18 贝叶斯分类

Bayesian Classification

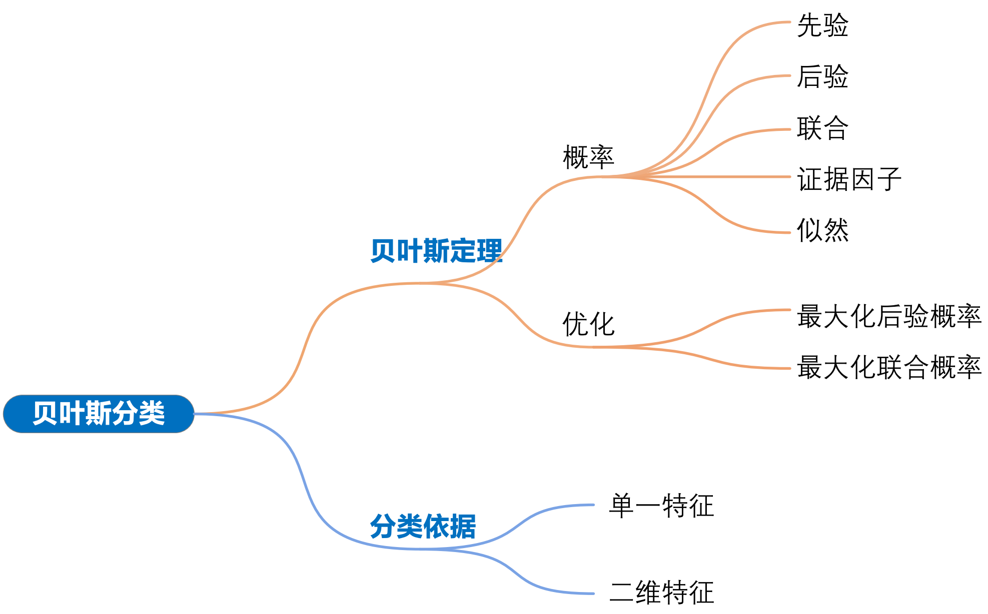
最大化后验概率，利用花萼长度分类鸢尾花

我们认为用最简单的假设来解释现象是一个很好的原则。

We consider it a good principle to explain the phenomena by the simplest hypothesis possible.

—— 托勒密 (Ptolemy) | 数学家、天文学家、地心说提出者 | 100 ~ 170

- ◀ matplotlib.pyplot.fill_between() 区域填充颜色
- ◀ seaborn.kdeplot() 绘制 KDE 概率密度估计曲线
- ◀ 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


贝叶斯分类
贝叶斯定理
概率
先验
后验
联合
证据因子
似然
优化
最大化后验概率
最大化联合概率
分类依据
单一特征
二维特征

18.1 贝叶斯定理：分类鸢尾花

本章和下一章和读者探讨采用贝叶斯定理对鸢尾花数据分类。本章采用鸢尾花数据中的花萼长度作为研究对象，利用 KDE 生成概率密度函数，预测鸢尾花分类。

以下是使用贝叶斯定理进行分类的一般步骤：

- 收集数据，并提取特征。
- 对于每个类别，计算其在所有样本中出现的概率，称之为先验概率。
- 对于每个特征，计算它在每个类别下的概率，称之为条件概率。
- 根据贝叶斯定理，计算给定特征下，每个类别出现的概率，称之为后验概率。
- 根据后验概率的大小判定分类。

具体实现过程中，可以使用不同的算法来计算条件概率和后验概率，如朴素贝叶斯算法、高斯朴素贝叶斯算法等。同时，为了避免过拟合和欠拟合问题，我们还需要使用交叉验证、平滑等技术来提高分类器的性能。

贝叶斯定理

大家知道鸢尾花数据分为三类——setosa、versicolour、virginica。我们分别用 C_1、C_2、C_3 作为标签代表这三类鸢尾花。

对于鸢尾花分类问题，贝叶斯定理可以按如下方式表达：

$$\underbrace{f_{Y|X}\left(C_k|x\right)}_{\text{Posterior}} = \frac{\overbrace{f_{X,Y}\left(x,C_k\right)}^{\text{Joint}}}{f_X\left(x\right)} = \frac{\overbrace{f_{X|Y}\left(x|C_k\right)}^{\text{Likelihood}}\overbrace{p_Y\left(C_k\right)}^{\text{Prior}}}{\underbrace{f_X\left(x\right)}_{\text{Evidence}}}, \quad k=1,2,3 \tag{1}$$

其中，X 代表鸢尾花花萼长度的连续随机变量，Y 代表分类的离散随机变量，Y 的取值为 C_1、C_2、C_3。

下面我们给 (1) 中几个概率值取名字：

$f_{Y|X}(C_k\,|\,x)$ 为**后验概率** (posterior)，又叫**成员值** (membership score)。在给定任意花萼长度 x 的条件下，比较三个后验概率 $f_{Y|X}(C_1|x)$、$f_{Y|X}(C_2|x)$、$f_{Y|X}(C_3|x)$ 大小，可以作为判定鸢尾花分类的依据。

$f_{X,Y}(x,C_k)$ 为**联合概率** (joint)，也可以记做 $f_{X\cap Y}(x\cap C_k)$。

$f_X(x)$ 为**证据因子** (evidence)，也叫证据。证据因子和分类无关，仅代表鸢尾花花萼长度 X 的概率分布情况。(1) 中，证据因子 $f_X(x)$ 对联合概率 $f_{X,Y}(x,C_k)$ 进行**归一化** (normalization) 处理。本章假设 $f_X(x) > 0$。

$p_Y(C_k)$ 为**先验概率** (prior)，表达样本集合中 C_k (k = 1, 2, 3) 类样本占比。注意，$p_Y(C_k)$ 为概率质量函数；这是因为随机变量 Y 为离散随机变量，取值为 $Y = C_1, C_2, C_3$。

$f_{X|Y}(x|C_k)$ 为**似然概率** (likelihood)。白话解释，给定类别 C_k 中 x 出现的可能性，比如给定鸢尾花为 setosa，花萼长度为 10 cm 的可能性可以写成 $f_{X|Y}(10 \mid \text{Setosa})$。

图 1 可视化三分类问题中的贝叶斯定理。下面，我们逐一讲解上述不同的概率，以及它们如何帮助我们完成鸢尾花分类。

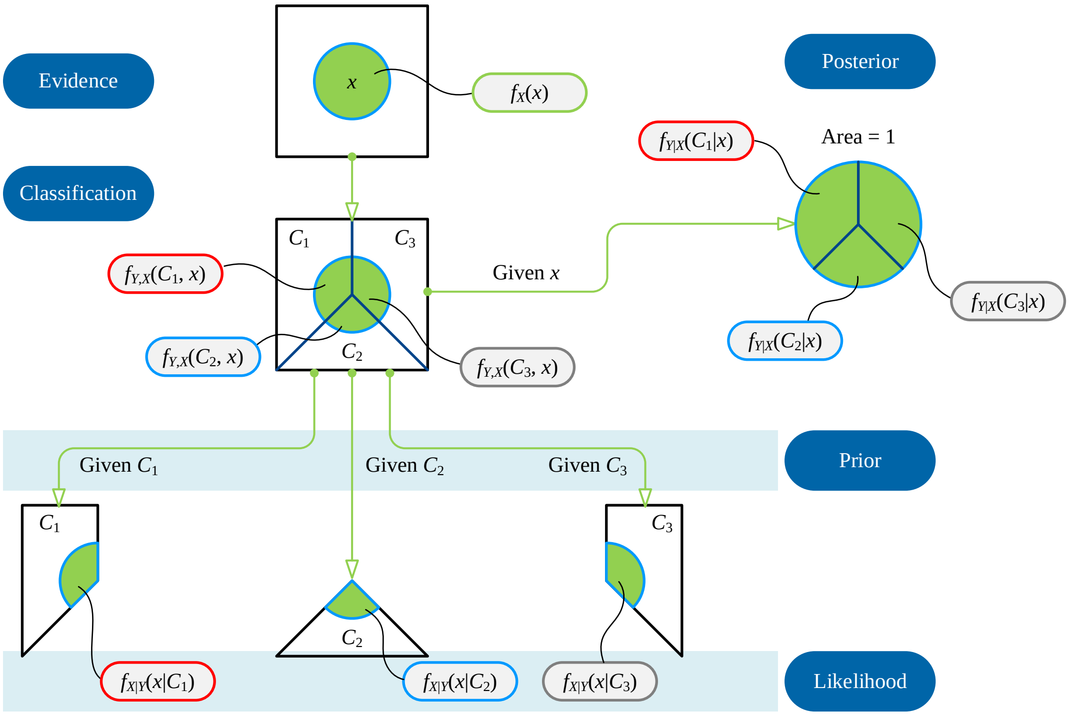


图 1. 利用贝叶斯定理，以花萼长度作为特征对鸢尾花进行分类

18.2 似然概率：给定分类条件下的概率密度

似然概率 $f_{X|Y}(x|C_k)$ 本身是条件概率，它描述的是给定类别 $Y = C_k$ 中 $X = x$ 出现的可能性。注意，$f_{X|Y}(x|C_k)$ 本身为概率密度函数 PDF。

