5 高斯朴素贝叶斯

Gaussian Naïve Bayes

假设特征之间条件独立，且条件概率分布服从高斯分布

给我带来极大的喜悦的是，求知的过程，而不是占有知识。

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment.

—— 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

- `matplotlib.pyplot.contour()` 绘制等高线线图
- `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.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

5.1 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 × 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 $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 两个后验概率大小，就可以预测分类。

图 1 所示为基于高斯分布得到的 $f_{Y|\chi}(C_1 \mid \boldsymbol{x})$ 和 $f_{Y|\chi}(C_2 \mid \boldsymbol{x})$ 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 1 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。

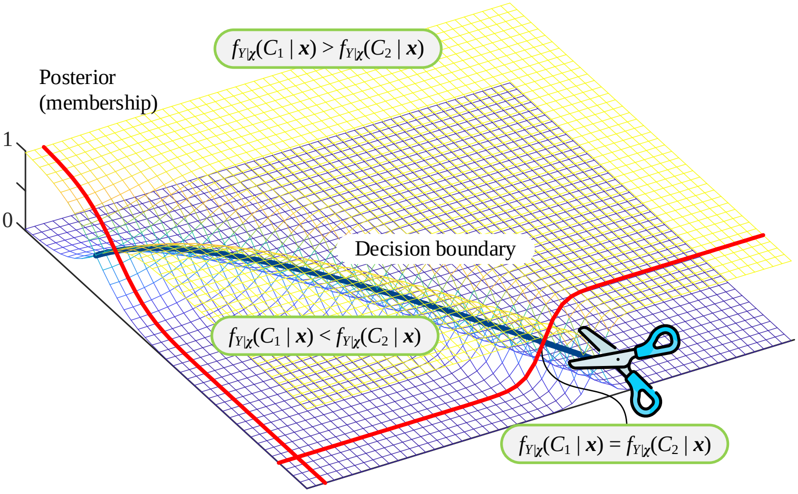

图 1. 二分类，比较后验概率大小，基于高斯分布

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 2 所示为，二分类问题中，比较联合概率 $f_{Y,\chi}(C_1, \boldsymbol{x})$ 和 $f_{Y,\chi}(C_2, \boldsymbol{x})$ 两个曲面高度，可以获得相同的决策边界。

