\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}\underbrace{f_{\chi|Y}\left(\boldsymbol{x}|C_k\right)}_{\text{Likelihood}} \tag{12}$$

注意，先验概率 $p_Y(C_k)$ 为概率质量函数 (probability mass function, PMF)。这是因为 Y 是离散随机变量，Y 的取值为分类标签 C_1、$C_2 \dots C_K$，并非连续。

将 (11) 代入 (12)，可以得到“特征条件独立”条件下，联合概率为：

$$\underbrace{f_{\chi,Y}\left(\boldsymbol{x},C_k\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}\underbrace{f_{\chi|Y}\left(\boldsymbol{x}|C_k\right)}_{\text{Likelihood}}=\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}\underbrace{\prod_{j=1}^{D}f_{X_j|Y}\left(x_j|C_k\right)}_{\text{Conditional independence}} \tag{13}$$

“朴素”贝叶斯优化问题

有了本节分析，基于 (13)，(7) 所示朴素贝叶斯优化问题可以写成：

$$\hat{y}=\underset{C_k}{\arg\max}\ p_Y\left(C_k\right)\prod_{j=1}^{D}f_{X_j|Y}\left(x_j|C_k\right) \tag{14}$$

这样，我们便解决了“朴素贝叶斯”中的“朴素”部分！

朴素贝叶斯分类流程

图 5 所示为朴素贝叶斯分类流程图，图中散点数据为鸢尾花前两个特征——花萼长度、花萼宽度。

图 5 中概率密度基于核密度估计 (Kernel Density Estimation, KDE)，《概率统计》第 18 章介绍过 KDE 方法。

请大家现在快速浏览这幅图，完成本章学习之后，再回过头来再仔细观察图 5 细节。此外，本章内容和《概率统计》最后一章有重叠，建议大家回顾。

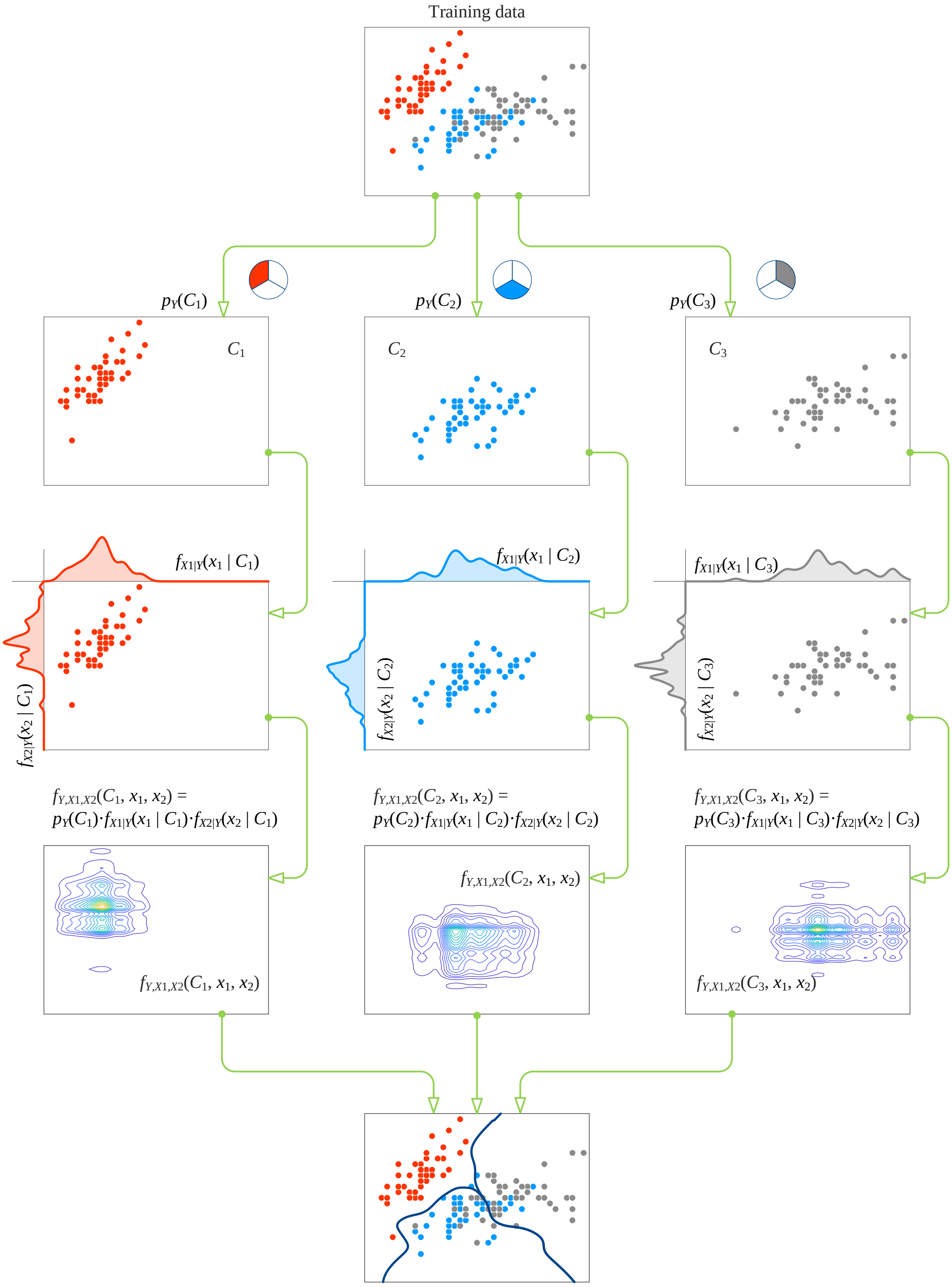