图 2 (a)、(b)、(c) 分别展示 $f_{X|Y}(x|C_1)$、$f_{X|Y}(x|C_2)$、$f_{X|Y}(x|C_3)$ 三个似然概率 PDF 曲线。这三条概率密度曲线采用高斯 KDE 估计得到。

在鸢尾花数据集所有 150 个样本数据中如果，我们只分析标签为 C_1 (Setosa) 的 50 个样本的话，$f_{X|Y}(x|C_1)$ 就是这 50 个样本数据得到花萼长度的概率密度函数 PDF。

$f_{X|Y}(x|C_2)$ 代表给定鸢尾花分类为 C_2 (Versicolour)，花萼长度的概率密度函数。同理，$f_{X|Y}(x|C_3)$ 代表给定鸢尾花分类为 C_3 (Virginica)，花萼长度的概率密度函数。图 2 (c) 比较 $f_{X|Y}(x|C_1)$、$f_{X|Y}(x|C_2)$、$f_{X|Y}(x|C_3)$ 三条曲线。

⚠ 注意，$f_{X|Y}(x|C_k)$ 和横轴包裹的面积为 1。

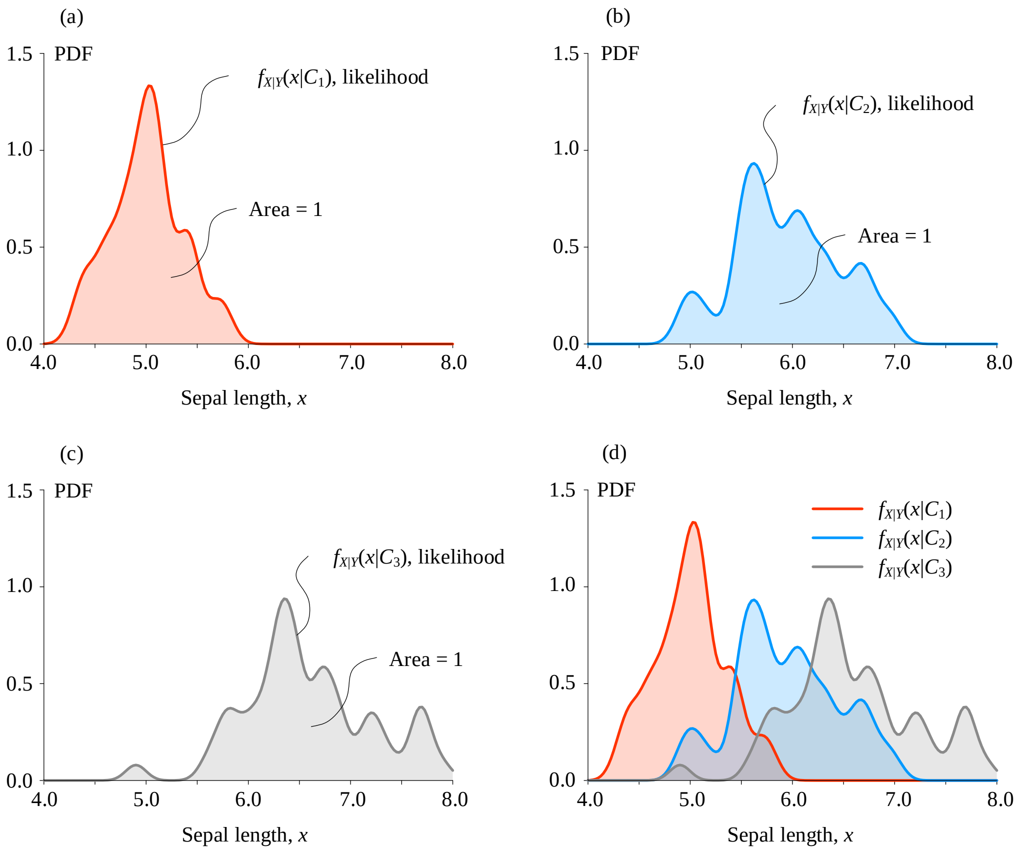


图 2. 三个似然概率 PDF 曲线 $f_{X|Y}(x|C_k)$

18.3 先验概率：鸢尾花分类占比

先验概率 $p_Y(C_k)$ 描述的是样本集合中 C_k 类样本占比。由于 Y 为离散随机变量，因此我们采用概率质量函数。$p_Y(C_k)$ 具体计算如下：

$$p_Y(C_k) = \frac{\text{count}(C_k)}{\text{count}(\Omega)}, \quad k = 1,2,3 \tag{2}$$

其中，count() 为计数运算符，count(C_k) 计算标签样本空间 Ω 中 C_k 类样本数据数量。

如图 3 所示，对于鸢尾花数据，每一类标签的样本数据都是 50，因此三类标签的先验概率都是 1/3：

$$p_Y\left(C_k\right)=\frac{50}{150}=\frac{1}{3},\quad k=1,2,3 \tag{3}$$

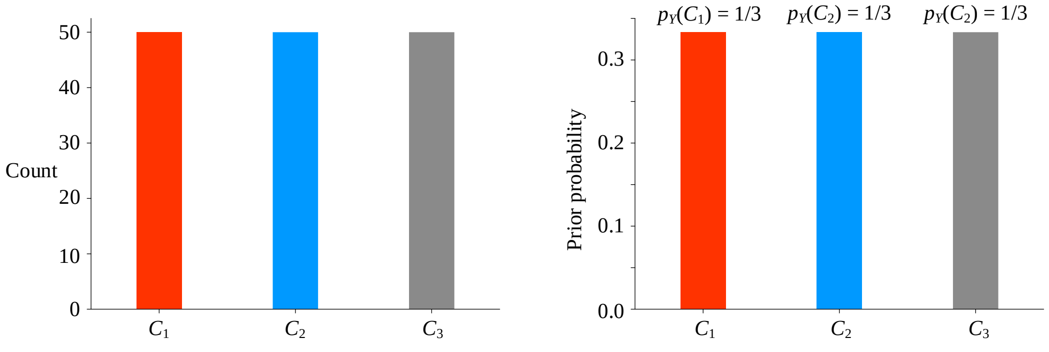


图 3. 150 个样本数据总三类的频数和先验概率

18.4 联合概率：可以作为分类标准

联合概率 $f_{X,Y}(x,C_k)$ 描述事件 $Y = C_k$ 和事件 $X = x$ 同时发生的可能性。

比如，花萼长度为 x = 5.6 cm 且鸢尾花分类为 $Y = C_1$ (Setosa) 的可能性可以用 $f_{X,Y}(5.6, C_1)$ 表达。

⚠ 注意，$f_{X,Y}(x,C_k)$ 为概率密度函数 PDF，并不是“概率”。

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

$$\overbrace{f_{X,Y}\left(x,C_k\right)}^{\text{Joint}}=\overbrace{f_{X|Y}\left(x|C_k\right)}^{\text{Likelihood}}\overbrace{p_Y\left(C_k\right)}^{\text{Prior}} \tag{4}$$

图 4 (a)、(b)、(c) 分别展示 $f_{X,Y}(x,C_1)$、$f_{X,Y}(x,C_2)$、$f_{X,Y}(x,C_3)$ 三个联合概率 PDF 曲线。这三幅图还展示从似然概率 $f_{X|Y}(x|C_k)$ 到联合概率 $f_{X,Y}(x,C_k)$ 的缩放过程。

似然概率 $f_{X|Y}(x|C_k)$ 和横轴包裹的面积为 1。而联合概率 $f_{X,Y}(x,C_k)$ 和横轴包裹的面积为 $p_Y(C_k)$。

图 4 (d) 比较 $f_{X,Y}(x,C_1)$、$f_{X,Y}(x,C_2)$、$f_{X,Y}(x,C_3)$ 三个联合概率 PDF 曲线，即“似然概率 × 先验概率”。实际上，这三条曲线的高低已经可以用来作为分类标准，这是本章后续要介绍的内容。