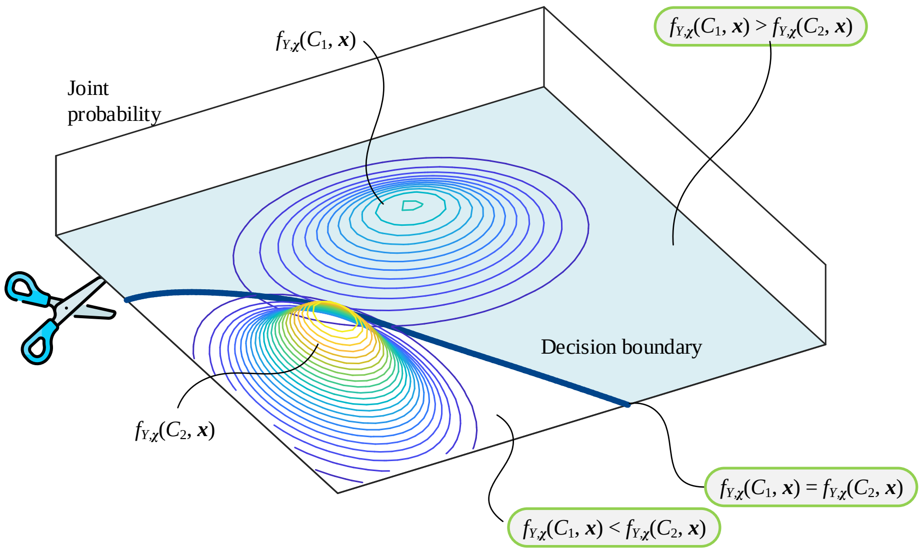

图 2. 二分类，比较联合概率大小，基于高斯分布

流程

图 3 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。

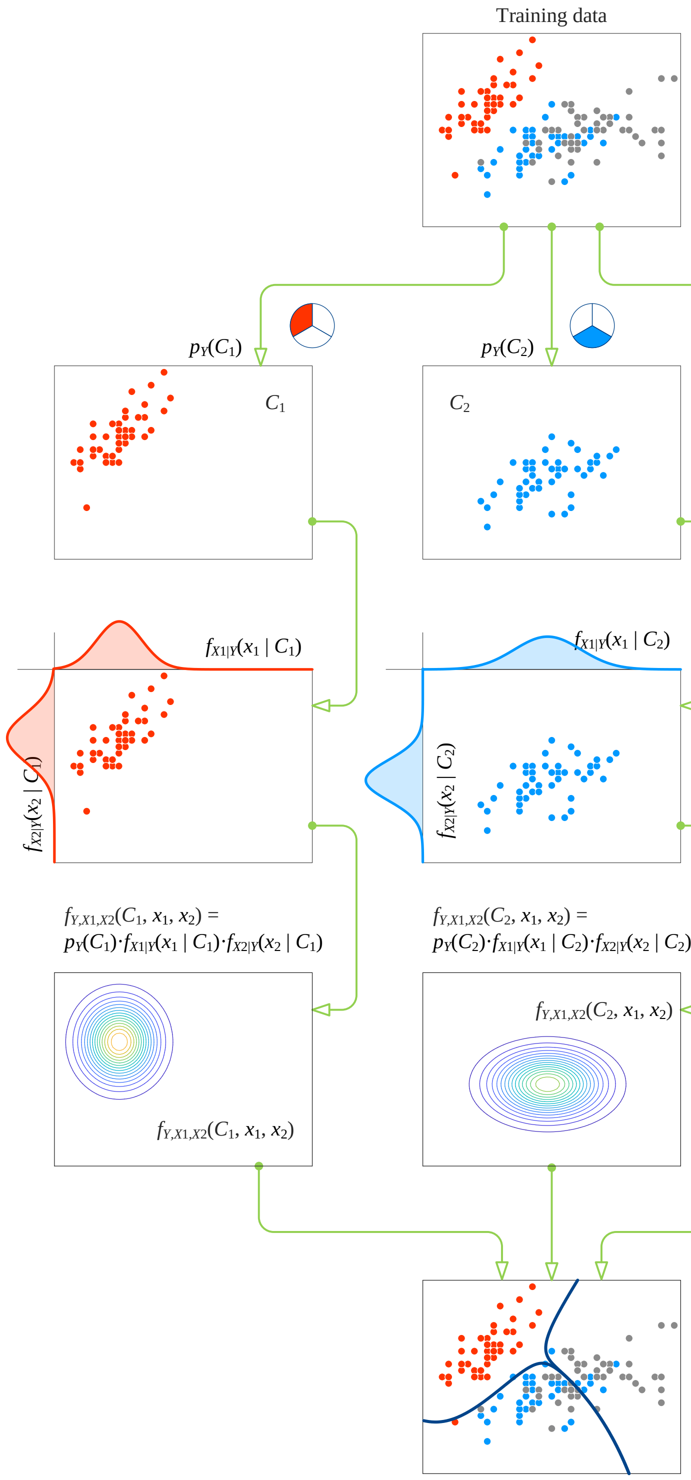

图 3. 高斯朴素贝叶斯分类过程

5.2 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：

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

比如，下式计算 C_1 似然概率密度：

$$\underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_1\right)}_{\text{Likelihood}}=\underbrace{f_{X1|Y}\left(x_1|C_1\right)f_{X2|Y}\left(x_2|C_1\right)}_{\text{Conditional independence}} \tag{2}$$