图 5. 朴素贝叶斯分类过程，基于 KDE

4.3 先验概率

先验概率计算最为简单。鸢尾花数据 C_1、C_2 和 C_3 三类对应的先验概率为：

$$p_Y(C_1) = \frac{\text{count}(C_1)}{\text{count}(\Omega)}, \quad p_Y(C_2) = \frac{\text{count}(C_2)}{\text{count}(\Omega)}, \quad p_Y(C_3) = \frac{\text{count}(C_3)}{\text{count}(\Omega)}, \tag{15}$$

鸢尾花数据共有 150 个数据点，count(Ω) = 150；而 C_1、C_2 和 C_3 三类各占 50，因此，

$$p_Y(C_1) = p_Y(C_2) = p_Y(C_3) = \frac{50}{150} = \frac{1}{3} \tag{16}$$

图 6 所示为鸢尾花数据先验概率结果。请注意，一般情况各类数据先验概率并不相等。

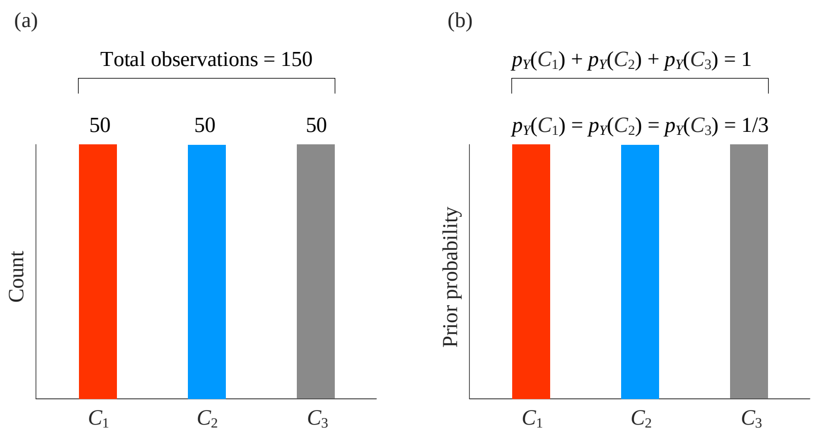

图 6. 鸢尾花数据先验概率

4.4 似然概率

根据前三节所述，朴素贝叶斯分类算法核心在于三方面：(1) 贝叶斯定理建立似然概率、先验概率和后验概率三者联系；(2) 估算似然概率时，假设特征之间条件独立；(3) 优化目标为，最大化后验概率，或最大化联合概率。

根据 (13)，想要获得联合概率，就先需要利用“特征条件独立”计算得到似然概率。

下面，我们利用花萼长度 (x_1) 和花萼宽度 (x_2) 两个特征 ($D = 2$)，解决鸢尾花三分类 ($K = 3$, C_1、C_2 和 C_3) 问题。本节先讨论如何获得 C_1、C_2 和 C_3 似然概率密度。

C_1 的似然概率

图 7 所示为求解似然概率密度 $f_{\chi|Y}(\boldsymbol{x} \mid C_1)$ 的过程。只考虑 setosa (C_1, $y = 0$) 样本数据点 ●，分别估算两个特征的条件边际分布 $f_{X1|Y}(x_1 \mid C_1)$ 和 $f_{X2|Y}(x_2 \mid C_1)$。

需要特别注意的是，图 7 中，$f_{X1|Y}(x_1 \mid C_1)$ 和 $f_{X2|Y}(x_2 \mid C_1)$ 曲线覆盖阴影区域面积均为 1。