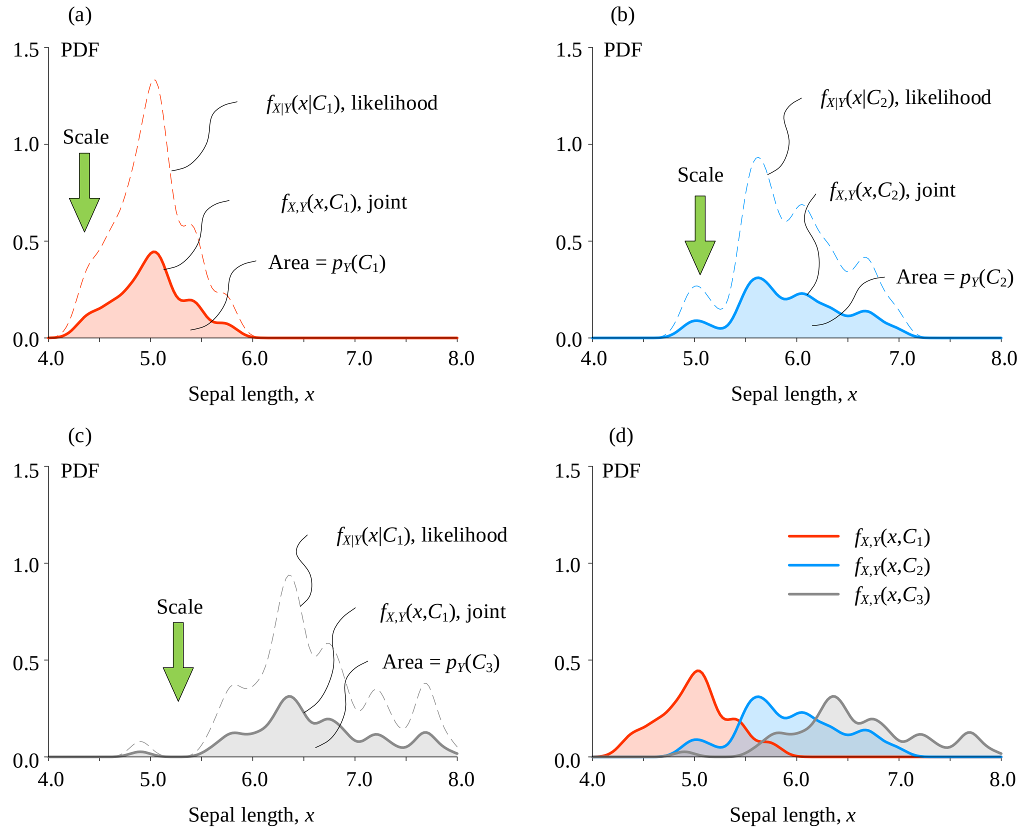


图 4. 先验概率和联合概率的关系

18.5 证据因子：和分类无关

证据因子 $f_X(x)$ 实际上就是 X 的边缘概率密度函数 PDF，证据因子和分类无关。对于本章鸢尾花花萼数据，$f_X(x)$ 就是根据样本数据利用 KDE 方法估计得到的概率密度函数。

显然，对于鸢尾花样本数据，C_1、C_2、C_3 为一组不相容分类，对样本空间 Ω 形成分割。根据全概率定理，下式成立：

$$\overbrace{f_X(x)}^{\text{Evidence}} = \sum_{k=1}^{3} \overbrace{f_{X,Y}(x,C_k)}^{\text{Joint}} = \sum_{k=1}^{3} \overbrace{f_{X|Y}(x|C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \tag{5}$$

也就是说，似然概率密度 $f_{X|Y}(x|C_k)$ 和先验概率 $p_Y(C_k)$，可以用来估算 $f_X(x)$。

对于鸢尾花三分类，(5) 可以展开来写：

$$f_X(x) = f_{X,Y}(x,C_1) + f_{X,Y}(x,C_2) + f_{X,Y}(x,C_3) \tag{6}$$

图 5 所示为利用联合概率 PDF 计算证据因子 PDF 的过程。

⚠ 注意，$f_X(x)$ 和横轴包裹的面积为 1。

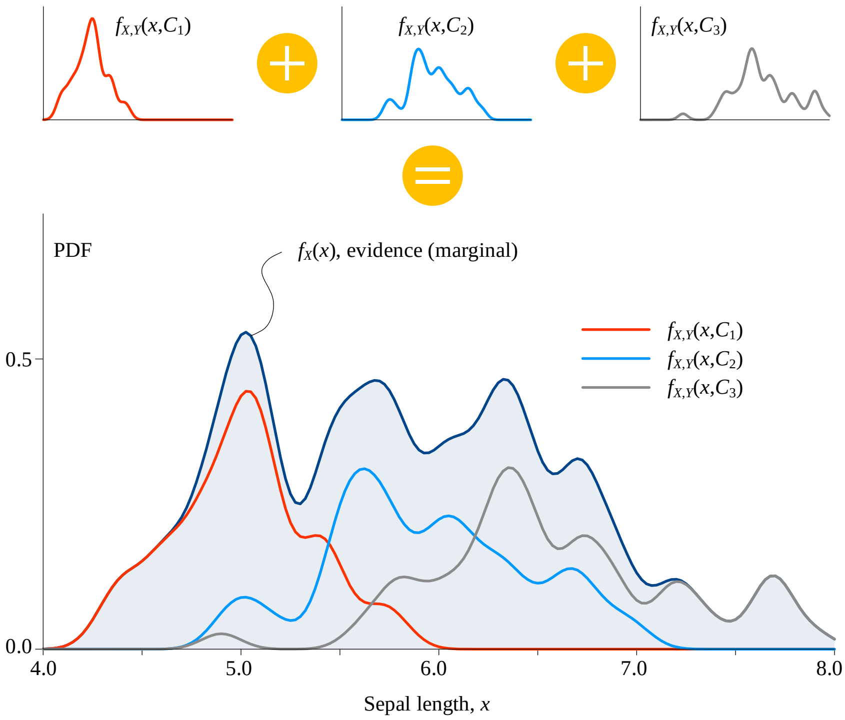


图 5. 叠加联合概率曲线，估算证据因子概率密度函数

18.6 后验概率：也是分类的依据

$f_{Y|X}(C_k\,|\,x)$ 指的是在事件 $X = x$ 发生条件下，事件 $Y = C_k$ 发生的概率。后验概率 $f_{Y|X}(C_k\,|\,x)$ 又叫成员值 (membership score)。白话来说，后验概率指的是在已知一些先验条件的情况下，通过贝叶斯定理计算得出的条件概率。换句话说，它是指在观测到某些数据或证据后，对于假设的某个事件发生的概率的更新。

比如，给定花萼的长度为 $x = 5.6$ cm，鸢尾花被分类为 $Y = C_1$ (Setosa) 的可能性，就可以用 $f_{Y|X}(C_1\,|\,5.6)$ 来描述。

⚠ 注意，后验概率实际上是概率，不是概率密度。因此，$f_{Y|X}(C_k\,|\,x)$ 的取值范围为 [0, 1]。

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

$$\overbrace{f_{Y|X}\left(C_k\,|\,x\right)}^{\text{Posterior}} = \frac{\overbrace{f_{X,Y}\left(x, C_k\right)}^{\text{Joint}}}{\underbrace{f_X\left(x\right)}_{\text{Evidence}}} \tag{7}$$

