

## 9

## Naive Bayes Classifier

## 朴素贝叶斯分类

假设特征之间条件独立，最大化后验概率



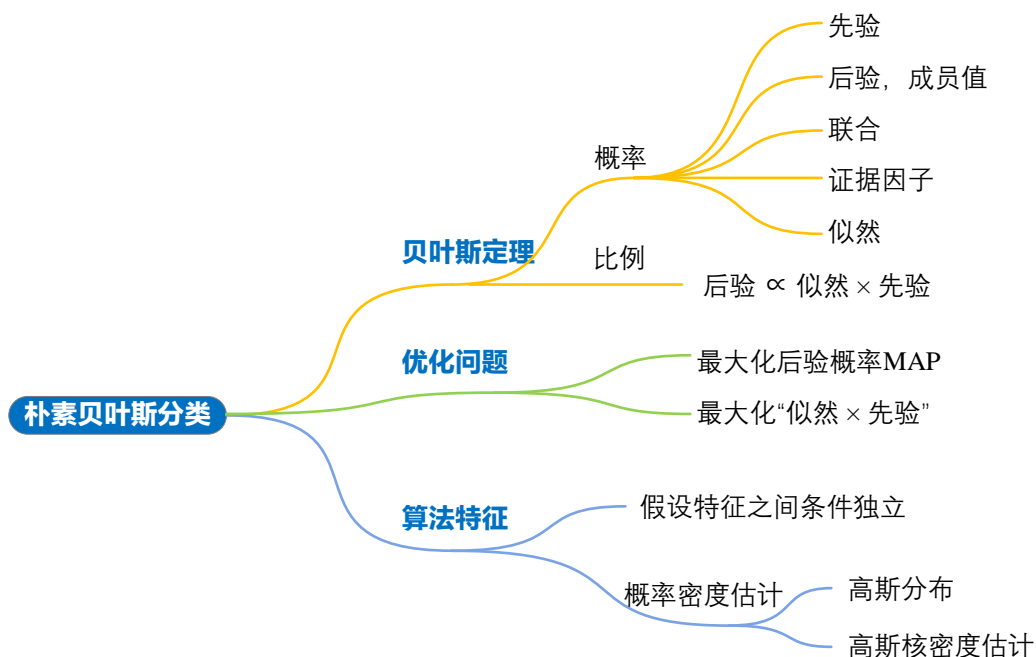
大家使用朴素贝叶斯分类器时，假设特征(条件)独立。之所以称之“朴素”，是因为那真是“天真”的假设。

*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
- ▶ matplotlib.pyplot.contour() 绘制等高线图
- ▶ matplotlib.pyplot.contourf() 绘制填充等高线图
- ▶ matplotlib.pyplot.scatter() 绘制散点图
- ▶ numpy.array() 创建 array 数据类型
- ▶ numpy.c\_() 按列叠加两个矩阵
- ▶ numpy.linspace() 产生连续均匀向量数值
- ▶ numpy.meshgrid() 创建网格化数据
- ▶ numpy.r\_() 按行叠加两个矩阵
- ▶ numpy.ravel() 将矩阵扁平化
- ▶ 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
- ▶ seaborn.scatterplot() 绘制散点图
- ▶ sklearn.datasets.load\_iris() 加载鸢尾花数据集
- ▶ sklearn.naive\_bayes.GaussianNB 高斯朴素贝叶斯分类算法函数



## 9.1 重逢贝叶斯



贝叶斯是我们的老朋友，《统计至简》薄频率派，厚贝叶斯派，这本书用了很大篇幅介绍了**贝叶斯定理** (Bayes' theorem) 和应用。《统计至简》第 18、19 章介绍贝叶斯分类的理论基础，第 20、21、22 章介绍贝叶斯统计推断。

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)。简单来说，朴素贝叶斯分类是一种基于贝叶斯定理和特征条件独立假设的分类方法，其原理是利用已知分类标记的训练数据，计算每个类别的条件概率分布，并根据贝叶斯定理计算未知样本属于每个类别的后验概率，最终将样本分配给具有最高概率的类别。其优点包括算法简单、计算高效、在处理大规模数据时表现良好，适用于多分类问题和高维数据；缺点是对特征的条件独立性要求较高，可能导致分类准确度下降，同时对于连续型变量的处理也存在一定困难。

本章和下一章共用一个思维导图。



**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

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  $\mathbf{x}$  处，**后验概率** (posterior)  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  的大小。

后验概率  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  本质上是**条件概率** (conditional probability)。白话说， $f_{Y|X}(C_1 | \mathbf{x})$  代表“给定  $\mathbf{x}$  被分类为  $C_1$  的概率”， $f_{Y|X}(C_2 | \mathbf{x})$  代表“给定  $\mathbf{x}$  被分类为  $C_2$  的概率”。

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

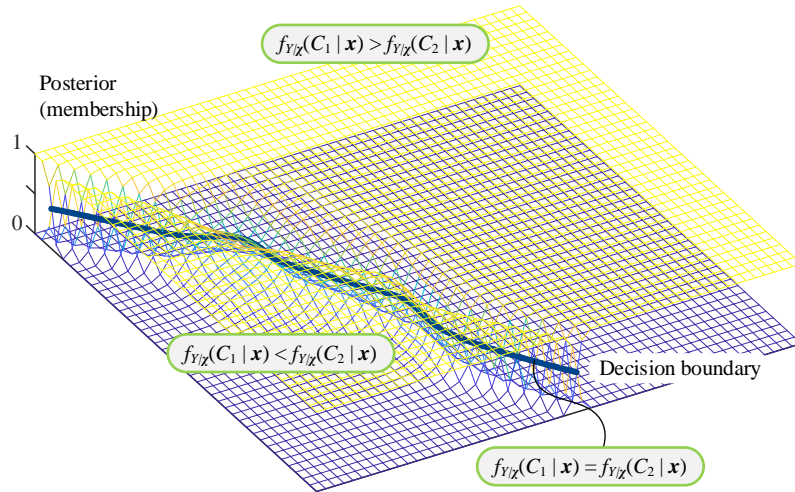


图 1. 二分类，比较后验概率大小，基于 KDE

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

根据  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  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， $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  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)， $f_X(\mathbf{x})$  可以通过下式计算得到：

$$\begin{aligned} f_X(\mathbf{x}) &= f_{X,Y}(\mathbf{x}, C_1) + f_{X,Y}(\mathbf{x}, C_2) \\ &= f_{Y|X}(C_1 | \mathbf{x}) f_X(\mathbf{x}) + f_{Y|X}(C_2 | \mathbf{x}) f_X(\mathbf{x}) \end{aligned} \quad (1)$$

$f_X(\mathbf{x})$  不为 0 时，(1) 左右消去  $f_X(\mathbf{x})$ ，得到：

$$1 = f_{Y|X}(C_1 | \mathbf{x}) + f_{Y|X}(C_2 | \mathbf{x}) \quad (2)$$

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

### 成员值：比较大小

后验概率值  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  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)。

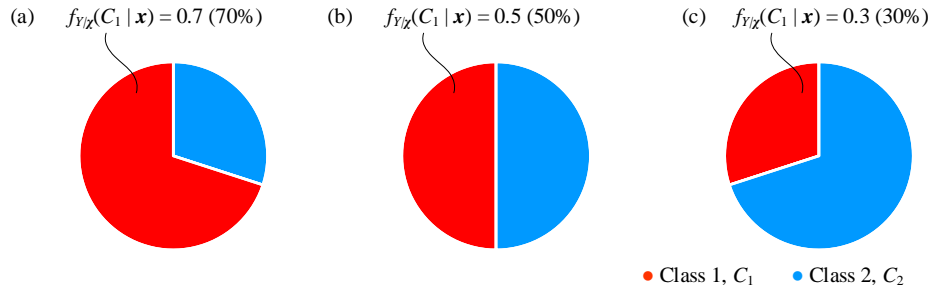


图 2. 二分类成员值

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

$f_{Y|X}(C_1 | \mathbf{x}) = 0.5$  (50%) 时, 对于二分类问题,  $\mathbf{x}$  应该位于决策边界上, 相当于“骑墙派”, 如图 2 (b) 所示。

若  $f_{Y|X}(C_1 | \mathbf{x}) = 0.3$  (30%),  $\mathbf{x}$  属于  $C_1$  成员值为 0.3。显然,  $\mathbf{x}$  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  $f_{Y|X}(C_1 | \mathbf{x}) > 0.5$ , 可以预测  $\mathbf{x}$  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  $f_X(\mathbf{x})$  不为 0 时, 后验概率  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  为:

$$\left\{ \begin{array}{l} \underbrace{f_{Y|X}(C_1 | \mathbf{x})}_{\text{Posterior}} = \frac{\overbrace{f_{X,Y}(\mathbf{x}, C_1)}^{\text{Joint}}}{\underbrace{f_X(\mathbf{x})}_{\text{Evidence}}} \\ \underbrace{f_{Y|X}(C_2 | \mathbf{x})}_{\text{Posterior}} = \frac{\overbrace{f_{X,Y}(\mathbf{x}, C_2)}^{\text{Joint}}}{\underbrace{f_X(\mathbf{x})}_{\text{Evidence}}} \end{array} \right. \quad (3)$$

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

$$\left\{ \begin{array}{l} \underbrace{f_{Y|X}(C_1 | \mathbf{x})}_{\text{Posterior}} \propto \underbrace{f_{X,Y}(\mathbf{x}, C_1)}_{\text{Joint}} \\ \underbrace{f_{Y|X}(C_2 | \mathbf{x})}_{\text{Posterior}} \propto \underbrace{f_{X,Y}(\mathbf{x}, C_2)}_{\text{Joint}} \end{array} \right. \quad (4)$$

对于二分类问题, 比较联合概率  $f_{X,Y}(C_1, \mathbf{x})$  和  $f_{X,Y}(C_2, \mathbf{x})$  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  $f_{X,Y}(C_1, \mathbf{x})$  和  $f_{X,Y}(C_2, \mathbf{x})$  两个曲面。通过比较  $f_{X,Y}(C_1, \mathbf{x})$  和  $f_{X,Y}(C_2, \mathbf{x})$  两个曲面高度, 我们可以得出和图 1 一样的分类结论。