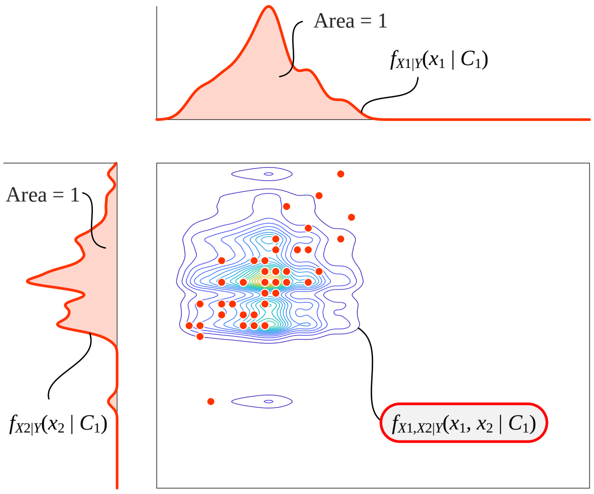

图 7. 分类 C_1 样本数据，鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率密度 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$

根据 (11)，似然概率 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 可以通过下式计算得到：

$$f_{\chi|Y}\left(\boldsymbol{x}|C_1\right) = f_{X1,X2|Y}\left(x_1,x_2|C_1\right) = f_{X1|Y}\left(x_1|C_1\right)\cdot f_{X2|Y}\left(x_2|C_1\right) \tag{17}$$

得到的 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 结果对应图 7 中等高线。而 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 曲面和水平面围成几何体的体积为 1，也就是说，$f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 在 $\mathbb{R}^2$ 的二重积分结果为 1，这个值是概率。而 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 的“偏积分”为条件边际分布 $f_{X1|Y}(x_1 \mid C_1)$ 或 $f_{X2|Y}(x_2 \mid C_1)$，它们还是概率密度，并非概率值。

《数学要素》第 14 章聊过“偏求和”，第 18 章聊过“偏积分”，建议大家回顾。

本章估算条件边际分布时用的是核密度估计方法。下一章则采用高斯分布 (Gaussian distribution) 来估算条件边际分布。因此，下一章的分类算法被称作，高斯朴素贝叶斯分类 (Gaussian Naïve Bayes classification)。

C_2 和 C_3 的似然概率

类似 (17)，C_2 和 C_3 似然概率可以通过下式估算得到：

$$\begin{cases} f_{X1,X2|Y}\left(x_1,x_2|C_2\right) = f_{X1|Y}\left(x_1|C_2\right)\cdot f_{X2|Y}\left(x_2|C_2\right) \\ f_{X1,X2|Y}\left(x_1,x_2|C_3\right) = f_{X1|Y}\left(x_1|C_3\right)\cdot f_{X2|Y}\left(x_2|C_3\right) \end{cases} \tag{18}$$

图 8 和图 9 等高线分别对应似然概率密度函数 $f_{X1,X2|Y}(x_1, x_2 | C_2)$ 和 $f_{X1,X2|Y}(x_1, x_2 | C_3)$ 结果。有了上一节的先验概率和本节得到的似然概率密度，我们可以求解联合概率。

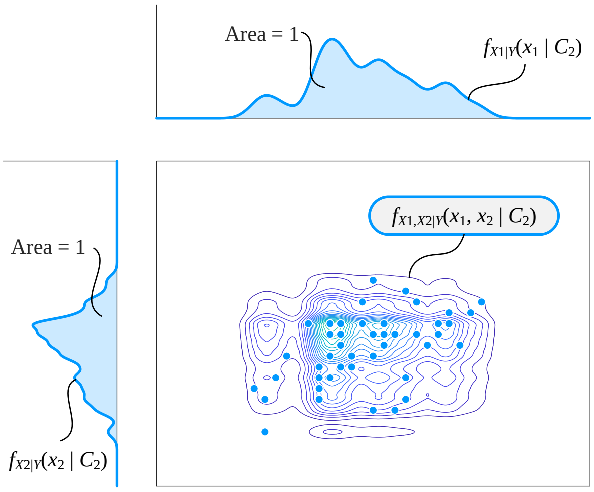

图 8. 分类 C_2 样本数据，鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率密度函数 $f_{X1,X2|Y}(x_1, x_2 | C_2)$

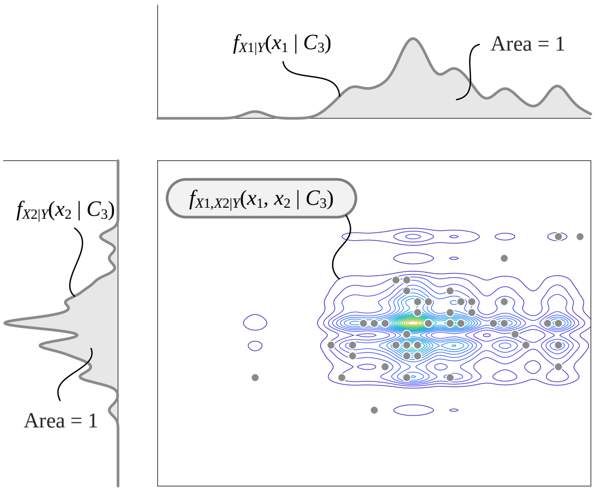

图 9. 分类 C_3 样本数据，鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率密度函数 $f_{X1,X2|Y}(x_1, x_2 | C_3)$