图 6 所示为后验概率 PDF 曲线 $f_{Y|X}(C_1|x)$ 的计算过程。图 7 则比较另外两组联合概率、证据因子、后验概率曲线。

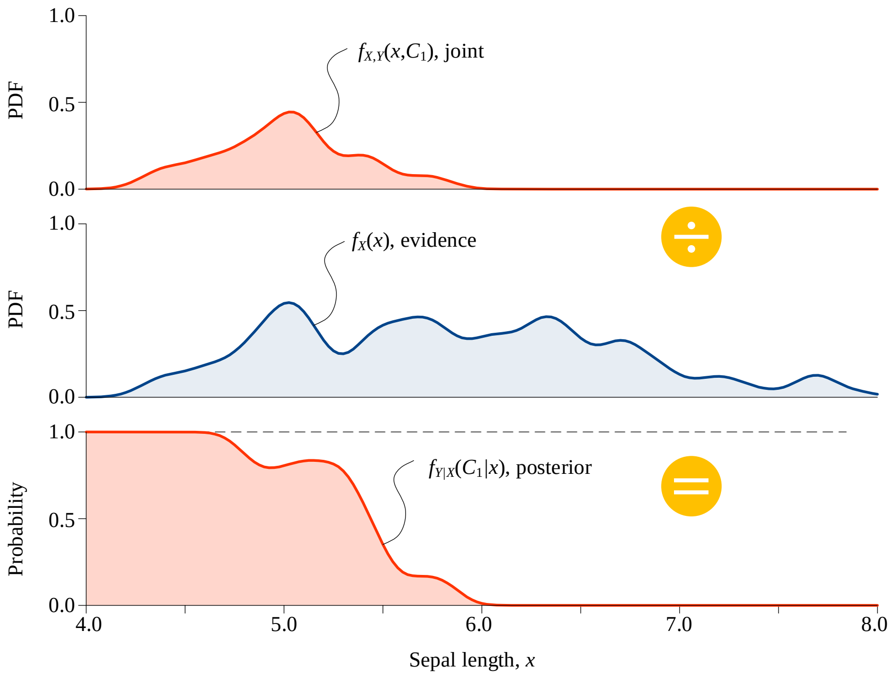


图 6. 计算后验概率 PDF 曲线 $f_{Y|X}(C_1|x)$

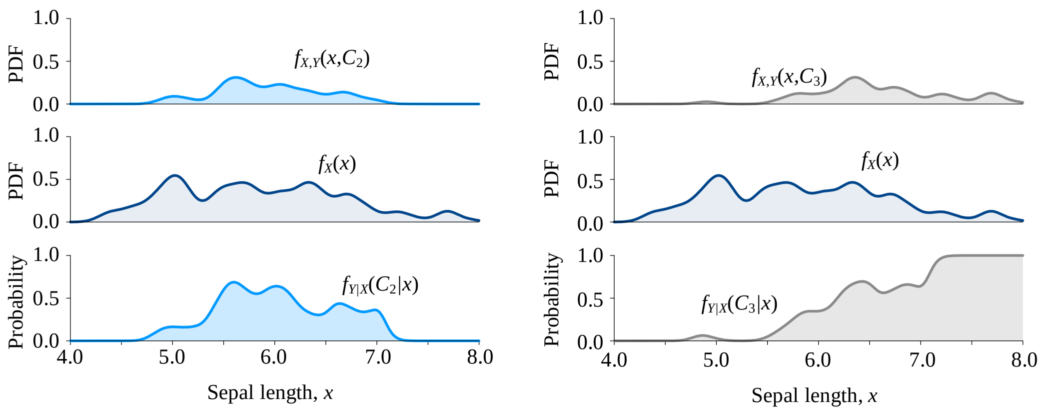


图 7. 比较联合概率、证据因子、后验概率曲线

成员值

后验概率之所以被称作"成员值"是因为：

$$\sum_{k=1}^{3}\underbrace{f_{Y|X}\left(C_k|x\right)}_{\text{Posterior}}=1 \tag{8}$$

这个式子不难推导。根据贝叶斯定理，下式成立：

$$\overbrace{f_X(x)}^{\text{Evidence}} = \sum_{k=1}^{3} \overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} = \sum_{k=1}^{3} \overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \overbrace{f_X(x)}^{\text{Evidence}} \tag{9}$$

即，

$$\overbrace{f_X(x)}^{\text{Evidence}} = \overbrace{f_X(x)}^{\text{Evidence}} \sum_{k=1}^{3} \overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \tag{10}$$

$f_X(x) > 0$ 时，(10) 左右消去 $f_X(x)$ 便获得 (8)。

分类依据

在给定任意花萼长度 x 的条件下，比较三个后验概率 $f_{Y|X}(C_1|x)$、$f_{Y|X}(C_2|x)$、$f_{Y|X}(C_3|x)$ 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。

举个例子，某朵鸢尾花花萼长度为 x = 5.6 cm 的前提下，它一定被分类为 C_1、C_2、C_3 任一标签。三种不同情况的可能性相加为 1，也就是说，这朵鸢尾花要么是 C_1，或者是 C_2，不然就是 C_3。

换个角度来看，比较图 8 三条不同颜色曲线的高度，我们就可以据此判断鸢尾花的分类。

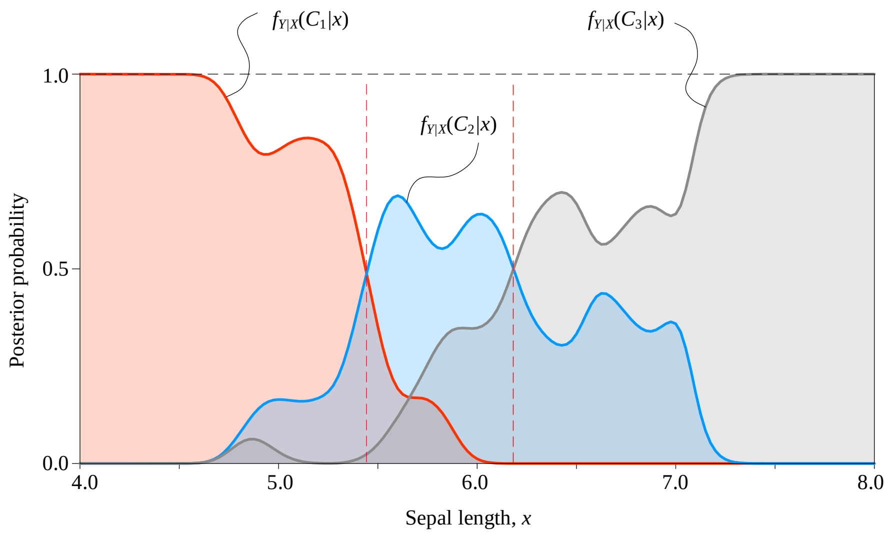


图 8. 比较三个后验概率 PDF 曲线 $f_{Y|X}(C_1|x)$、$f_{Y|X}(C_2|x)$、$f_{Y|X}(C_3|x)$

观察 (7)，可以发现后验概率 $f_{Y|X}(C_1|x)$ 正比于联合概率 $f_{X,Y}(x,C_k)$，证据因子 $f_X(x)$ 仅仅起到缩放作用：

$$\overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \propto \overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} \tag{11}$$

实际上，没有必要计算后验概率 $f_{Y|X}(C_1|x)$，比较联合概率 $f_{X,Y}(x,C_k)$ 就可以对鸢尾花进行分类。上式实际上是贝叶斯推断中最重要的比例关系——后验 ∝ 先验 × 似然。

比较四条曲线

本节最后，我们把**似然概率** (likelihood)、**联合概率** (joint)、**证据因子** (evidence)、**后验概率** (posterior) 这四条曲线放在一幅中加以比较，具体如图 9、图 10、图 11 所示。

请大家注意以下几点：

- 似然概率 (likelihood) 曲线为条件概率密度，和横轴围成图形的面积为 1；
- 似然概率 (likelihood) 经过先验概率 (prior) 缩放得到联合概率 (joint)；
- 比较联合概率密度 (即先验 × 似然) 大小，可以预测分类；
- 后验 ∝ 先验 × 似然；
- 联合概率曲线面积为对应先验概率；
- 联合概率叠加得到证据因子 (evidence)；
- 联合概率 (joint) 除以证据因子得到后验概率 (posterior)，证据因子起到归一化作用；
- 后验概率，也叫成员值 (membership score)，本质上是概率值，取值范围在 [0, 1] 之间；
- 比较后验概率/成员值大小，可以预测分类，方便可视化。

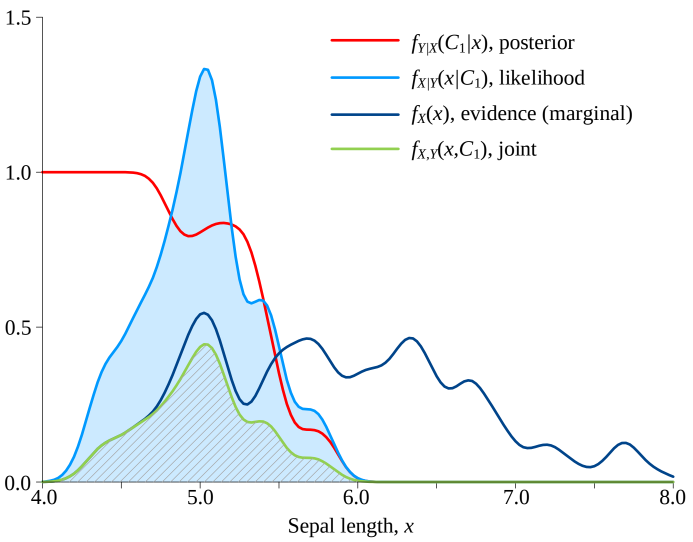


图 9. 比较后验概率 $f_{Y|X}(C_1 \mid x)$、似然概率 $f_{X|Y}(x|C_1)$、证据因子 $f_X(x)$、联合概率 $f_{X,Y}(x,C_1)$

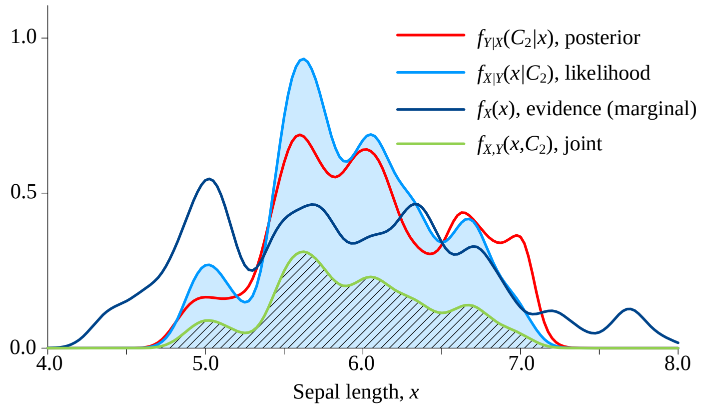


图 10. 比较后验概率 $f_{Y|X}(C_2\,|\,x)$、似然概率 $f_{X|Y}(x|C_2)$、证据因子 $f_X(x)$、联合概率 $f_{X,Y}(x,C_2)$