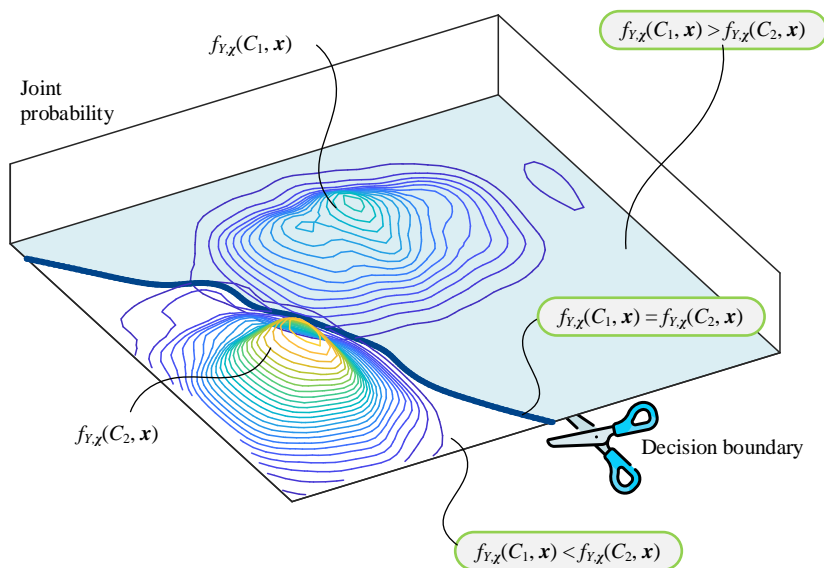


图 3. 二分类，比较联合概率大小，基于 KDE

### 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：

$$\hat{y} = \arg \max_{C_k} f_{Y|Z}(C_k | \mathbf{x}) \quad (5)$$

其中， $k = 1, 2, \dots, K$ 。

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

$$\underbrace{f_{Y|Z}(C_k | \mathbf{x})}_{\text{Posterior}} \propto \underbrace{f_{Z,Y}(\mathbf{x}, C_k)}_{\text{Joint}} \quad (6)$$

因此，(5) 等价于：

$$\hat{y} = \arg \max_{C_k} f_{Z,Y}(\mathbf{x}, C_k) \quad (7)$$

由于后验  $\propto$  似然  $\times$  先验，最大化后验概率等价于最大化“似然  $\times$  先验”。

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《统计至简》第 18、19 章内容。

## 9.2 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)!

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](mailto:jiang.visualize.ml@gmail.com)

独立指两个事件  $A$ 、 $B$  之间没有任何关联，即  $A$  的发生与  $B$  的发生互不影响，可以表示为  $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$ 。条件独立指在已知某些条件下，两个事件  $A$ 、 $B$  之间没有任何关联，比如给定条件  $C$  下， $A$  的发生与  $B$  的发生互不影响，可以表示为  $\Pr(A \cap B | C) = \Pr(A | C) \times \Pr(B | C)$ 。

## 特征独立

对于  $x_1$  和  $x_2$  两特征情况，“特征独立”指的是：

$$f_{\mathbf{x}}(\mathbf{x}) = f_{x_1, x_2}(x_1, x_2) = f_{x_1}(x_1) f_{x_2}(x_2) \quad (8)$$

$f_{x_1}(x_1)$  和  $f_{x_2}(x_2)$  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：

$$f_{\mathbf{x}}(\mathbf{x}) = f_{x_1}(x_1) f_{x_2}(x_2) \dots f_{x_D}(x_D) = \prod_{j=1}^D f_{x_j}(x_j) \quad (9)$$

图 4 中等高线为“特征独立”条件下，证据因子  $f_{\mathbf{x}}(\mathbf{x})$  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。

$f_{x_1}(x_1)$  和  $f_{x_2}(x_2)$  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  $f_{\mathbf{x}}(\mathbf{x})$  时，偏差很大。比较图 4 等高线和散点分布就可以看出来。

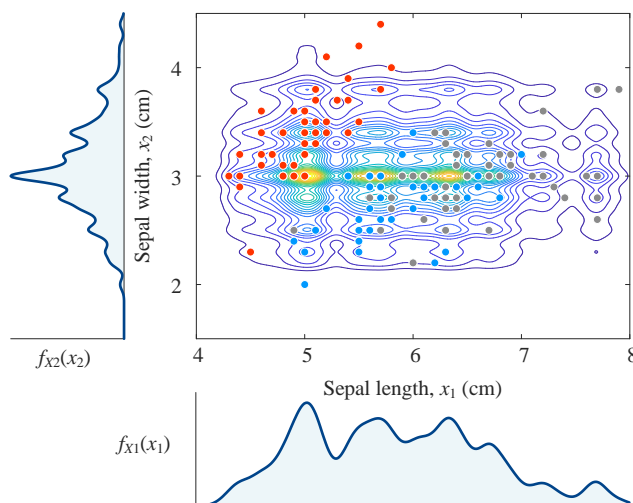


图 4. “特征独立”条件下，证据因子  $f_{\mathbf{x}}(\mathbf{x})$  概率密度，基于 KDE

## 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：

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

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](mailto:jiang.visualize.ml@gmail.com)



推广到  $D$  个特征情况，“特征条件独立”假设下，似然概率为：

$$\underbrace{f_{\mathbf{X}|Y}(\mathbf{x}|C_k)}_{\text{Likelihood}} = f_{X_1|Y}(x_1|C_k) f_{X_2|Y}(x_2|C_k) \dots f_{X_D|Y}(x_D|C_k) = \prod_{j=1}^D f_{X_j|Y}(x_j|C_k) \quad (11)$$

⚠ 请大家格外注意， $A$  和  $B$  相互独立，无法推导得到  $A$  和  $B$  条件独立。而  $A$  和  $B$  条件独立，也无法推导得到  $A$  和  $B$  相互独立。《统计至简》第 3 章专门介绍过条件独立。

### 特征条件独立 → 联合概率

根据贝叶斯定理，联合概率为：

$$\underbrace{f_{\mathbf{X},Y}(\mathbf{x}, C_k)}_{\text{Joint}} = \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{f_{\mathbf{X}|Y}(\mathbf{x}|C_k)}_{\text{Likelihood}} \quad (12)$$

⚠ 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：

$$\underbrace{f_{\mathbf{X},Y}(\mathbf{x}, C_k)}_{\text{Joint}} = \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{f_{\mathbf{X}|Y}(\mathbf{x}|C_k)}_{\text{Likelihood}} = \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{\prod_{j=1}^D f_{X_j|Y}(x_j|C_k)}_{\text{Conditional independence}} \quad (13)$$

### “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：

$$\hat{y} = \arg \max_{C_k} p_Y(C_k) \prod_{j=1}^D f_{X_j|Y}(x_j|C_k) \quad (14)$$

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)。《统计至简》第 17 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。



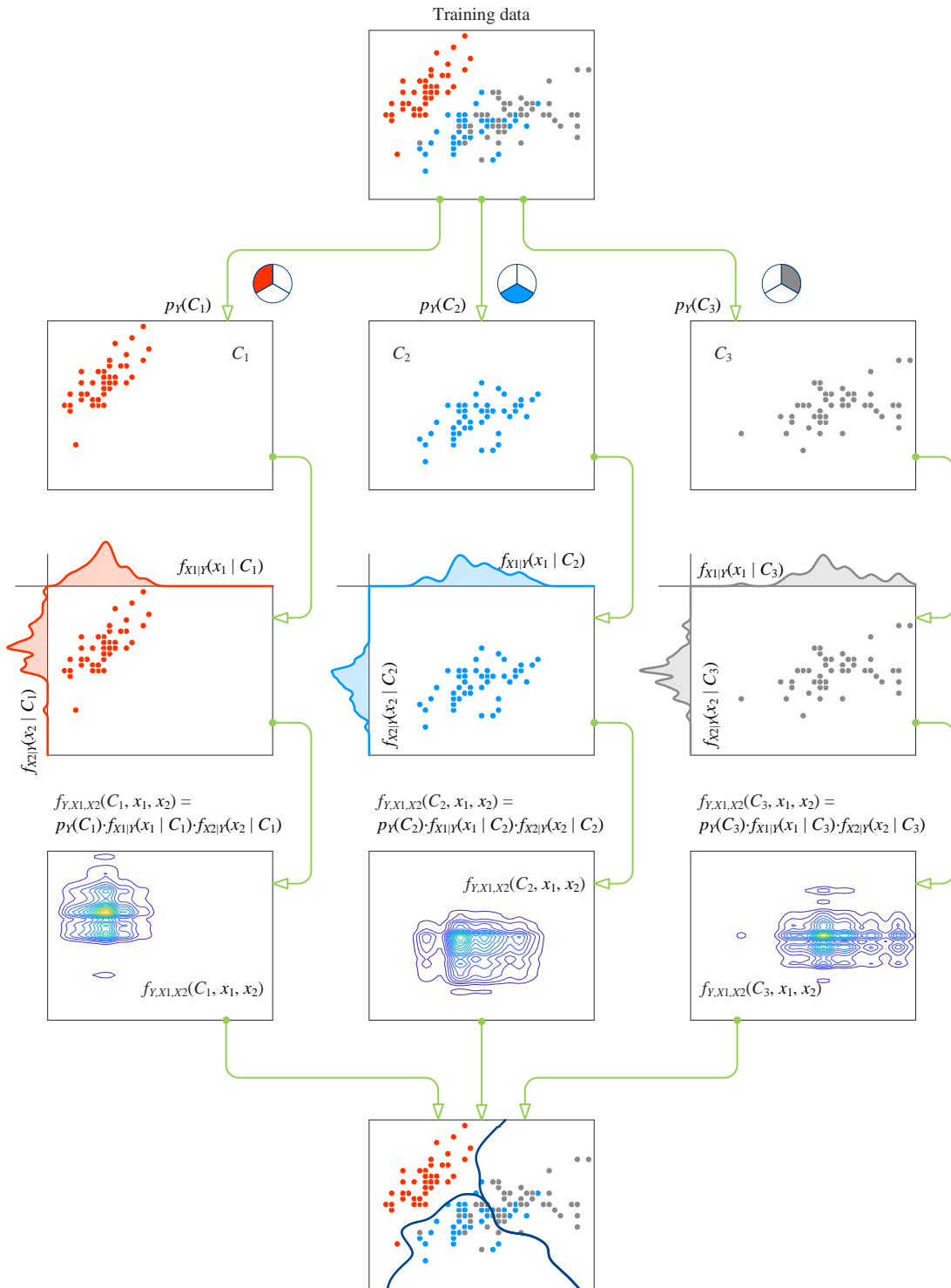


图 5. 朴素贝叶斯分类过程，基于 KDE

## 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：

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

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](mailto:jiang.visualize.ml@gmail.com)

鸢尾花数据共有 150 个数据点， $\text{count}(\Omega) = 150$ ；而  $C_1$ 、 $C_2$  和  $C_3$  三类各占 50，因此，

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

图 6 所示为鸢尾花数据先验概率结果。

⚠ 注意，一般情况各类数据先验概率并不相等。

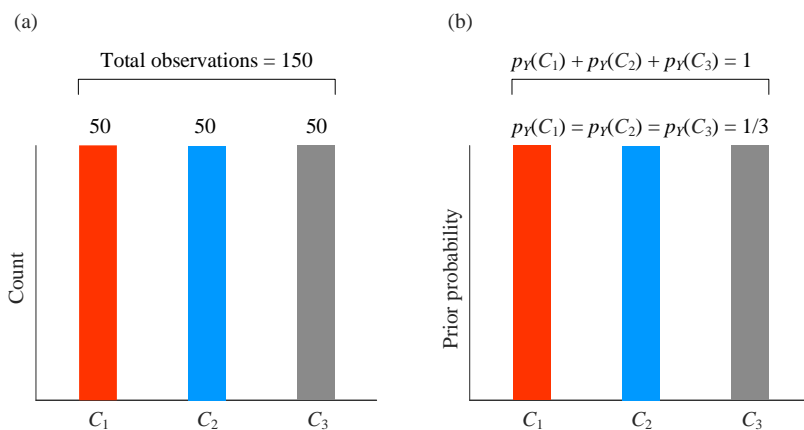


图 6. 鸢尾花数据先验概率

## 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 (似然  $\times$  先验)。

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  $f_{X|Y}(\mathbf{x} | C_1)$  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  $f_{X_1|Y}(x_1 | C_1)$  和  $f_{X_2|Y}(x_2 | C_1)$ 。