4.5 联合概率

C_1 的联合概率

根据 (13) 可以计算得到联合概率。对于鸢尾花三分类问题，假设“特征条件独立”，利用贝叶斯定理，联合概率 $f_{X1,X2,Y}(x_1, x_2, C_1)$ 可以通过下式得到：

$$\begin{aligned}\underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_1\right)}_{\text{Joint}} &= \underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_1\right)}_{\text{Likelihood}}\underbrace{p_Y\left(C_1\right)}_{\text{Prior}} \\ &= \underbrace{f_{X1|Y}\left(x_1|C_1\right)\cdot f_{X2|Y}\left(x_2|C_1\right)}_{\text{Conditional independence}}\underbrace{p_Y\left(C_1\right)}_{\text{Prior}}\end{aligned} \tag{19}$$

利用 (17)，我们已经得到似然概率密度曲面 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$。(16) 给出先验概率 $p_Y(C_1)$，代入 (19) 可以求得联合概率 $f_{X1,X2,Y}(x_1, x_2, C_1)$：

$$\underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_1\right)}_{\text{Joint}} = \underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_1\right)}_{\text{Likelihood}}\times\underbrace{\frac{1}{3}}_{\text{Prior}} \tag{20}$$

容易发现，先验概率 $p_Y(C_1) = 1/3$ 相当于一个缩放系数。

图 10 所示为联合概率 $f_{X1,X2,Y}(x_1, x_2, C_1)$ 概率密度曲面。图 10 的 z 轴数值为概率密度值，并非概率。

我们知道似然概率密度曲面 $f_{X1,X2|Y}(x_1, x_2 \mid C_1)$ 和水平面围成三维形状的体积为 1。而图 10 中联合概率 $f_{X1,X2,Y}(x_1, x_2, C_1)$ 和水平面围成体积为 $p_Y(C_1) = 1/3$。也就是说，$f_{X1,X2,Y}(x_1, x_2, C_1)$ 在 $\mathbb{R}^2$ 的二重积分结果为 1/3，这个值是概率值。

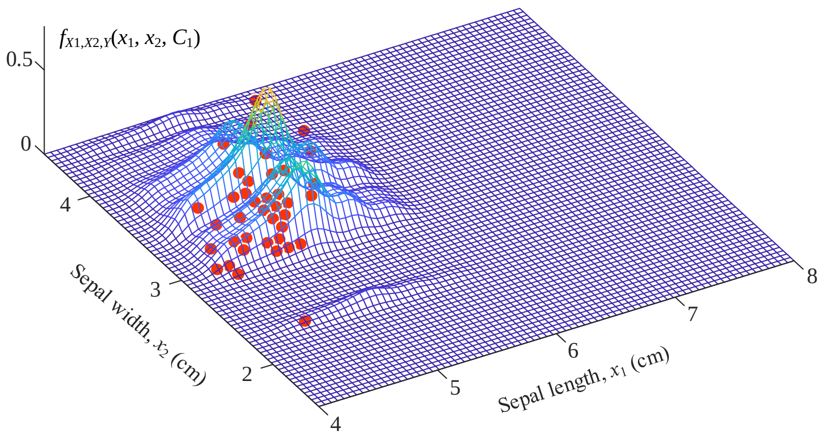

图 10. $f_{X1,X2,Y}(x_1, x_2, C_1)$ 概率密度曲面，基于 KDE

C_1 和 C_2 的联合概率

类似地，我们可以计算得到另外两个联合概率 $f_{X1,X2|Y}(x_1, x_2 \mid C_2)$ 和 $f_{X1,X2|Y}(x_1, x_2 \mid C_3)$，对应曲面分别如图 11 和图 12 所示。