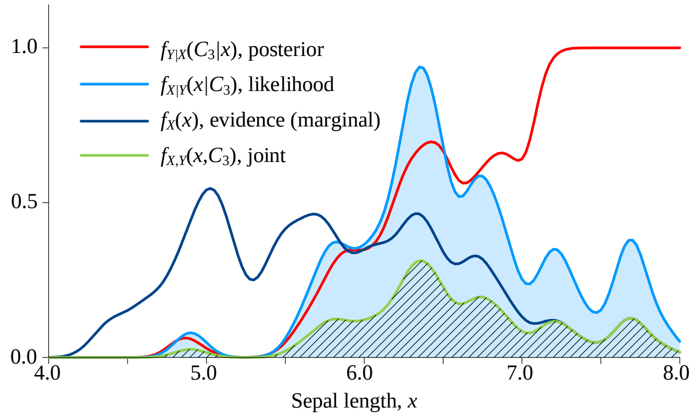


图 11. 比较后验概率 $f_{Y|X}(C_3\,|\,x)$、似然概率 $f_{X|Y}(x|C_3)$、证据因子 $f_X(x)$、联合概率 $f_{X,Y}(x,C_3)$

Bk5_Ch18_01.py 代码绘制本章前文大部分图像。

18.7 单一特征分类：基于 KDE

似然概率 → 联合概率

图 12 总结以花萼长度为单一特征，计算似然概率和联合概率的过程。

鸢尾花数据较为特殊，前文介绍过，鸢尾花数据共有 150 个数据点，C_1、C_2 和 C_3 三类各占 50，因此三个先验概率相等。因此，图 12 中，从似然概率密度 $f_{X|Y}(x\,|\,C_k)$ 到联合概率 $f_{X,Y}(x, C_k)$，高度缩放比例相同。一般情况下，相同缩放比例这种情况几乎不存在。

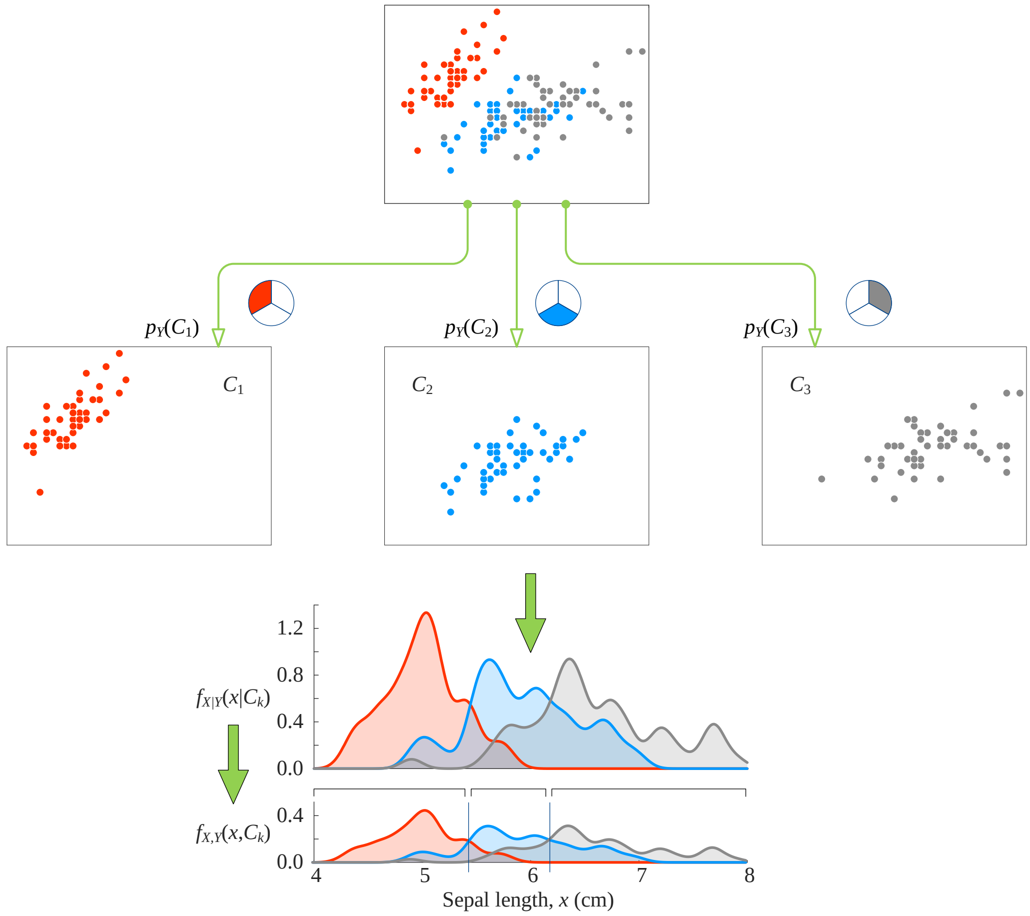


图 12. 似然概率到联合概率，花萼长度特征 x，基于 KDE

比较后验概率

有了本节前文联合概率和证据因子，我们可以获得后验概率密度曲线，如图 13。后验概率也叫成员值，后验概率更容易分类可视化。

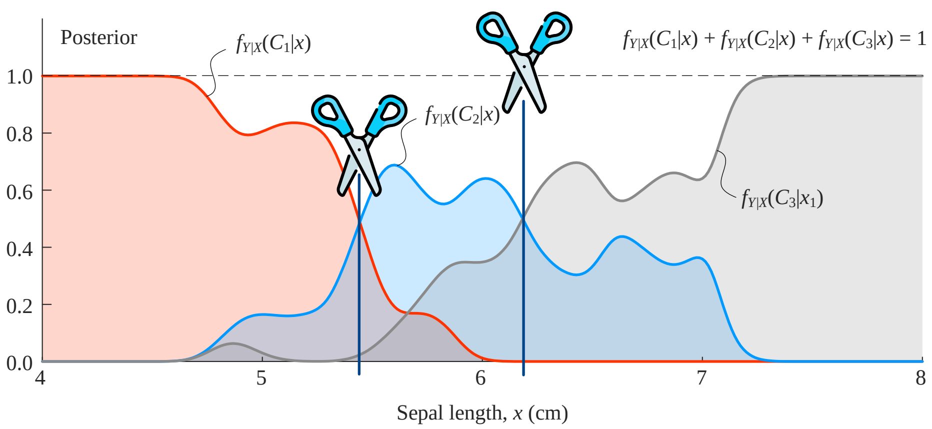


图 13. 后验概率，花萼长度特征，基于 KDE

举个例子

如图 14 所示，比较花萼长度特征后验概率大小，可以很容易预测 A、B、C、D 和 E 五点分类。A 的预测分类为 C_1；B 为决策边界；C 的预测分类为 C_2；D 为**决策边界** (decision boundary)；E 预测分类为 C_3。

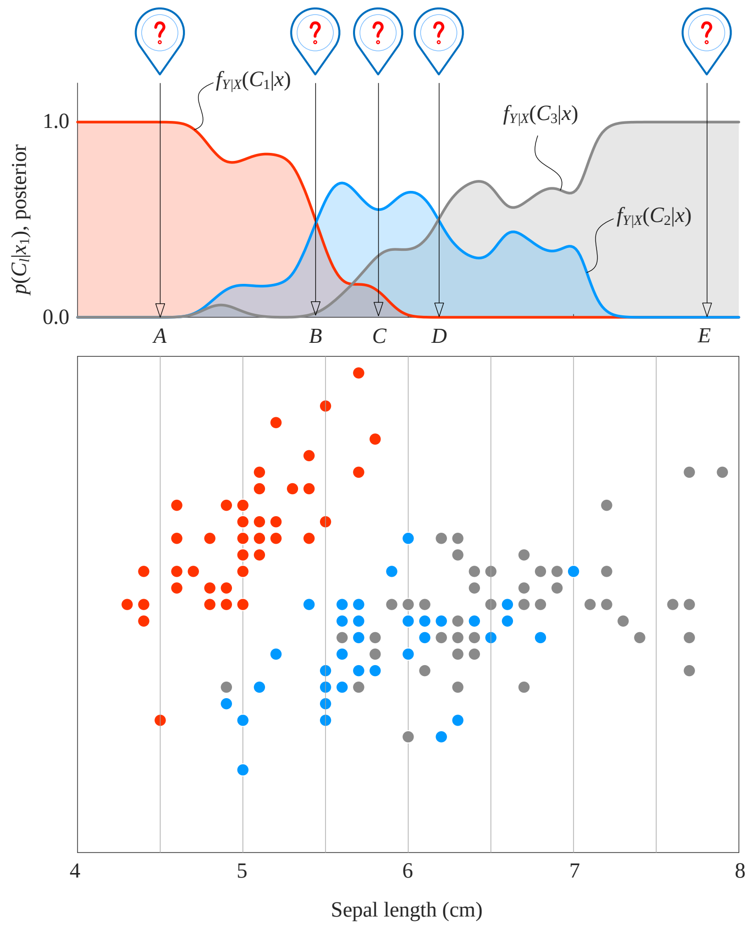


图 14. 利用花萼长度特征后验概率，进行分类预测

堆积直方图、饼图

图 15 所示为另外两种成员值 (后验概率) 的可视化方案——**堆积直方图** (stacked bar chart) 和**饼图** (pie chart)。通过这两个可视化方案，大家可以清楚看到不同类别成员值随特征变化。