需要特别注意的是，图 7 中， $f_{X_1|Y}(x_1 | C_1)$  和  $f_{X_2|Y}(x_2 | C_1)$  曲线覆盖阴影区域面积均为 1。

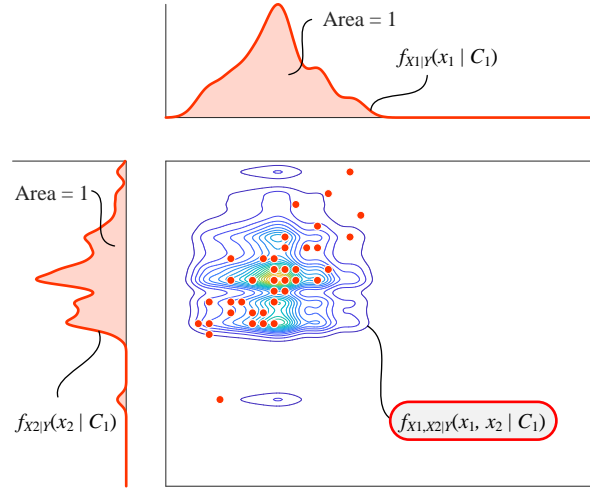


图 7. 分类  $C_1$  样本数据，鸢尾花花萼长度  $x_1$  和花萼宽度  $x_2$  条件独立，得到似然概率密度  $f_{X1,X2|Y}(x_1, x_2 | C_1)$

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

$$f_{X|Y}(\mathbf{x}|C_1) = f_{X1,X2|Y}(x_1, x_2 | C_1) = f_{X1|Y}(x_1 | C_1) \cdot f_{X2|Y}(x_2 | C_1) \quad (17)$$

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

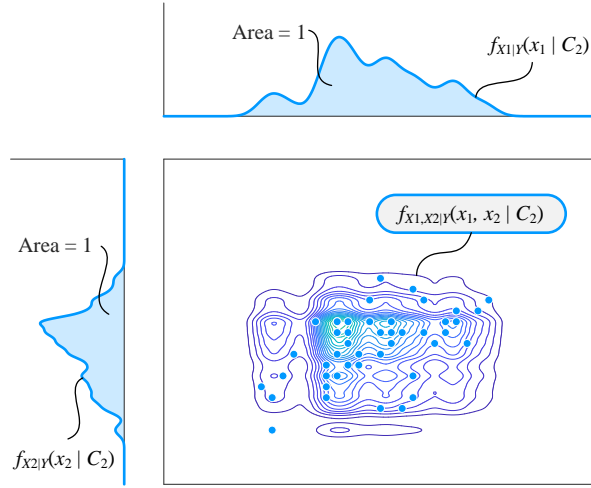
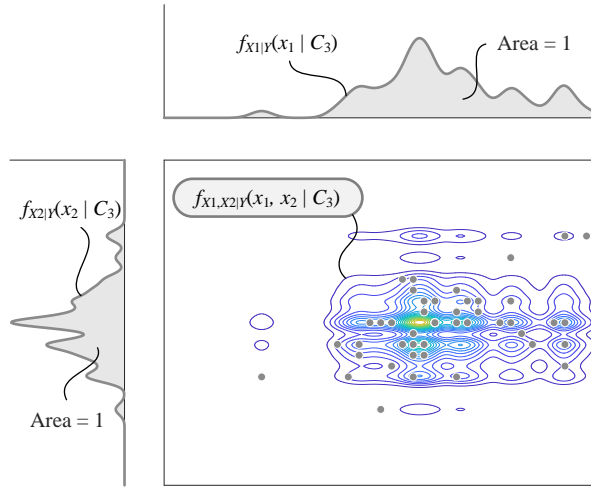
本章估算条件边际分布时用的是高斯核密度估计方法。下一章则采用高斯分布 (Gaussian distribution) 来估算条件边际分布。因此，下一章的分类算法被称作，**高斯朴素贝叶斯分类** (Gaussian Naïve Bayes classification)。

## $C_2$ 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：

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

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  $f_{X1,X2|Y}(x_1, x_2 | C_3)$ 

## 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：

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

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

$$\underbrace{f_{X_1, X_2, Y}(x_1, x_2, C_1)}_{\text{Joint}} = \underbrace{f_{X_1, X_2 | Y}(x_1, x_2 | C_1)}_{\text{Likelihood}} \times \underbrace{\frac{1}{3}}_{\text{Prior}} \quad (20)$$

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  $f_{X_1, X_2, Y}(x_1, x_2, C_1)$  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  $f_{X_1, X_2 | Y}(x_1, x_2 | C_1)$  和水平面围成三维形状的体积为 1。而图 10 中联合概率  $f_{X_1, X_2, Y}(x_1, x_2, C_1)$  和水平面围成体积为  $p_Y(C_1) = 1/3$ 。也就是说， $f_{X_1, X_2, Y}(x_1, x_2, C_1)$  在  $\mathbb{R}^2$  的二重积分结果为  $1/3$ ，这个值是概率值。

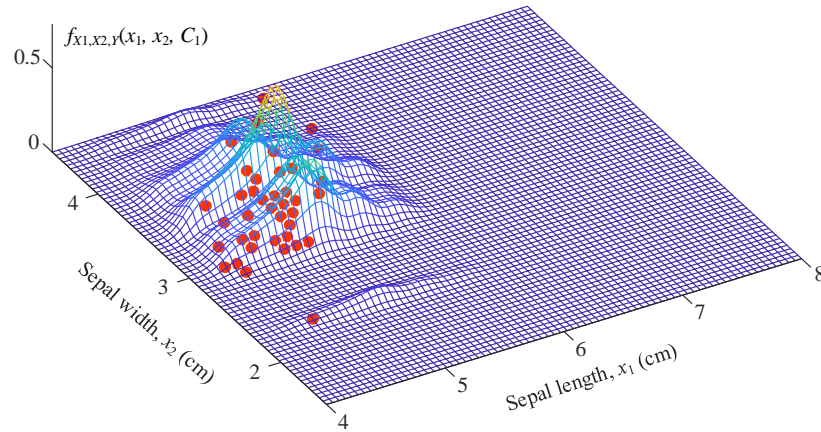


图 10.  $f_{X_1, X_2, Y}(x_1, x_2, C_1)$  概率密度曲面，基于 KDE

### $C_1$ 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  $f_{X_1, X_2 | Y}(x_1, x_2 | C_2)$  和  $f_{X_1, X_2 | Y}(x_1, x_2 | C_3)$ ，对应曲面分别如图 11 和图 12 所示。

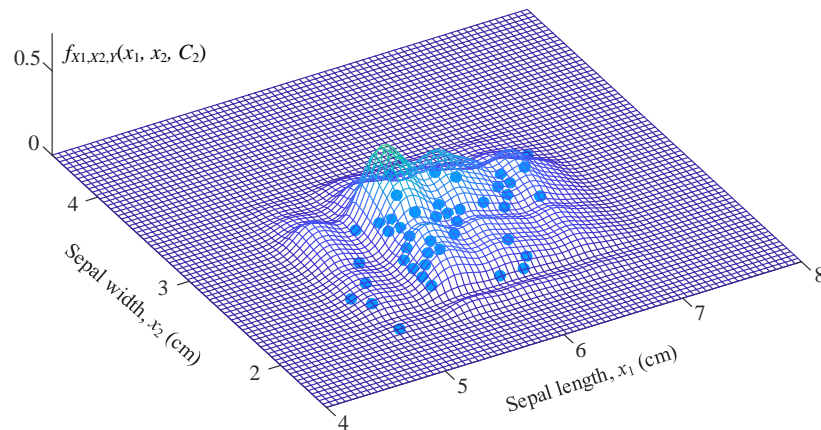
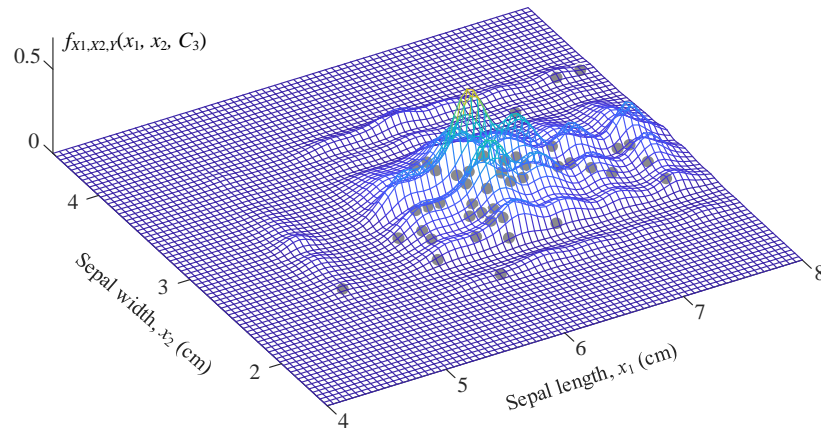