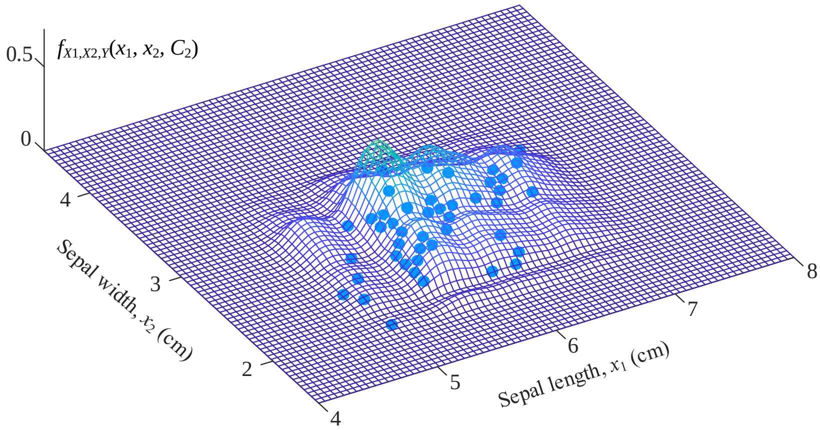

图 11. $f_{X1,X2,Y}(x_1, x_2, C_2)$ 概率密度曲面，基于 KDE

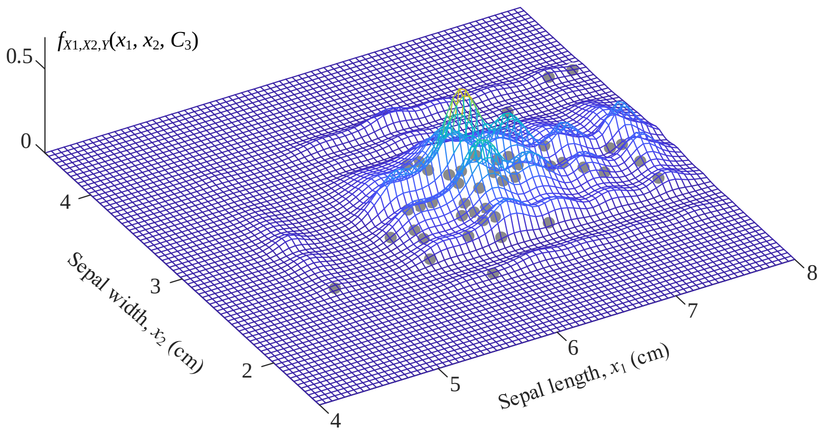

图 12. $f_{X1,X2,Y}(x_1, x_2, C_3)$ 概率密度曲面，基于 KDE

分类

至此，根据 (7) 我们可以比较上述三个联合概率密度曲面高度，从而获得决策边界。图 13 所示为采用朴素贝叶斯分类算法，基于 KDE 估算条件边际概率密度，得到的鸢尾花三分类边界。

请大家注意，目前 Python 的 Scikit-learn 工具包暂时不支持基于 KDE 的朴素贝叶斯分类。Scikit-learn 提供基于高斯分布的朴素贝叶斯分类器，这是下一章要介绍的内容。另外，KDE 朴素贝叶斯分类得到的决策边界不存在解析解。而高斯朴素贝叶斯分类得到的决策边界存在解析解。

利用 (7) 思想——比较联合概率大小——我们已经完成分类问题。但是，一般情况我们都会求出证据因子，并求得后验概率。如前文所述，后验概率又叫成员值，可以直接表达分类可能性百分比，便于可视化和解释结果。根据贝叶斯公式，要想得到后验概率，需要求得证据因子，这是下一节介绍的内容。

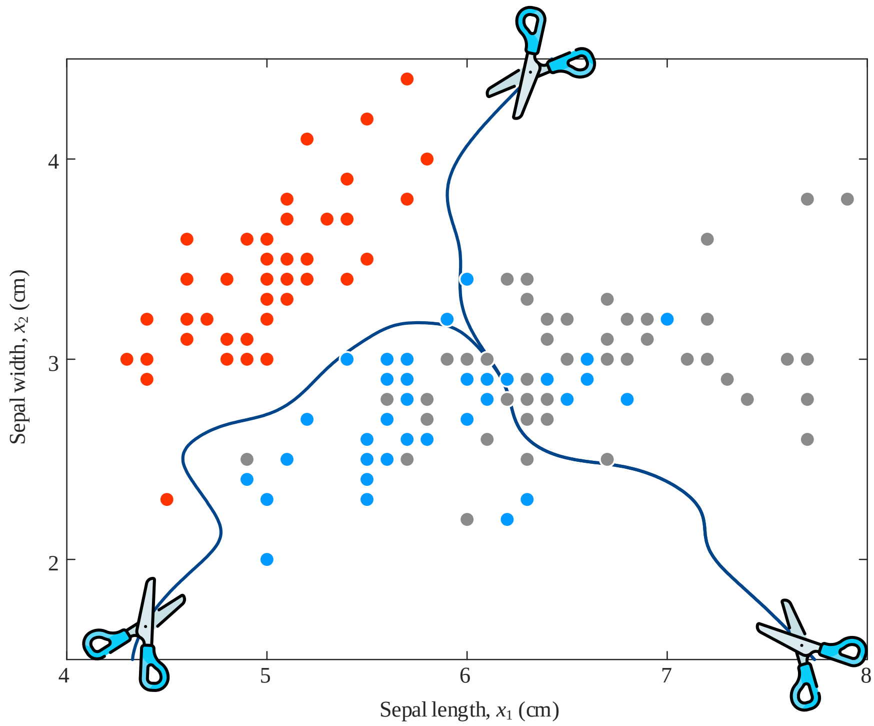

图 13. 朴素贝叶斯决策边界，基于核密度估计 KDE

4.6 证据因子

假设特征条件独立，利用全概率定理和 (13)，证据因子 $f_\chi(\boldsymbol{x})$ 概率密度可以通过下式计算得到：

$$\underbrace{f_\chi(\boldsymbol{x})}_{\text{Evidence}} = \sum_{k=1}^{K}\left\{\underbrace{f_{\chi,Y}(\boldsymbol{x}, C_k)}_{\text{Joint}}\right\} = \sum_{k=1}^{K}\left\{\underbrace{p_Y(C_k)}_{\text{Prior}}\underbrace{f_{\chi|Y}(\boldsymbol{x}|C_k)}_{\text{Likelihood}}\right\} = \sum_{k=1}^{K}\left\{\underbrace{p_Y(C_k)}_{\text{Prior}}\underbrace{\prod_{j=1}^{D} f_{X_j|Y}(x_j|C_k)}_{\text{Conditional independence}}\right\} \tag{21}$$