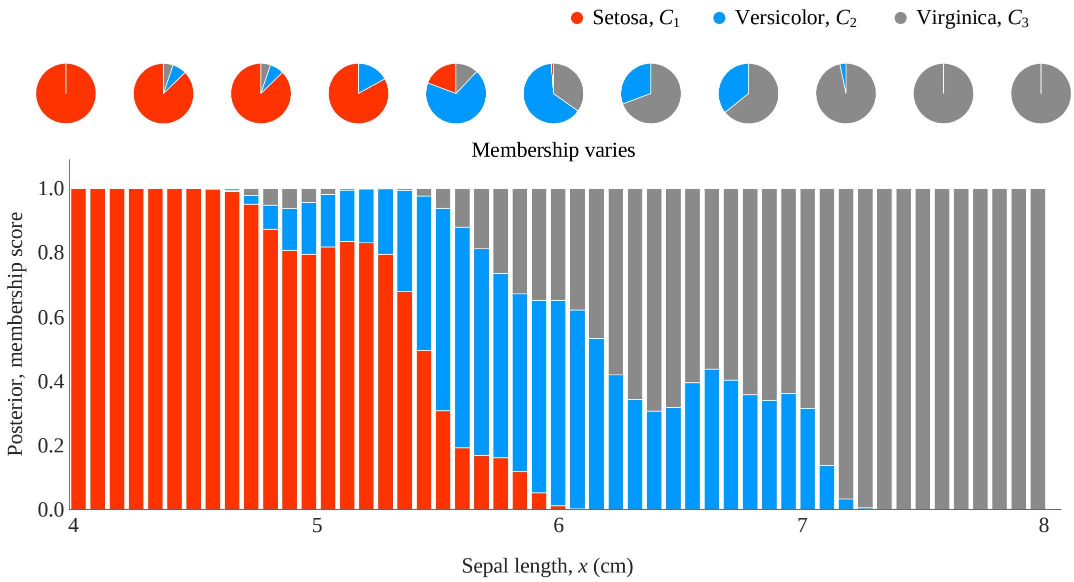


图 15. 堆积直方图和饼图，利用花萼长度特征成员值确定分类，基于 KDE

花萼宽度

本章前文都是基于花萼长度这个单一特征来判断鸢尾花分类，我们当然可以使用鸢尾花其他特征判断其分类。本节最后展示利用鸢尾花宽度作为依据判断鸢尾花分类。

图 16 所示为对于花萼宽度特征，从似然概率到联合概率的计算过程。

同理，比较花萼宽度特征的后验概率大小，可以决定图 17 中 A、B、C 和 D 点分类预测。A 的预测分类为 C_1；B 为决策边界；C 为决策边界；D 的预测分类为 C_2。

图 18 所示为利用花萼宽度特征成员值堆积直方图和饼图。大家可能会问，如何同时利用鸢尾花花萼长度、花萼宽度作为分类依据？这个问题，我们下一章回答。

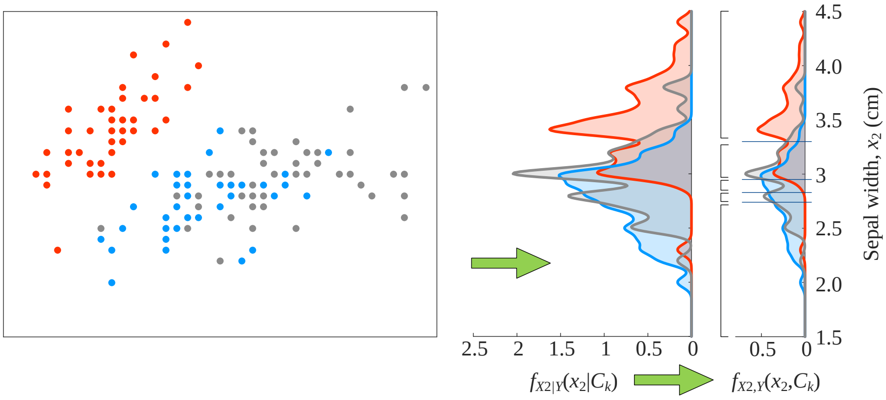


图 16. 似然概率到联合概率，花萼宽度特征 x_2，基于 KDE

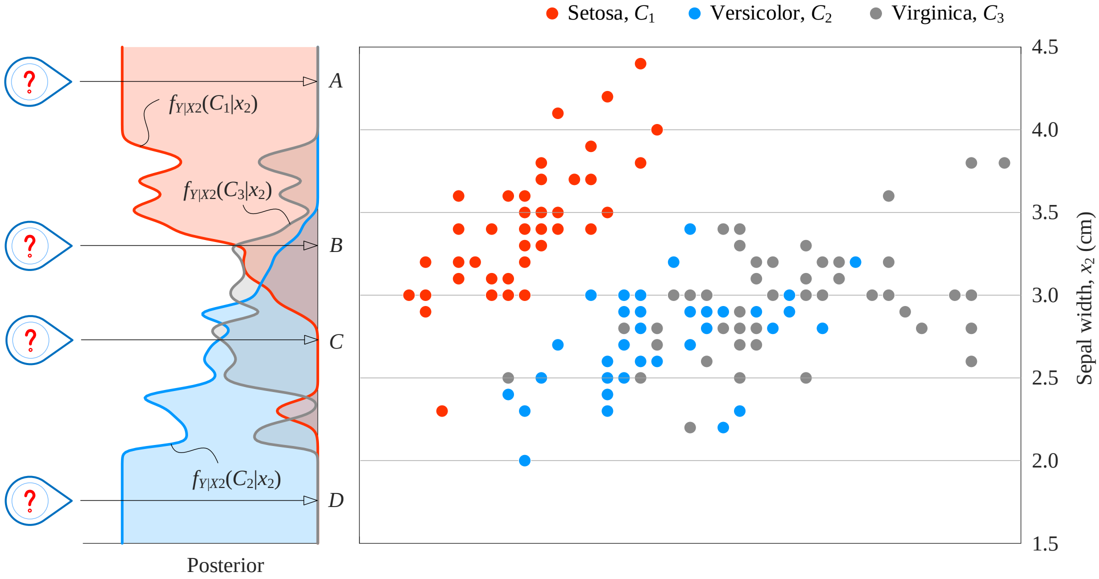


图 17. 利用花萼宽度特征后验概率，进行分类预测

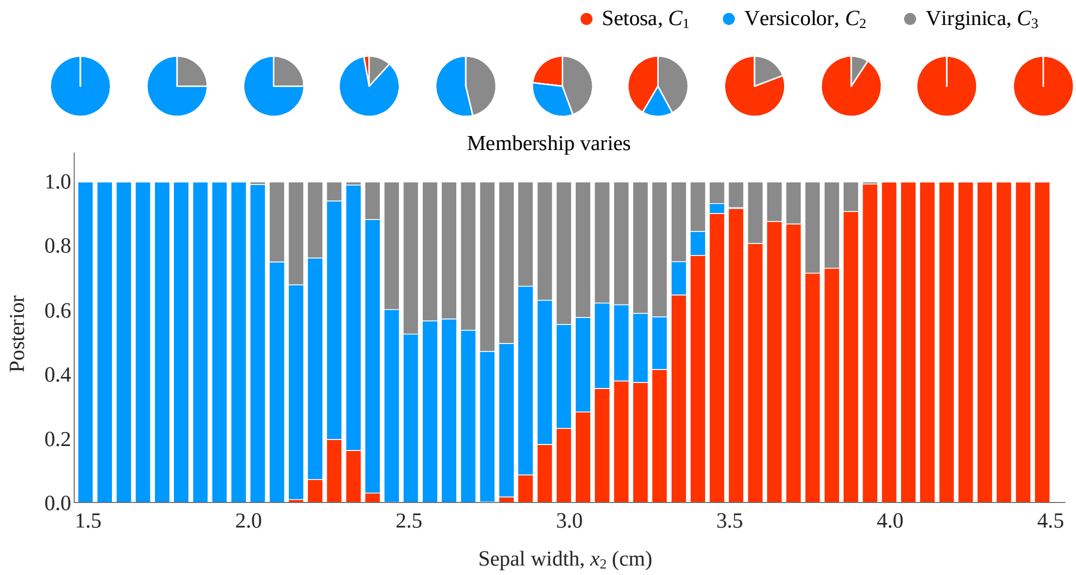


图 18. 堆积直方图和饼图，利用花萼宽度特征成员值确定分类，基于 KDE

18.8 单一特征分类：基于高斯

本章前文利用 KDE 方法估计似然概率，本章最后一节利用高斯分布估计似然概率。这一节，我们还是单独研究花萼长度特征 x_1、花萼宽度特征 x_2。

似然概率 → 联合概率

图 19 所示为花萼长度特征 x_1 上，利用高斯分布估算似然概率，然后计算联合概率；最后获得以特征 x_1 为依据决策边界。比较图 19 联合概率曲线高度，鸢尾花数据被划分为三个区域。这三个区域的位置和本章前文基于 KDE 估算稍有不同。