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 $f_{X1|Y}(x_1 | C_1)$ 和 $f_{X2|Y}(x_2 | C_1)$ 可以写成：

$$\begin{cases} f_{X1|Y}\left(x_1|C_1\right)=\dfrac{1}{\sqrt{2\pi}\sigma_{1|C1}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x_1-\mu_{1|C1}}{\sigma_{1|C1}}\right)^2\right) \\ f_{X2|Y}\left(x_2|C_1\right)=\dfrac{1}{\sqrt{2\pi}\sigma_{2|C1}}\exp\left(-\dfrac{1}{2}\left(\dfrac{x_2-\mu_{2|C1}}{\sigma_{2|C1}}\right)^2\right) \end{cases} \tag{3}$$

对于鸢尾花数据，$\mu_{1|C1}$ 为标签为 C_1 数据在花萼长度 x_1 特征上的均值，$\sigma_{1|C1}$ 为 C_1 数据在 x_1 特征上标准差；$\mu_{2|C1}$ 为标签为 C_1 数据在花萼宽度 x_2 特征上的均值，$\sigma_{2|C1}$ 为 C_1 数据在 x_2 特征上标准差。

图 4 中给出 $f_{X1|Y}(x_1 | C_1)$ 和 $f_{X2|Y}(x_2 | C_1)$ 两个概率密度函数曲线，以及 $\mu_{1|C1}$ 和 $\mu_{2|C1}$ 所在位置。

将 (3) 代入 (2)，可以得到 $f_{X1,X2|Y}(x_1, x_2| C_1)$：

$$\begin{aligned} f_{X1,X2|Y}\left(x_1,x_2|C_1\right)\cdot &= f_{X1|Y}\left(x_1|C_1\right)\cdot f_{X2|Y}\left(x_2|C_1\right) \\ &= \frac{\exp\left(-\dfrac{1}{2}\left(\dfrac{x_1-\mu_{1|C1}}{\sigma_{1|C1}}\right)^2\right)}{\sqrt{2\pi}\sigma_{1|C1}}\times\frac{\exp\left(-\dfrac{1}{2}\left(\dfrac{x_2-\mu_{2|C1}}{\sigma_{2|C1}}\right)^2\right)}{\sqrt{2\pi}\sigma_2} \\ &= \frac{\exp\left(-\dfrac{1}{2}\left(\dfrac{\left(x_1-\mu_{1|C1}\right)^2}{\sigma_{1|C1}^2}+\dfrac{\left(x_2-\mu_{2|C1}\right)^2}{\sigma_{2|C1}^2}\right)\right)}{\left(\sqrt{2\pi}\right)^2\sigma_{1|C1}\sigma_{2|C1}} \end{aligned} \tag{4}$$

图 4 中等高线便是 $f_{X1,X2|Y}(x_1, x_2| C_1)$ 曲面等高线。

大家可能已经发现，图 4 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。

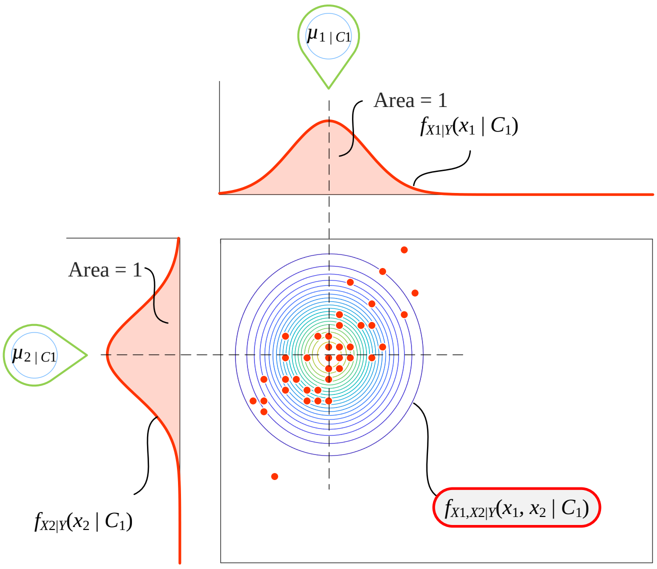

图 4. 分类 C_1 样本数据，假设鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率 $f_{X1,X2|Y}(x_1, x_2\,|\,C_1)$，基于高斯分布