两特征、三分类问题

当 $K = 3$ 时，对于两特征分类问题，证据因子 $f_{X1,X2}(x_1, x_2)$ 可以利用下式求得：

$$\begin{aligned}\underbrace{f_{X1,X2}(x_1,x_2)}_{\text{Evidence}} &= \underbrace{f_{X1,X2,Y}(x_1,x_2,C_1)}_{\text{Joint}} + \underbrace{f_{X1,X2,Y}(x_1,x_2,C_2)}_{\text{Joint}} + \underbrace{f_{X1,X2,Y}(x_1,x_2,C_3)}_{\text{Joint}} \\ &= \underbrace{p_Y(C_1)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}(x_1,x_2|C_1)}_{\text{Likelihood}} + \underbrace{p_Y(C_2)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}(x_1,x_2|C_2)}_{\text{Likelihood}} + \underbrace{p_Y(C_3)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}(x_1,x_2|C_3)}_{\text{Likelihood}}\end{aligned} \tag{22}$$

这步计算很容易理解，对于鸢尾花数据，上一节得到的三个联合概率曲面 (图 10 ~ 图 12) 叠加便得到证据因子 $f_{X1,X2}(x_1, x_2)$ 概率密度曲面。图 14 所示为运算过程。图 14 实际上也是一种概率密度估算的方法。