图 20 所示为花萼宽度特征 x_2 上同样过程。比较图 20 联合概率曲线高度，同样发现鸢尾花数据被划分为三个区域。

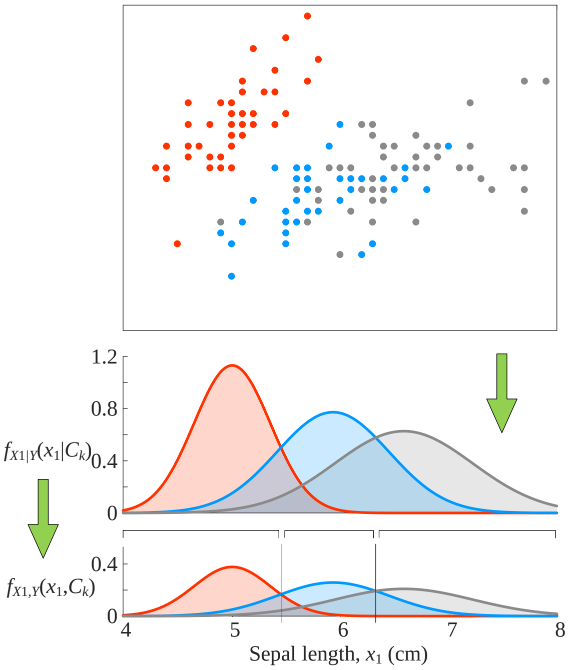


图 19. 似然概率到联合概率，花萼长度特征 x_1，基于高斯分布

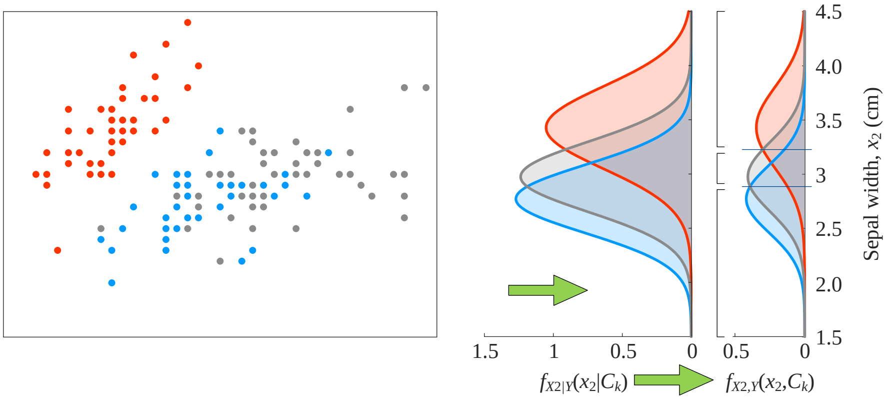


图 20. 似然概率到联合概率，花萼宽度特征 x_2，基于高斯分布

证据因子

图 21 和图 22 所示为利用全概率定理，获得 $f(x_1)$ 和 $f(x_2)$ 两个证据因子的概率密度函数。这实际上也是一种概率密度估算的方法。

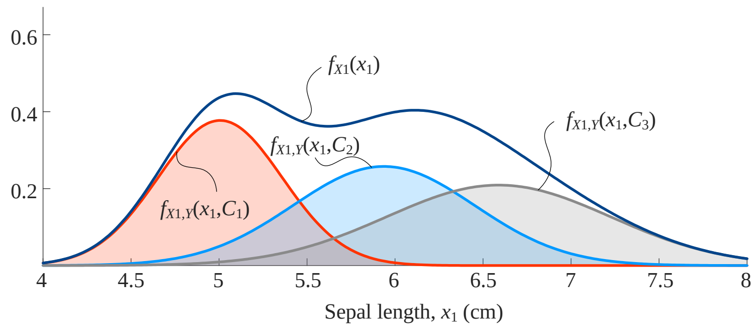


图 21. 证据因子/边缘概率，花萼长度特征 x_1，基于高斯分布

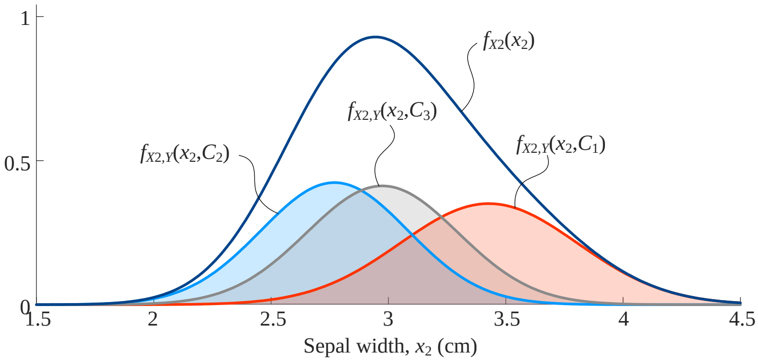


图 22. 证据因子/边缘概率，花萼宽度特征 x_2，基于高斯分布

后验概率

图 23 和图 24 比较两组后验概率曲线，以及如何据此得到的决策边界。

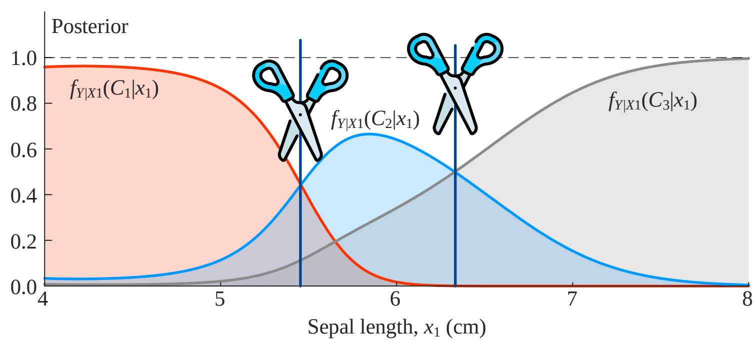


图 23. 后验概率，花萼长度特征 x_1，基于高斯分布

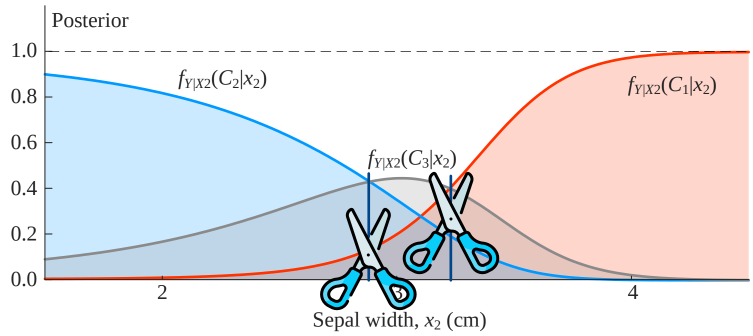


图 24. 后验概率，花萼宽度特征 x_2，基于高斯分布

后验概率：分类预测

图 25 所示为利用花萼长度特征后验概率曲线，进行分类预测。比较后验概率值大小可以判断：*A* 点预测分类为 C_1；*B* 点为 C_1 和 C_2 之间决策边界；*C* 点预测分类为 C_2；*D* 点为 C_2 和 C_3 之间决策边界；*E* 点预测分类为 C_3。