图 11.  $f_{X_1, X_2, Y}(x_1, x_2, C_2)$  概率密度曲面，基于 KDE

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。

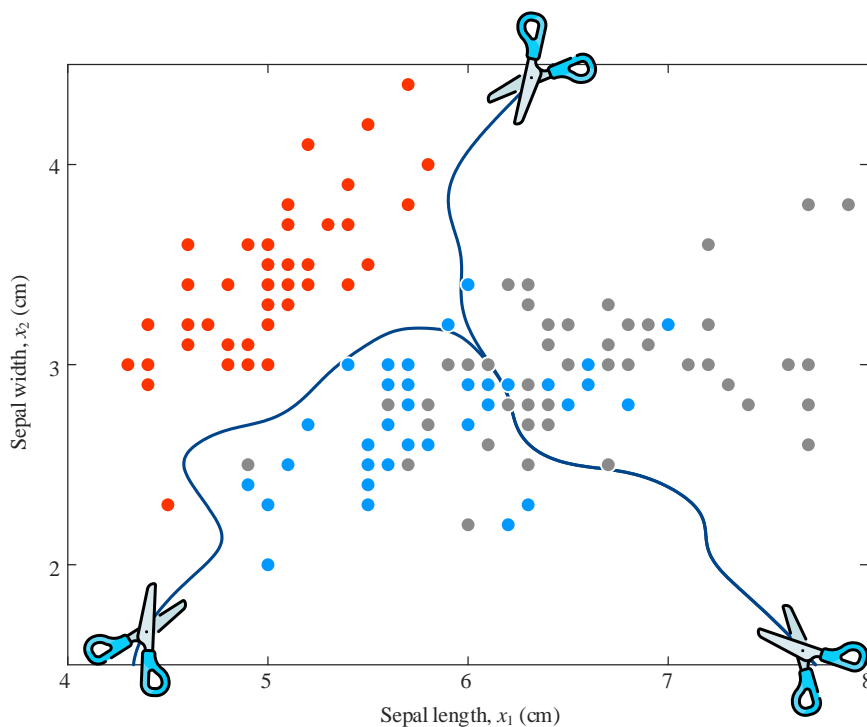


图 13. 朴素贝叶斯决策边界，基于核密度估计 KDE

## 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  $f_X(\mathbf{x})$  概率密度可以通过下式计算得到：

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

## 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：

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

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。



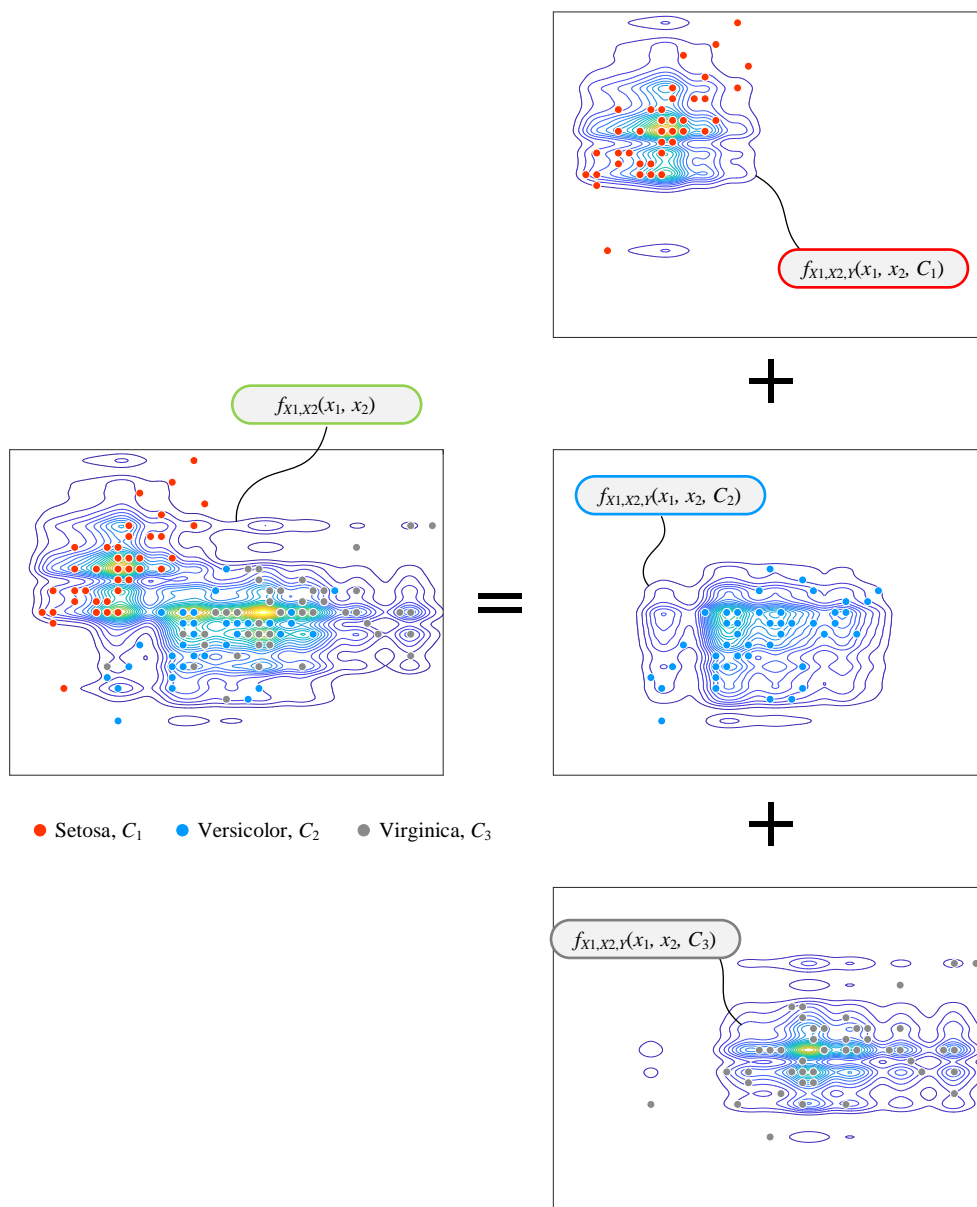


图 14. 估算证据因子概率密度，基于 KDE

## 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。

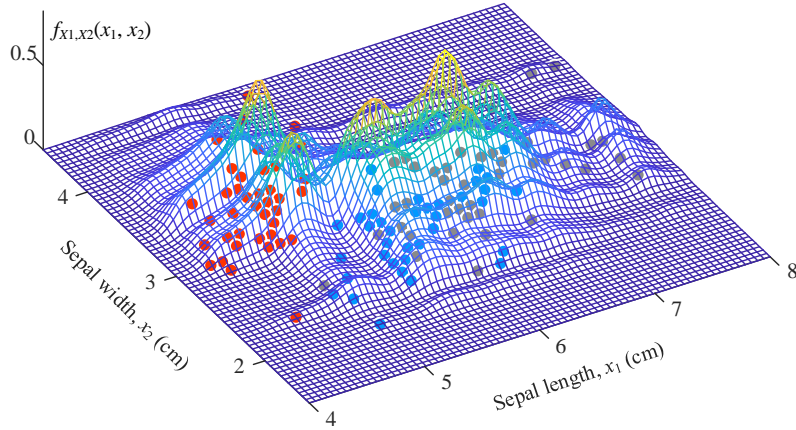


图 15. 估算得到的概率密度曲面，特征条件独立，基于 KDE

### 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  $f_{X1,X2}(x1, x2)$  不为 0，后验概率  $f_{Y|X1,X2}(C1 | x1, x2)$  可以通过下式得到：

$$\underbrace{f_{Y|X1,X2}(C1 | x1, x2)}_{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}(x1, x2, C1)}^{\text{Joint}}}{\underbrace{f_{X1,X2}(x1, x2)}_{\text{Evidence}}} \quad (23)$$

白话来讲，后验概率  $f_{Y|X1,X2}(C1 | x1, x2)$  的含义是，给定  $(x1, x2)$  的具体值，分类标签为  $C1$  的可能性多大？所以， $f_{Y|X1,X2}(C1 | x1, x2)$  并不是概率密度， $f_{Y|X1,X2}(C1 | x1, x2)$  是概率。

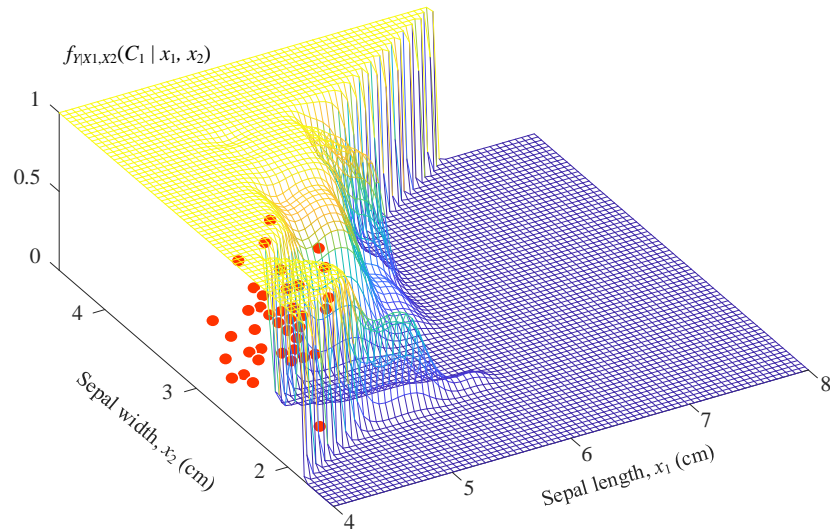
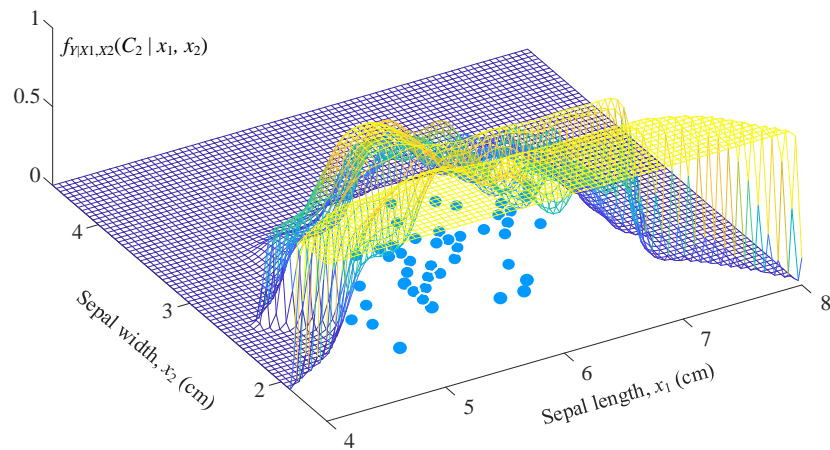
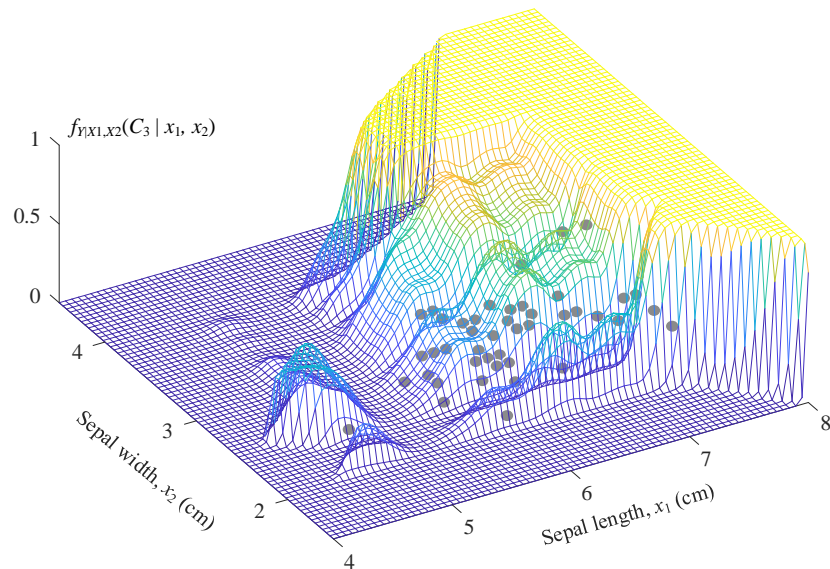
图 16 所示为后验概率  $f_{Y|X1,X2}(C1 | x1, x2)$  曲面，容易发现曲面高度在  $[0, 1]$  之间。

同理，可以计算得到另外两个后验概率  $f_{Y|X1,X2}(C2 | x1, x2)$  和  $f_{Y|X1,X2}(C3 | x1, x2)$ 。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：

$$\underbrace{f_{Y|X1,X2}(C1 | x1, x2)}_{\text{Posterior}} + \underbrace{f_{Y|X1,X2}(C2 | x1, x2)}_{\text{Posterior}} + \underbrace{f_{Y|X1,X2}(C3 | x1, x2)}_{\text{Posterior}} = 1 \quad (24)$$