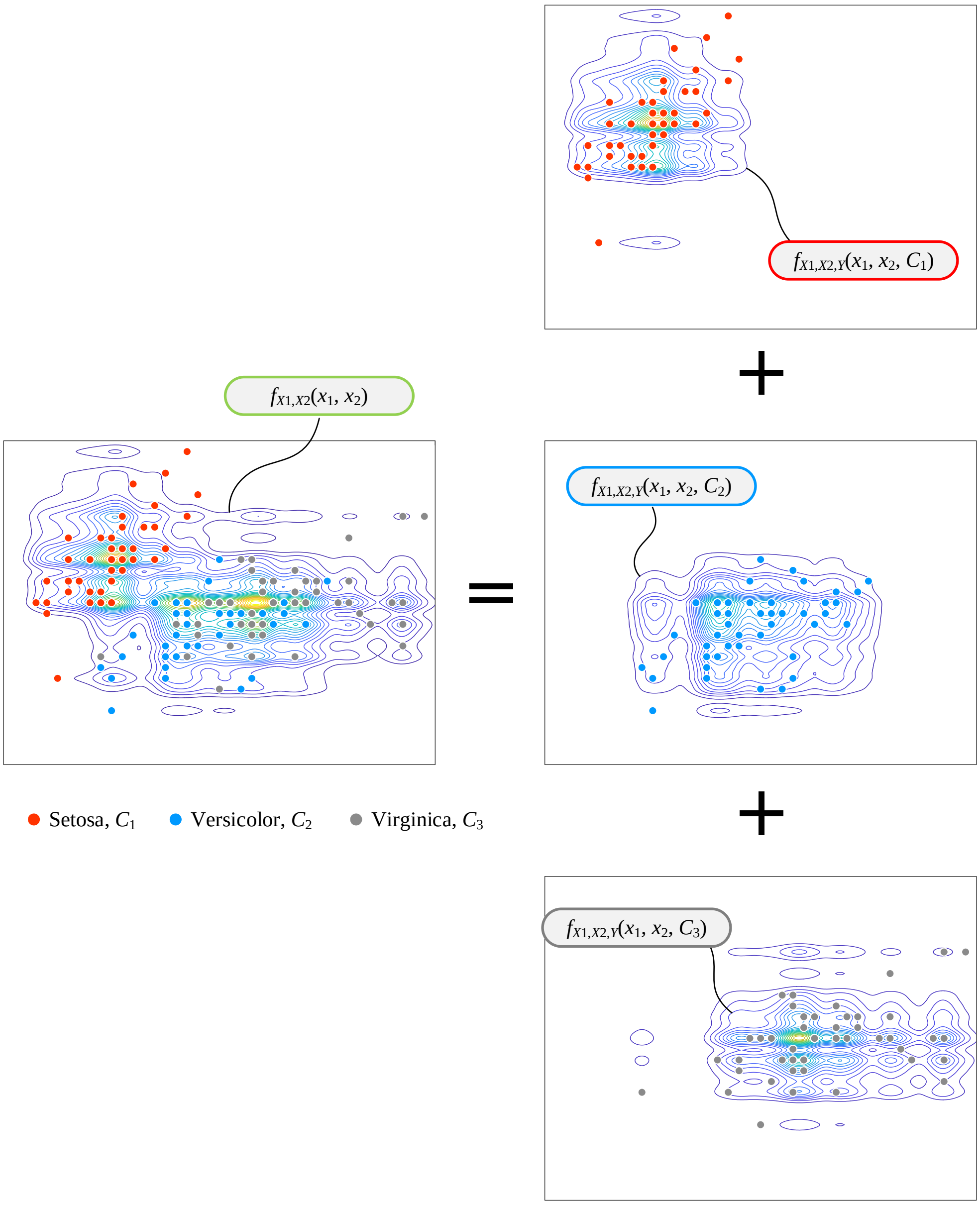

图 14. 估算证据因子概率密度，基于 KDE

概率密度估算

图 15 所示为利用"特征条件独立"构造得到的证据因子 $f_{X1,X2}(x_1, x_2)$ 概率密度曲面。$f_{X1,X2}(x_1, x_2)$ 概率密度曲面和水平面构成的几何形体体积为 1。

图 4 所示为假设“特征独立”条估算的证据因子概率密度曲面。前文提过，图 4 这个曲面没有准确捕捉样本数据分布特点；然而，图 15 曲面较为准确描述样本数据分布。

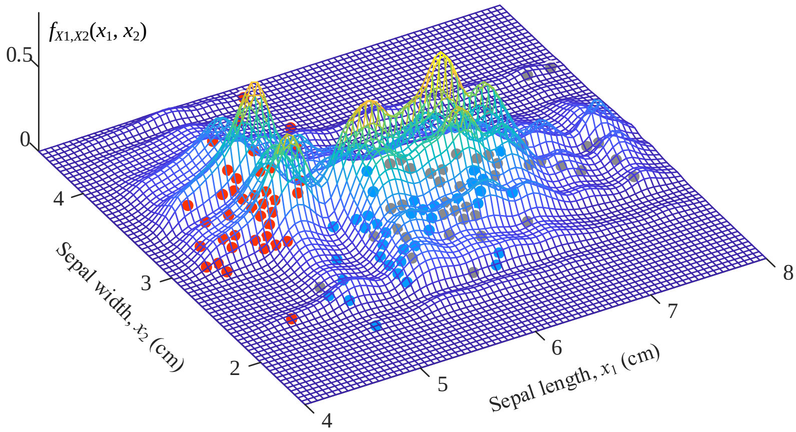

图 15. 估算得到的概率密度曲面，特征条件独立，基于 KDE

4.7 后验概率：成员值

有了前两节计算得到联合概率和证据因子，本节我们计算后验概率。

当 $K = 3$ 时，如果证据因子 $f_{X1,X2}(x_1, x_2)$ 不为 0，后验概率 $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 可以通过下式得到：

$$\underbrace{f_{Y|X1,X2}\left(C_1 \middle| x_1, x_2\right)}_{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1, x_2, C_1\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1, x_2\right)}_{\text{Evidence}}} \tag{23}$$

白话来讲，后验概率 $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 的含义是，给定 (x_1, x_2) 的具体值，分类标签为 C_1 的可能性多大？所以，$f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 并不是概率密度，$f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 是概率。

图 16 所示为后验概率 $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 曲面，容易发现曲面高度在 $[0, 1]$ 之间。

同理，可以计算得到另外两个后验概率 $f_{Y|X1,X2}(C_2 \mid x_1, x_2)$ 和 $f_{Y|X1,X2}(C_3 \mid x_1, x_2)$。比较三个后验概率曲面高度关系，可以得到和图 13 完全一致的决策边界。

对于三分类问题，后验概率 (成员值) 存在以下关系：

$$\underbrace{f_{Y|X1,X2}\left(C_1 \middle| x_1, x_2\right)}_{\text{Posterior}} + \underbrace{f_{Y|X1,X2}\left(C_2 \middle| x_1, x_2\right)}_{\text{Posterior}} + \underbrace{f_{Y|X1,X2}\left(C_3 \middle| x_1, x_2\right)}_{\text{Posterior}} = 1 \tag{24}$$