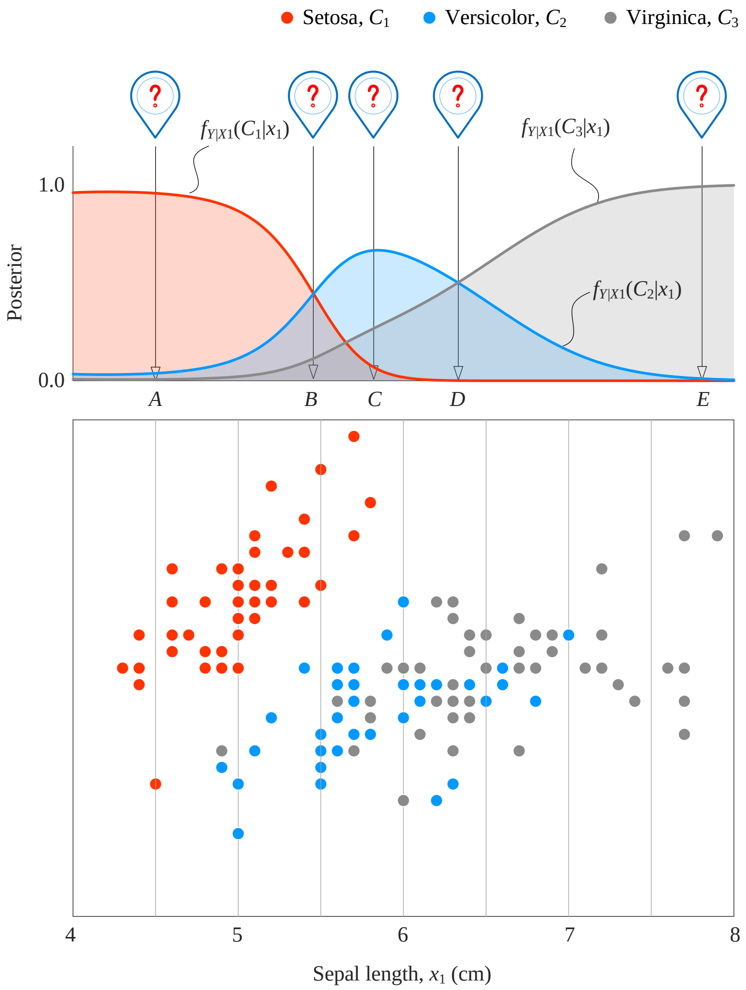


图 25. 利用花萼长度特征后验概率，进行分类预测

图 26 所示为利用花萼宽度特征后验概率曲线，进行分类预测。比较后验概率值大小可以判断：A 点预测分类为 C_1；B 点预测分类为 C_3；C 点为 C_2 和 C_3 之间决策边界；D 点预测分类为 C_2。

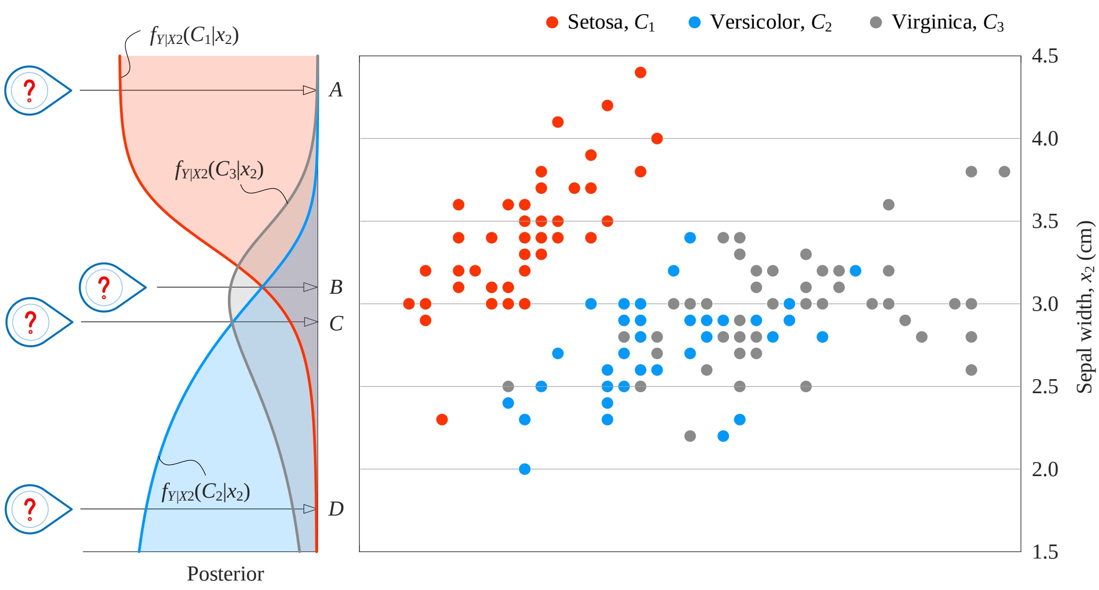


图 26. 利用花萼宽度特征后验概率，进行分类预测

图 27 和图 28 所示为利用堆积直方图和饼图表达成员值/后验概率随特征变化。对比图 15 和图 18，可以发现，基于高斯分类的成员值/后验概率变化过程更为平滑。

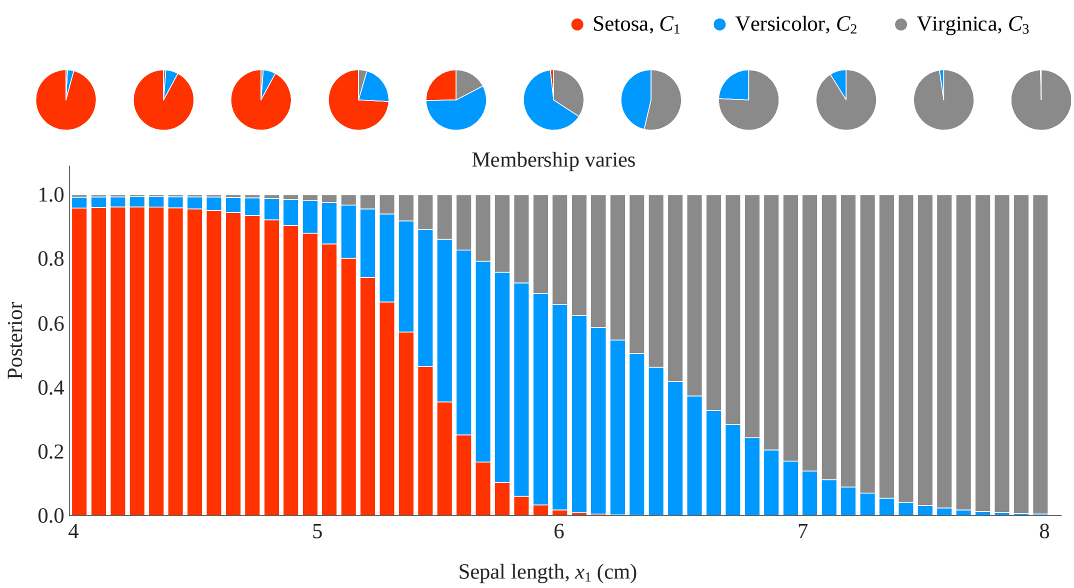


图 27. 堆积直方图和饼图，利用花萼长度特征成员值确定分类，基于高斯分布

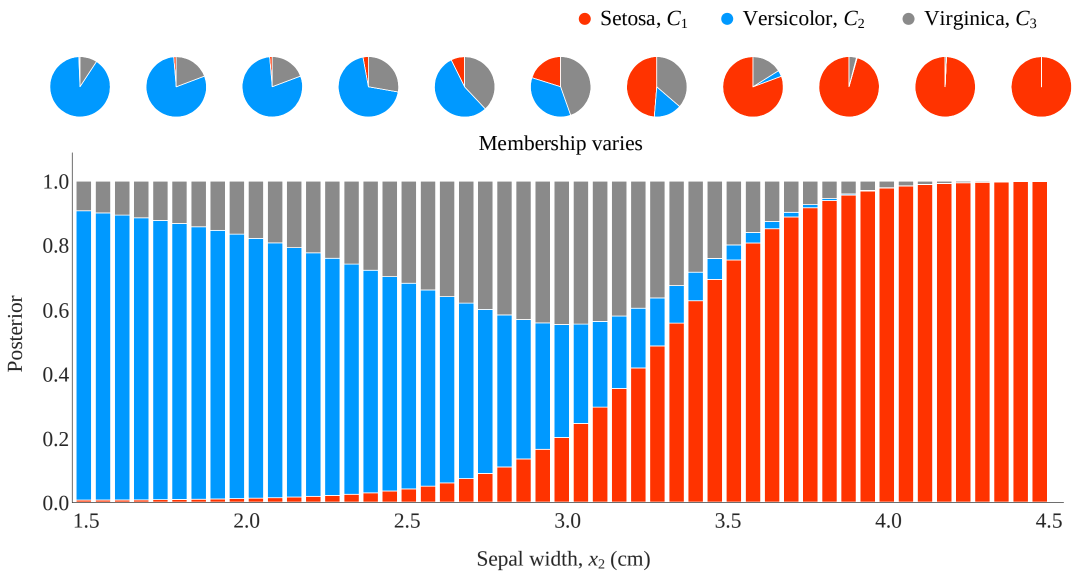


图 28. 堆积直方图和饼图，利用花萼宽度特征成员值确定分类，基于高斯分布

这一章中，大家必须要掌握的是贝叶斯定理中的先验概率、后验概率、证据因子、似然概率等概念。而贝叶斯分类是一种基于贝叶斯定理的分类方法。在贝叶斯分类算法中，优化问题可以最大化后验概率，也可以最大化联合概率，即“先验 × 似然”。

请大家务必掌握比例关系——后验 ∝ 先验 × 似然。这是贝叶斯推断中最重要的比例关系。

下一章，我们将分类的依据从单一特征提高到二维，让大家更清楚地看到先验概率、后验概率、证据因子、似然概率的“样子”。