白话说，给定平面上任意一点  $(x1, x2)$ ，它的分类可能性只有三个—— $C1$ 、 $C2$ 、 $C3$ 。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

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；

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](mailto:jiang.visualize.ml@gmail.com)

- 朴素贝叶斯分类器的“朴素”来自假设“特征条件独立”；
- 后验  $\propto$  似然  $\times$  先验；
- 比较联合概率 (似然  $\times$  先验) 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。

## 9.3 高斯你好

高斯的足迹几乎踏遍数学的每个角落，他所到之处都留下了自己的名字。哪怕在机器学习算法中，“高斯”这个金字招牌也反复出现。比如，本书提到几种算法：

- 高斯朴素贝叶斯 (Gaussian Naive Bayes)
- 高斯判别分析 (Gaussian discriminant analysis)
- 高斯过程 (Gaussian process)
- 高斯混合模型 (Gaussian mixture model)

并不是高斯发明了这些算法；而是，后来人在创造这些算法时，都利用了**高斯分布** (Gaussian distribution)。



卡尔·弗里德里希·高斯 (Carl Friedrich Gauss)

德国数学家、物理学家、天文学家 | 1777 ~ 1855

常被称作“数学王子”，在数学的每个领域开疆拓土。丛书关键词：● 等差数列 ● 高斯分布  
● 最小二乘法 ● 高斯朴素贝叶斯 ● 高斯判别分析 ● 高斯过程 ● 高斯混合模型 ● 高斯核函数

### 原理

上一章介绍了朴素贝叶斯分类，这种分类算法思路核心在于如下三点：

- 贝叶斯定理；
- 假设特征之间条件独立 (朴素之处)；
- 优化目标为最大化后验概率，或最大化联合概率 (似然  $\times$  先验)。

上一章在估算单一特征条件边际分布时，采用高斯核密度估计 KDE。而本章介绍的**高斯朴素贝叶斯** (Gaussian Naive Bayes) 最大不同在于，采用高斯分布估计单一特征条件边际分布。

### 最大化后验概率

朴素贝叶斯分类的优化目标可以是——最大化后验概率。对于二分类问题，直接比较  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  两个后验概率大小，就可以预测分类。

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](mailto:jiang.visualize.ml@gmail.com)



图 19 所示为基于高斯分布得到的  $f_{Y|X}(C_1 | \mathbf{x})$  和  $f_{Y|X}(C_2 | \mathbf{x})$  两个后验概率曲面。比较上一章基于 KDE 的后验概率曲面，可以发现高斯后验概率曲面非常平滑。图 19 中深蓝色曲线就是决策边界。这条决策边界实际上是二次曲线。

➔ 这一点，我们将会在下一章讲解**高斯判别分析** (Gaussian Discriminant Analysis, GDA) 时深入介绍。

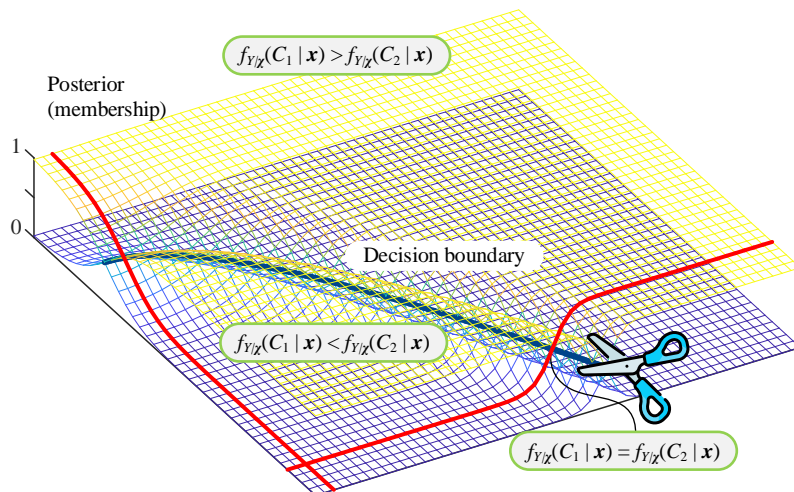


图 19. 二分类，比较后验概率大小，基于高斯分布

### 最大化联合概率

上一章提到，朴素贝叶斯分类的优化目标同样可以是——最大化联合概率。原因是，联合概率正比于后验概率。图 20 所示为，二分类问题中，比较联合概率  $f_{Y,X}(C_1, \mathbf{x})$  和  $f_{Y,X}(C_2, \mathbf{x})$  两个曲面高度，可以获得相同的决策边界。

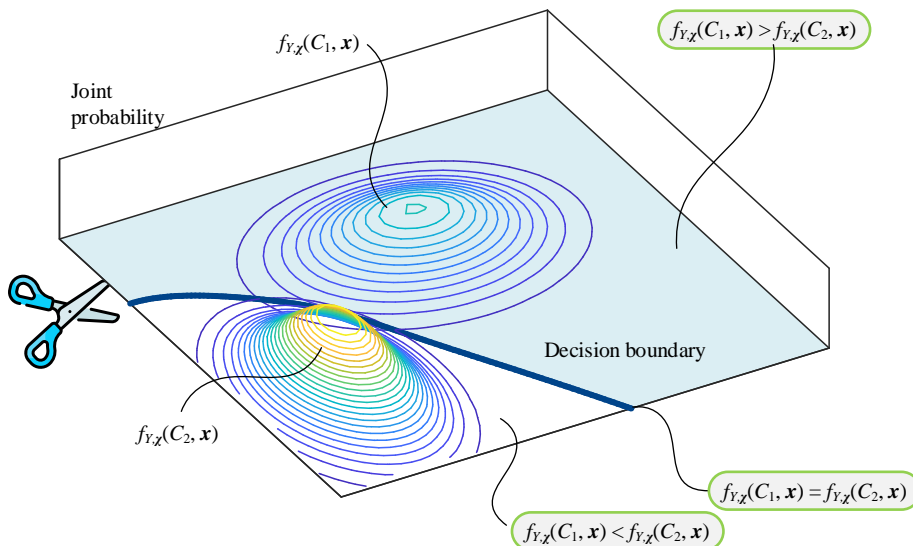
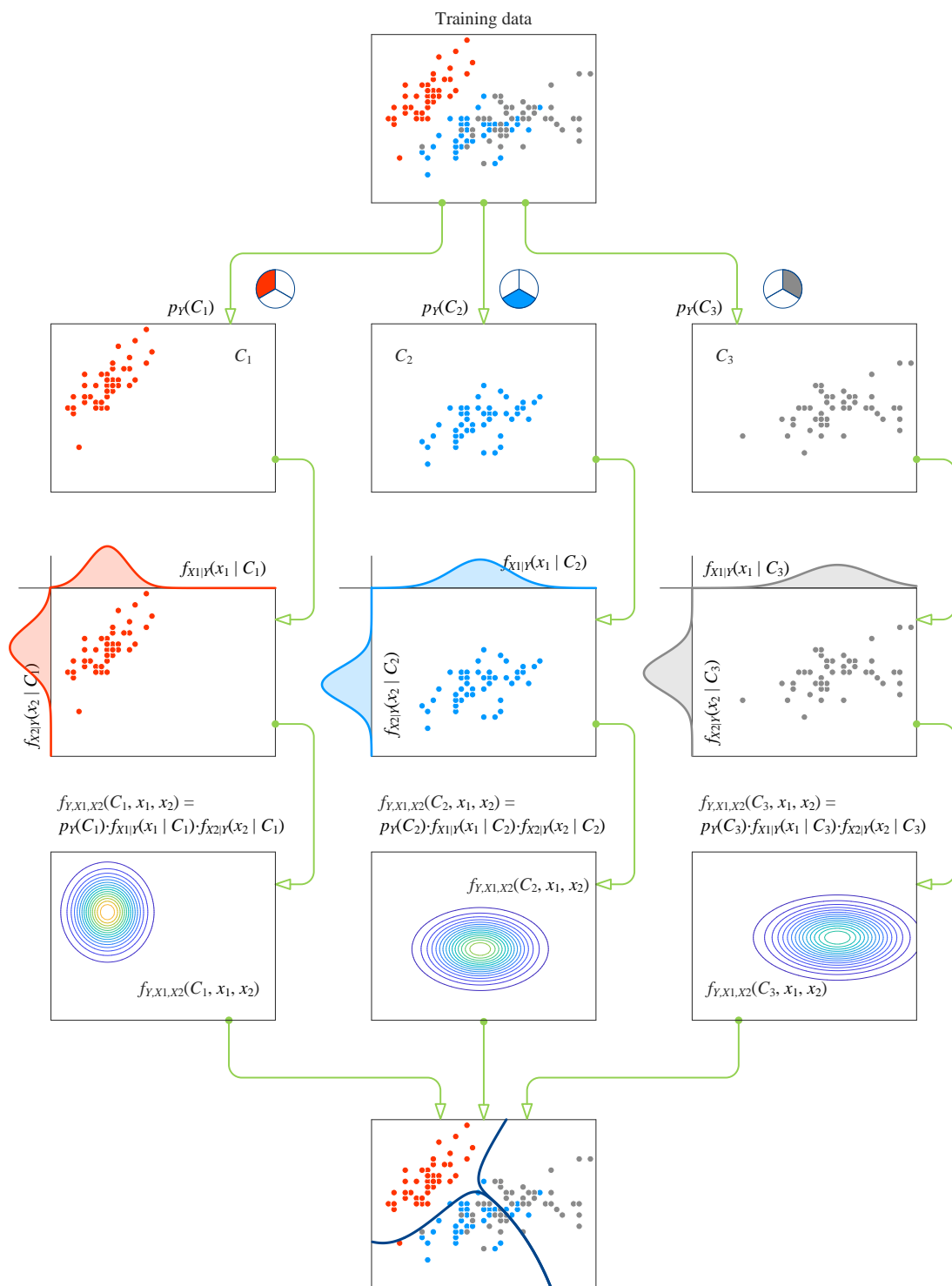


图 20. 二分类，比较联合概率大小，基于高斯分布

### 流程

图 5 所示为高斯朴素贝叶斯分类流程图。这一流程和上一章介绍的朴素贝叶斯分类流程完全一致。前文已经指出，高斯朴素贝叶斯分类器的特点是，估算特征条件边际分布时，高斯朴素贝叶斯分类采用高斯分布。

为方便大家学习，本章采用和上一章几乎一样结构。建议大家阅读本章时，平行对比上一章，对照边际分布曲线变化趋势，比较各种概率曲面特征，特别是对比决策边界形态。此外，本章帮助读者回顾高斯分布，让大家了解到在机器学习算法中如何引入高斯分布，以及明白高斯分布对决策边界形态有怎样的影响。



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](mailto:jiang.visualize.ml@gmail.com)

图 21. 高斯朴素贝叶斯分类过程

## 似然概率

朴素贝叶斯分类算法在估算似然概率时，假设特征之间条件独立：

$$\underbrace{f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{C}_k)}_{\text{Likelihood}} = \prod_{j=1}^D f_{x_j|\mathbf{y}}(x_j|\mathbf{C}_k) \quad (25)$$