白话说，给定平面上任意一点 (x_1, x_2)，它的分类可能性只有三个——C_1、C_2、C_3。因此，上式中，三个条件概率之和为 1。

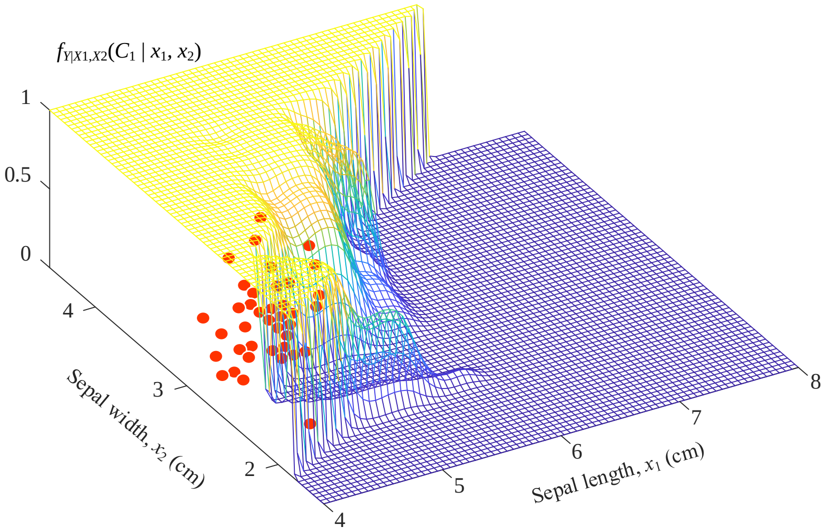

图 16. $f_{Y|X1,X2}(C_1 | x_1, x_2)$ 后验概率曲面，基于 KDE

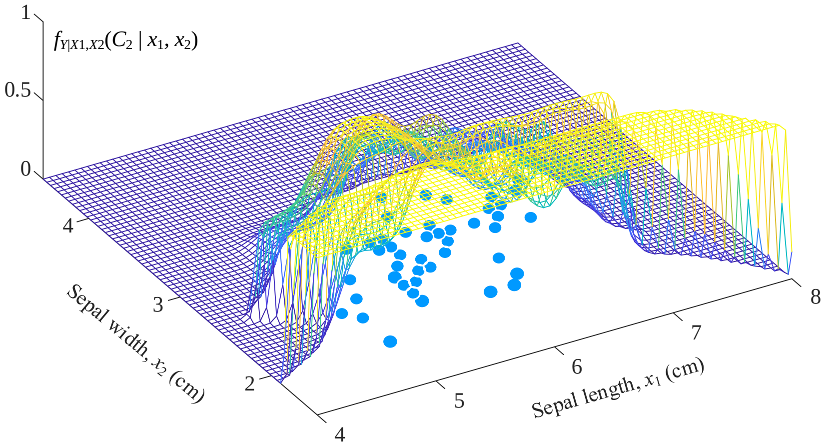

图 17. $f_{Y|X1,X2}(C_2 | x_1, x_2)$ 后验概率曲面，基于 KDE

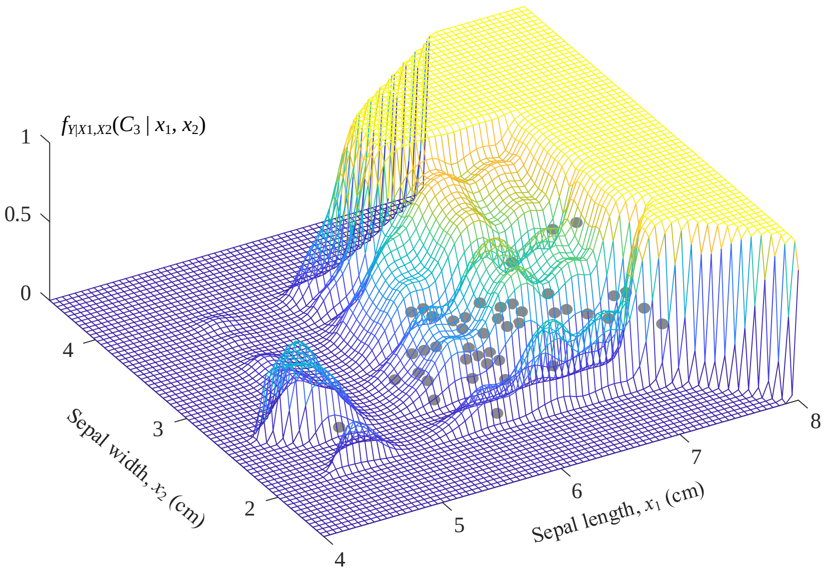

图 18. $f_{Y|X1,X2}(C_3 | x_1, x_2)$ 后验概率曲面，基于 KDE

本章最后请大家特别注意以下几点：

- 贝叶斯定理和全概率定理是朴素贝叶斯分类器的理论基础；
- 朴素贝叶斯分类器的“朴素”来自假设“特征条件独立”；
- 比较联合概率密度大小，可以预测分类；
- 假设“特征条件独立”，联合概率叠加得到证据因子，这是一种概率密度估算方法；
- 后验概率，本身就是概率值，取值范围在 [0, 1] 之间；
- 比较后验概率大小，同样可以预测分类。

下一章介绍高斯朴素贝叶斯。下一章采用和本章几乎一致的内容安排，请大家对照阅读。