比如，下式计算  $C_1$  似然概率密度：

$$\underbrace{f_{x_1, x_2|\mathbf{y}}(x_1, x_2|\mathbf{C}_1)}_{\text{Likelihood}} = \underbrace{f_{x_1|\mathbf{y}}(x_1|\mathbf{C}_1)}_{\text{Conditional independence}} f_{x_2|\mathbf{y}}(x_2|\mathbf{C}_1) \quad (26)$$

## 引入高斯分布

高斯朴素贝叶斯中条件边际分布采用的是高斯分布估计。比如，上式中的  $f_{x_1|\mathbf{y}}(x_1|\mathbf{C}_1)$  和  $f_{x_2|\mathbf{y}}(x_2|\mathbf{C}_1)$  可以写成：

$$\begin{cases} f_{x_1|\mathbf{y}}(x_1|\mathbf{C}_1) = \frac{1}{\sqrt{2\pi}\sigma_{1|\mathbf{C}_1}} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_{1|\mathbf{C}_1}}{\sigma_{1|\mathbf{C}_1}}\right)^2\right) \\ f_{x_2|\mathbf{y}}(x_2|\mathbf{C}_1) = \frac{1}{\sqrt{2\pi}\sigma_{2|\mathbf{C}_1}} \exp\left(-\frac{1}{2}\left(\frac{x_2 - \mu_{2|\mathbf{C}_1}}{\sigma_{2|\mathbf{C}_1}}\right)^2\right) \end{cases} \quad (27)$$

对于鸢尾花数据， $\mu_{1|\mathbf{C}_1}$  为标签为  $C_1$  数据在花萼长度  $x_1$  特征上的均值， $\sigma_{1|\mathbf{C}_1}$  为  $C_1$  数据在  $x_1$  特征上标准差； $\mu_{2|\mathbf{C}_1}$  为标签为  $C_1$  数据在花萼宽度  $x_2$  特征上的均值， $\sigma_{2|\mathbf{C}_1}$  为  $C_1$  数据在  $x_2$  特征上标准差。

图 22 中给出  $f_{x_1|\mathbf{y}}(x_1|\mathbf{C}_1)$  和  $f_{x_2|\mathbf{y}}(x_2|\mathbf{C}_1)$  两个概率密度函数曲线，以及  $\mu_{1|\mathbf{C}_1}$  和  $\mu_{2|\mathbf{C}_1}$  所在位置。

将 (27) 代入 (17)，可以得到  $f_{x_1, x_2|\mathbf{y}}(x_1, x_2|\mathbf{C}_1)$ ：

$$\begin{aligned} f_{x_1, x_2|\mathbf{y}}(x_1, x_2|\mathbf{C}_1) &= f_{x_1|\mathbf{y}}(x_1|\mathbf{C}_1) \cdot f_{x_2|\mathbf{y}}(x_2|\mathbf{C}_1) \\ &= \frac{\exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_{1|\mathbf{C}_1}}{\sigma_{1|\mathbf{C}_1}}\right)^2\right)}{\sqrt{2\pi}\sigma_{1|\mathbf{C}_1}} \times \frac{\exp\left(-\frac{1}{2}\left(\frac{x_2 - \mu_{2|\mathbf{C}_1}}{\sigma_{2|\mathbf{C}_1}}\right)^2\right)}{\sqrt{2\pi}\sigma_{2|\mathbf{C}_1}} \\ &= \frac{\exp\left(-\frac{1}{2}\left(\frac{(x_1 - \mu_{1|\mathbf{C}_1})^2}{\sigma_{1|\mathbf{C}_1}^2} + \frac{(x_2 - \mu_{2|\mathbf{C}_1})^2}{\sigma_{2|\mathbf{C}_1}^2}\right)\right)}{(\sqrt{2\pi})^2 \sigma_{1|\mathbf{C}_1} \sigma_{2|\mathbf{C}_1}} \end{aligned} \quad (28)$$

图 22 中等高线便是  $f_{x_1, x_2|\mathbf{y}}(x_1, x_2|\mathbf{C}_1)$  曲面等高线。

大家可能已经发现，图 22 中等高线为正椭圆！对于本节情况，正椭圆说明特征条件独立。



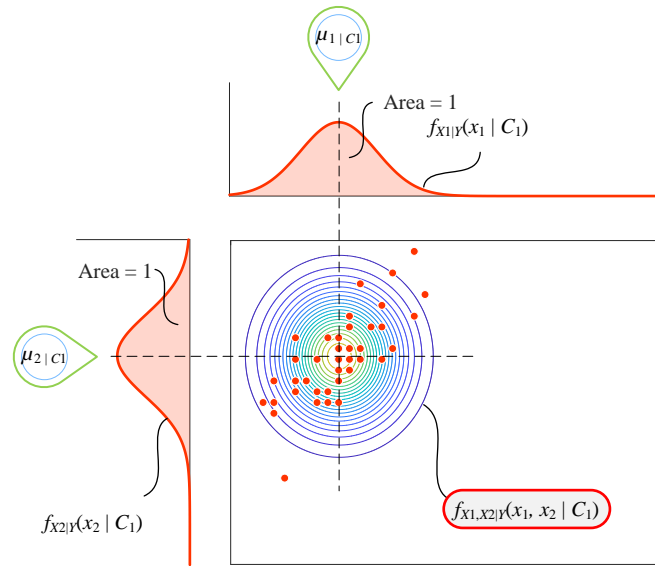


图 22. 分类  $C_1$  样本数据，假设鸢尾花花萼长度  $x_1$  和花萼宽度  $x_2$  条件独立，得到似然概率  $f_{x_1,x_2|Y}(x_1, x_2 | C_1)$ ，基于高斯分布

图 23 和图 24 所示为似然概率  $f_{x_1,x_2|Y}(x_1, x_2 | C_2)$  和  $f_{x_1,x_2|Y}(x_1, x_2 | C_3)$  结果。

**⚠** 再次提醒大家注意，“特征条件独立”不同于“特征独立”。

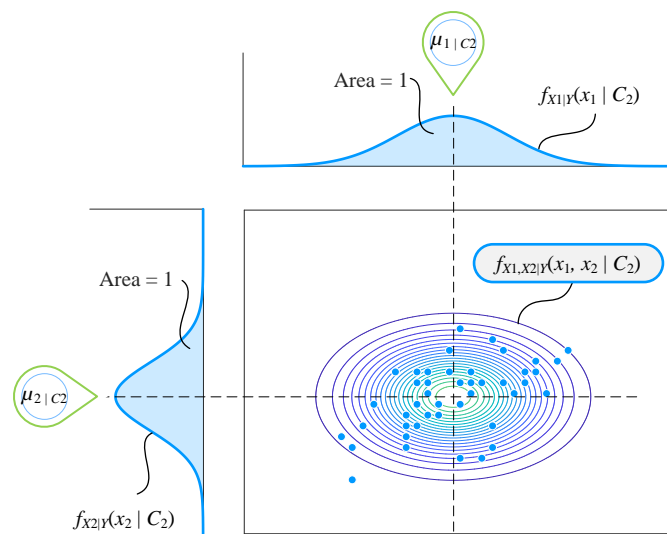


图 23. 分类  $C_2$  样本数据，假设鸢尾花花萼长度  $x_1$  和花萼宽度  $x_2$  条件独立，得到似然概率  $f_{x_1,x_2|Y}(x_1, x_2 | C_2)$ ，基于高斯分布

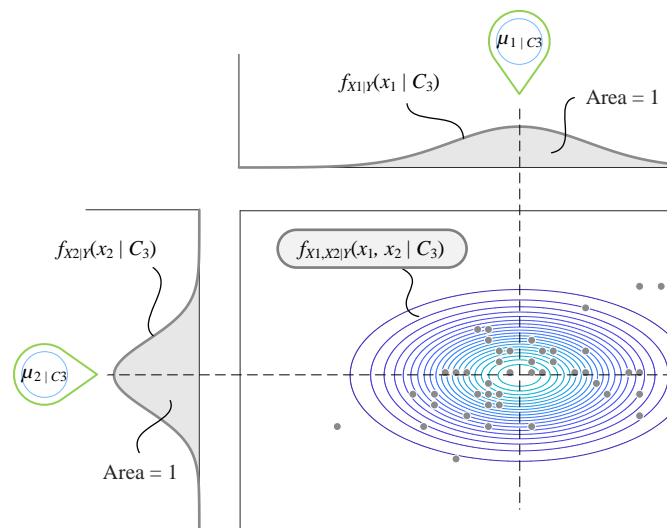


图 24. 分类  $C_3$  样本数据，假设鸢尾花花萼长度  $x_1$  和花萼宽度  $x_2$  条件独立，得到似然概率  $f_{x1,x2|y}(x_1, x_2 | C_3)$ ，基于高斯分布

## 联合概率

这一节利用下式估算联合概率：

$$\underbrace{f_{\mathbf{x},Y}(\mathbf{x}, C_k)}_{\text{Joint}} = \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{f_{\mathbf{x}|Y}(\mathbf{x} | C_k)}_{\text{Likelihood}} = \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{\prod_{j=1}^D f_{x_j|Y}(x_j | C_k)}_{\text{Conditional independence}} \quad (29)$$

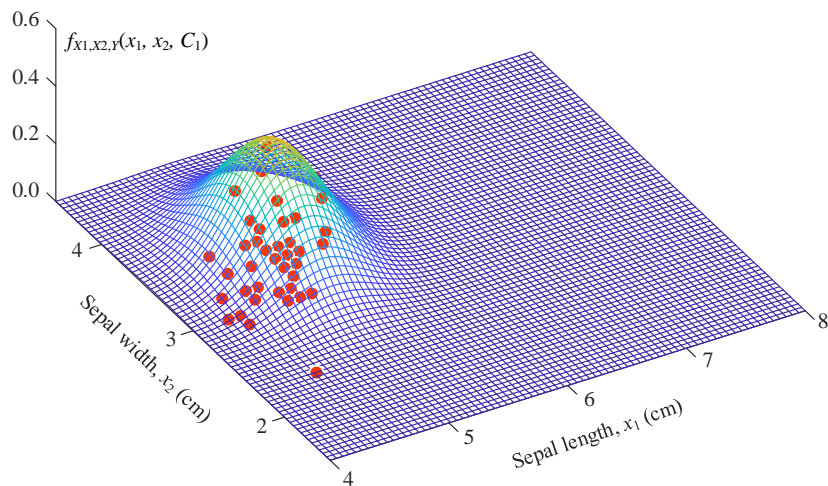
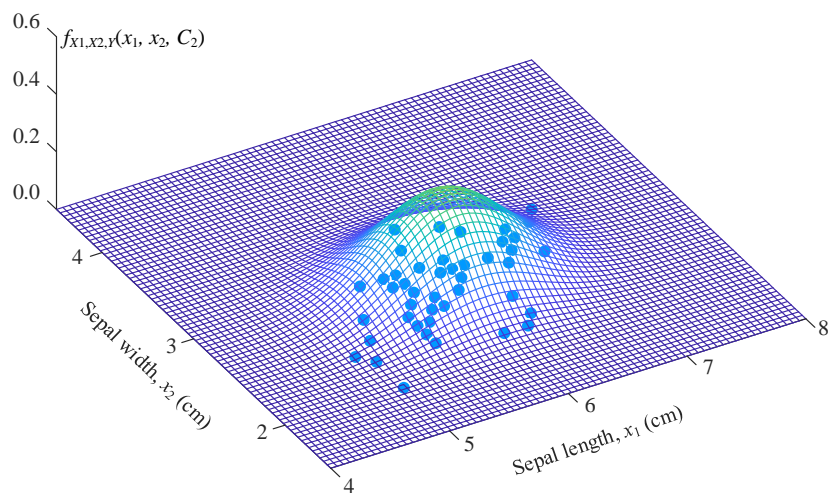
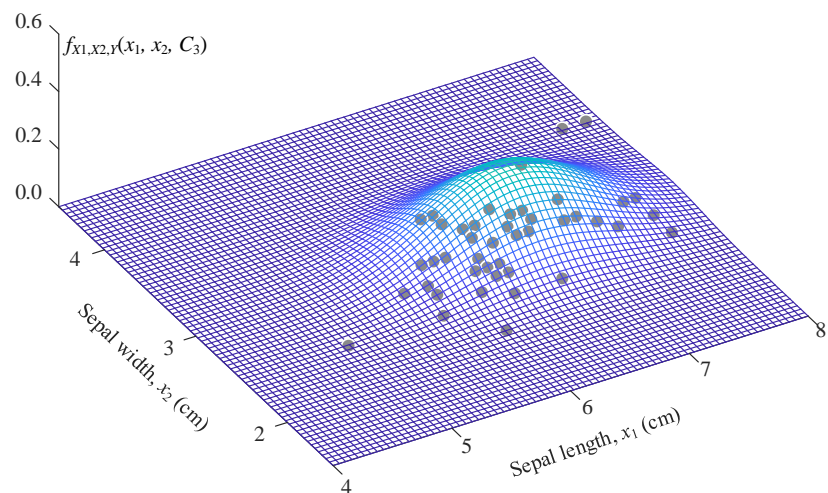
## 三分类问题

对于鸢尾花三分类 ( $K = 3$ ) 问题，联合概率  $f_{x1,x2,y}(x_1, x_2, C_k)$  ( $k = 1, 2, 3$ ) 可以通过下式得到：

$$\underbrace{f_{x1,x2,y}(x_1, x_2, C_k)}_{\text{Joint}} = \underbrace{f_{x1,x2|y}(x_1, x_2 | C_k)}_{\text{Likelihood}} \underbrace{p_Y(C_k)}_{\text{Prior}} = \underbrace{f_{x1|y}(x_1 | C_k)}_{\text{Conditional independence}} \cdot \underbrace{f_{x2|y}(x_2 | C_k)}_{\text{Conditional independence}} \underbrace{p_Y(C_k)}_{\text{Prior}} \quad (30)$$

再次注意，先验概率  $p_Y(C_k)$  相当于一个缩放系数。

图 25 ~ 图 27 所示为  $f_{x1,x2,y}(x_1, x_2, C_1)$ 、 $f_{x1,x2,y}(x_1, x_2, C_2)$  和  $f_{x1,x2,y}(x_1, x_2, C_3)$  三个联合概率密度函数曲面。

图 25.  $f_{X1,X2,Y}(x_1, x_2, C_1)$  概率密度曲面，基于高斯分布图 26.  $f_{X1,X2,Y}(x_1, x_2, C_2)$  概率密度曲面，基于高斯分布图 27.  $f_{X1,X2,Y}(x_1, x_2, C_3)$  概率密度曲面，基于高斯分布

### 分类依据：最大化联合概率

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](mailto:jiang.visualize.ml@gmail.com)

根据上一章介绍的高斯朴素贝叶斯优化目标之一——最大化联合概率；考虑到特征条件独立这一假设，高斯朴素贝叶斯目标函数为：

$$\hat{y} = \arg \max_{C_k} p_Y(C_k) \prod_{j=1}^D f_{X_j|Y}(x_j|C_k) \quad (31)$$

因此，比较图 25 ~ 图 27 三个曲面高度，可以进行鸢尾花分类预测。

sklearn 工具包高斯朴素贝叶斯分类算法的函数为 `sklearn.naive_bayes.GaussianNB`。同样，这个函数常用的 methods 为 `fit(X, y)` 和 `predict(q)`。`fit(X, y)` 用来加载样本数据，`predict(q)` 用来预测查询点  $q$  的分类。

通过 sklearn 高斯朴素贝叶斯分类算法得到的分类预测和决策边界。比较上一章的决策边界，可以发现高斯朴素贝叶斯分类算法得到的决策边界，形态上更简洁。图 28 中决策边界实际上是二次曲线。也就是说，协方差矩阵为对角阵，即特征条件独立时，高斯判别分析算法得到的决策边界等同于高斯朴素贝叶斯。这一点，下一章讲详细讲解。

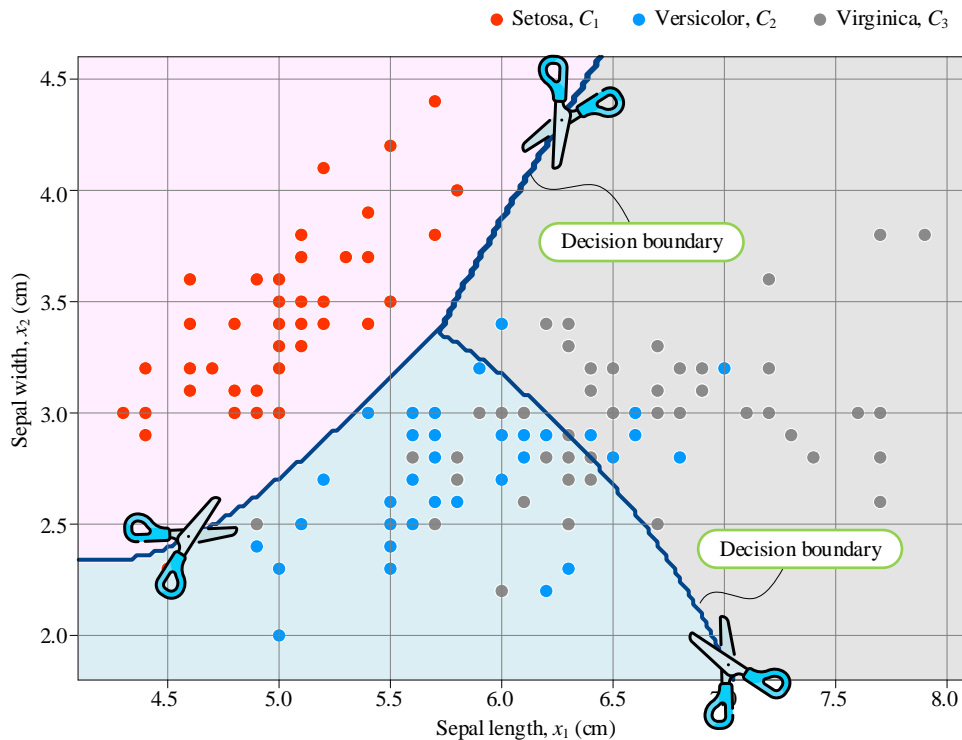


图 28. 鸢尾花分类预测，朴素贝叶斯决策边界，基于高斯分布



代码 Bk7\_Ch05\_01.py 利用高斯朴素贝叶斯分类鸢尾花数据集并绘制图 28。

## 证据因子：一种概率估算方法

根据全概率定理以及假设特征条件独立，证据因子  $f(x)$  可以这样计算：

$$\underbrace{f_Z(\mathbf{x})}_{\text{Evidence}} = \sum_{k=1}^K \left\{ \underbrace{p_Y(C_k)}_{\text{Prior}} \underbrace{\prod_{j=1}^D f_{X_j|Y}(x_j|C_k)}_{\text{Conditional independence}} \right\} \quad (32)$$

上一章提到，上式本身是一种概率密度估算方法，具体如图 29 所示。

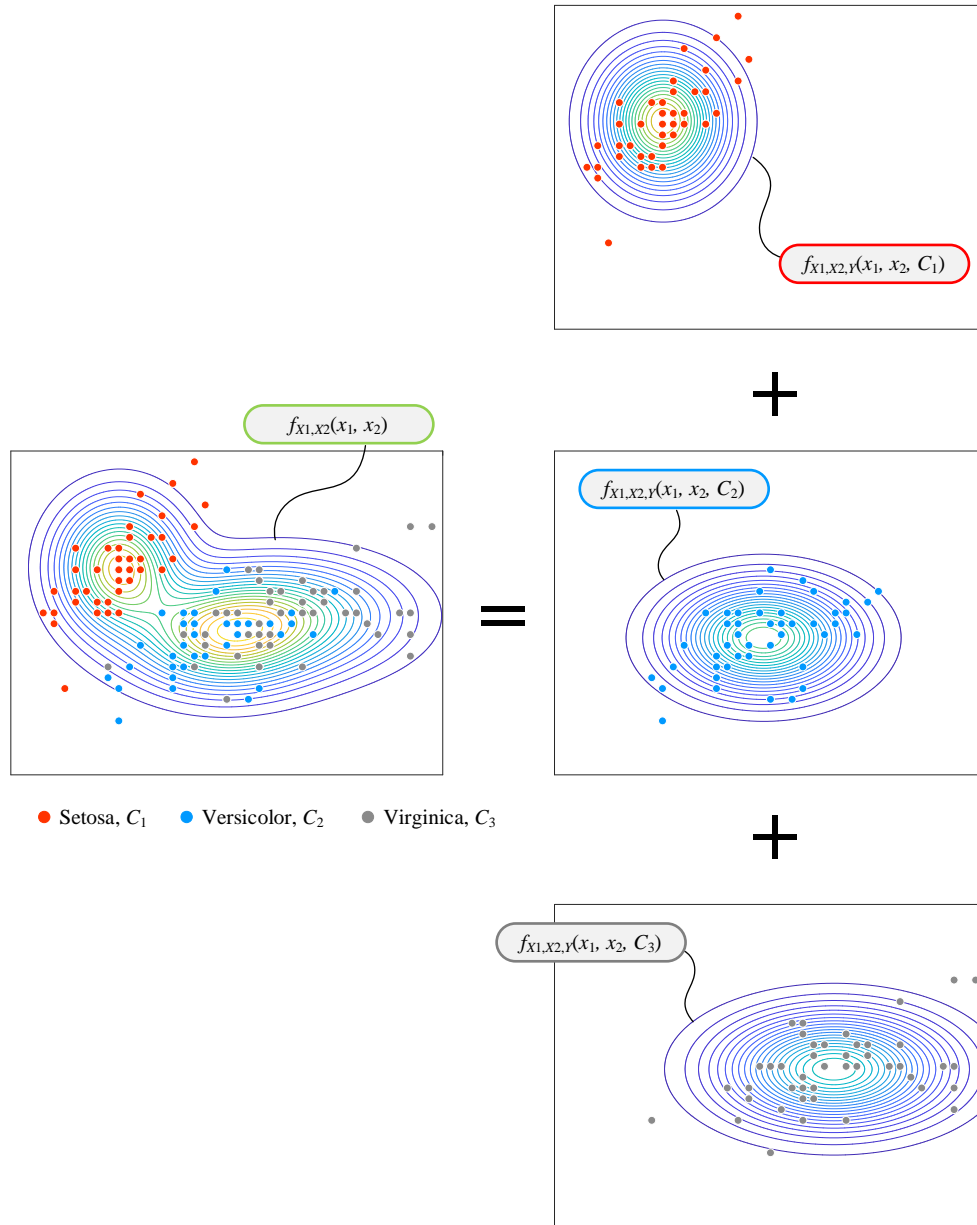


图 29. 估算证据因子概率密度，基于高斯分布

图 30 所示为估算得到的二元概率密度曲面  $f_{X1,X2}(x1, x2)$ 。注意，这个概率密度曲面主要基于以下两点：(1) 假设特征条件独立；(2) 条件边际概率通过一元高斯分布估计。

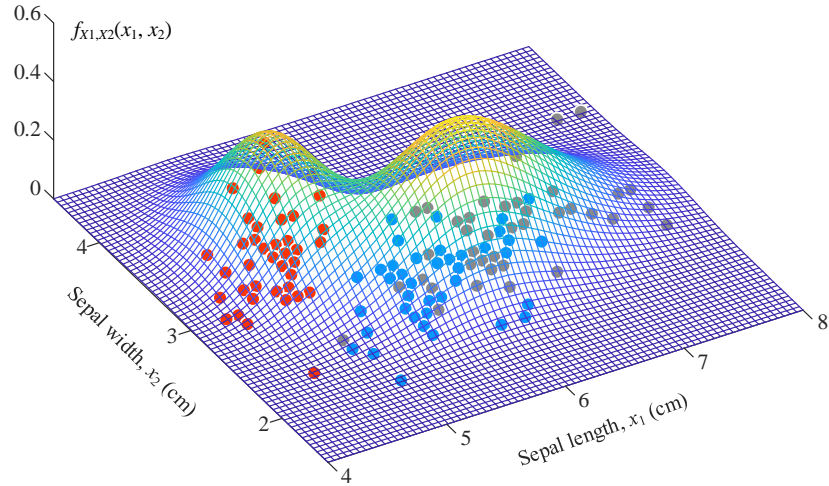


图 30. 估算得到的二元概率密度曲面，特征条件独立，基于高斯分布

### 后验概率：成员值

利用先验概率、似然概率和证据因子，根据贝叶斯定理计算得到后验概率，即成员值：

$$\underbrace{f_{Y|X}(C_k | \mathbf{x})}_{\text{Posterior}} = \frac{\overbrace{f_{X|Y}(\mathbf{x} | C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}}}{\underbrace{f_X(\mathbf{x})}_{\text{Evidence}}} \quad (33)$$

其中，假设分母中的证据因子不为 0。

如果假设“特征条件独立”，上式可以写成：

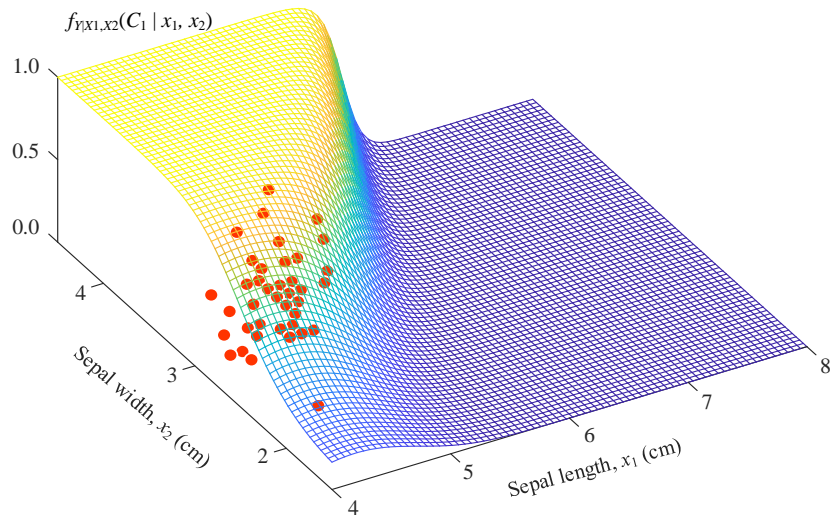
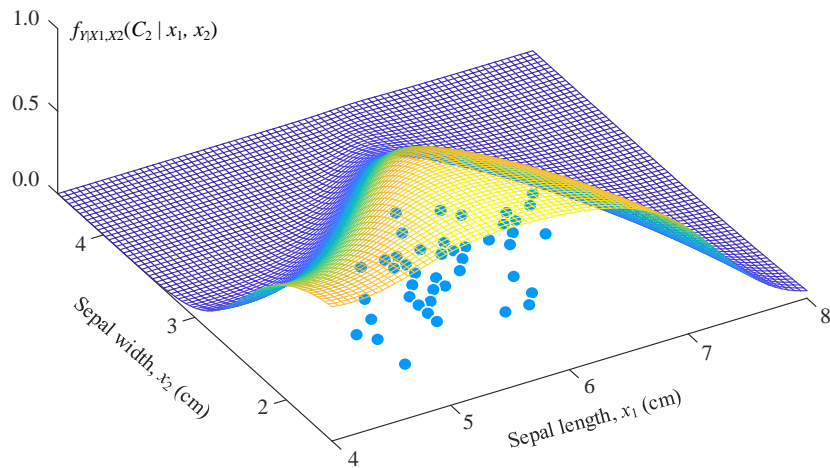
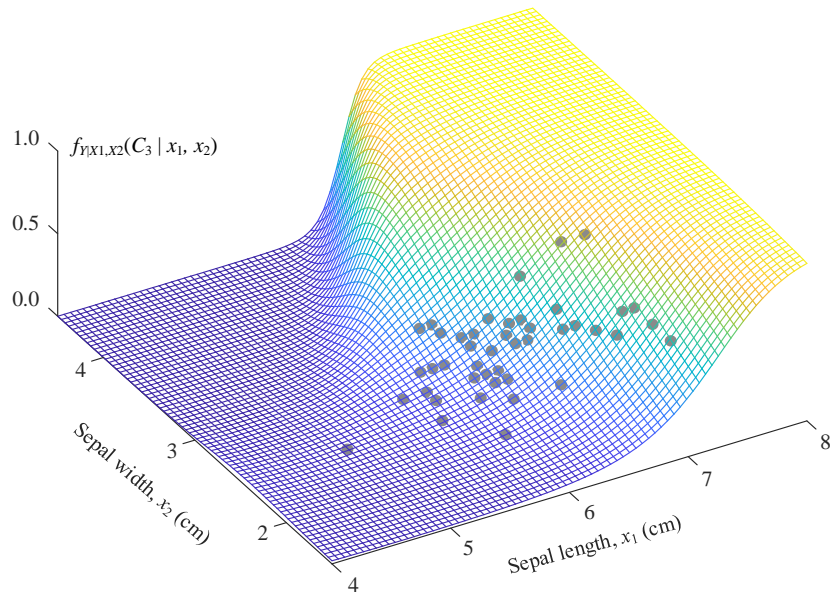
$$\underbrace{f_{Y|X}(C_k | \mathbf{x})}_{\text{Posterior}} = \frac{\prod_{j=1}^D \overbrace{f_{X_j|Y}(x_j | C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}}}{\underbrace{f_X(\mathbf{x})}_{\text{Evidence}}} \quad (34)$$

上一章介绍过，朴素贝叶斯分类优化目标也可以是——最大化后验概率：

$$\hat{y} = \arg \max_{C_k} f_{Y|X}(C_k | \mathbf{x}) \quad (35)$$

图 31 ~ 图 33 所示为  $f_{Y|X1,X2}(C_1 | x_1, x_2)$ 、 $f_{Y|X1,X2}(C_2 | x_1, x_2)$  和  $f_{Y|X1,X2}(C_3 | x_1, x_2)$  三个后验概率曲面。



图 31.  $f_{Y|X1,X2}(C_1 | x_1, x_2)$  后验概率曲面，基于高斯分布图 32.  $f_{Y|X1,X2}(C_2 | x_1, x_2)$  后验概率曲面，基于高斯分布图 33.  $f_{Y|X1,X2}(C_3 | x_1, x_2)$  后验概率曲面，基于高斯分布

对于鸢尾花三分类问题，如图 34 所示，比较  $f_{Y|X1,X2}(C_1 | x_1, x_2)$ 、 $f_{Y|X1,X2}(C_2 | x_1, x_2)$  和  $f_{Y|X1,X2}(C_3 | x_1, x_2)$  三个后验概率密度曲面高度，可以预测分类，并获得决策边界。



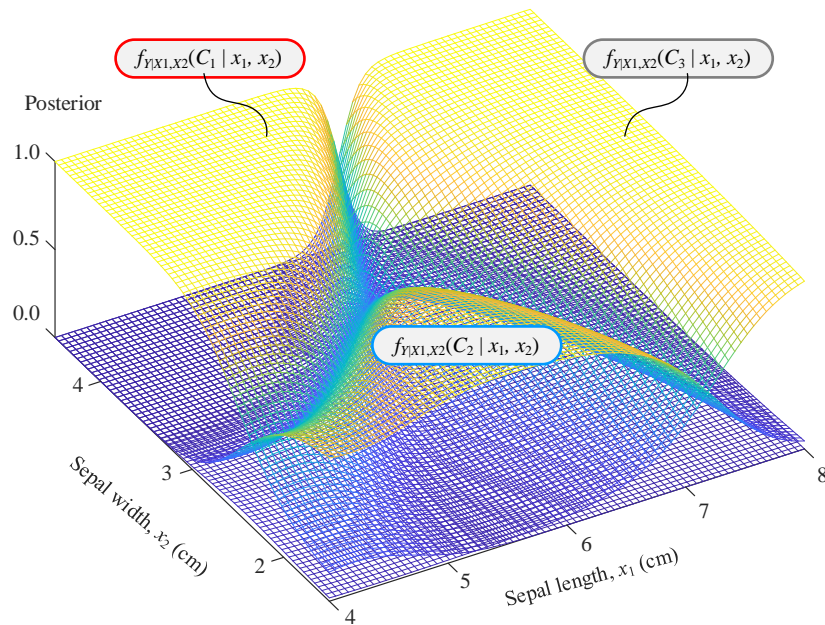


图 34. 比较三个后验概率曲面，基于高斯分布

贝叶斯定理是机器学习和深度学习中重要的概率论工具，广泛应用于分类、聚类、推荐系统等领域。本章和上一章介绍的朴素贝叶斯分类是贝叶斯定理的众多应用之一。我们在《统计至简》还介绍过贝叶斯统计推断，在《数据有道》聊过贝叶斯回归。

贝叶斯派思想强调我们对未知事物的认识应该是不断修正和更新的。它通过贝叶斯定理将已有的先验知识和新的实验数据结合起来，不断修正我们对未知事件的概率估计，实现对真实概率的逼近。贝叶斯派思想应用于机器学习和人工智能领域，可以用于推断和预测，解决实际问题，例如自然语言处理、图像识别、推荐系统等。贝叶斯派思想的优点是可以有效处理不确定性和噪声，具有广泛的应用前景。