图 5 和图 6 所示为似然概率 $f_{X1,X2|Y}(x_1, x_2\,|\,C_2)$ 和 $f_{X1,X2|Y}(x_1, x_2\,|\,C_3)$ 结果。

⚠ 再次提醒大家注意，“特征条件独立”不同于“特征独立”。

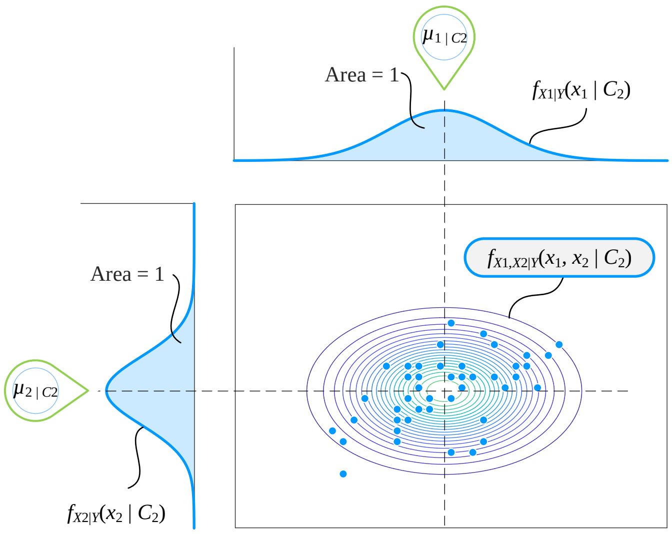

图 5. 分类 C_2 样本数据，假设鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率 $f_{X1,X2|Y}(x_1, x_2\,|\,C_2)$，基于高斯分布

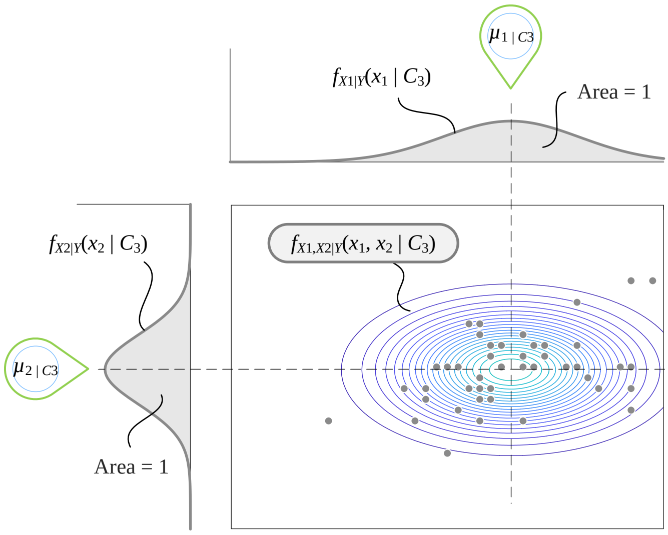

图 6. 分类 C_3 样本数据，假设鸢尾花花萼长度 x_1 和花萼宽度 x_2 条件独立，得到似然概率 $f_{X1,X2|Y}(x_1, x_2 \mid C_3)$，基于高斯分布

5.3 联合概率

这一节利用下式估算联合概率：

$$\underbrace{f_{\chi,Y}\left(\boldsymbol{x},C_k\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}\underbrace{f_{\chi|Y}\left(\boldsymbol{x}\middle|C_k\right)}_{\text{Likelihood}}=\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}\underbrace{\prod_{j=1}^{D}f_{X_j|Y}\left(x_j\middle|C_k\right)}_{\text{Conditional independence}} \tag{5}$$

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：

$$\underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_k\right)}_{\text{Joint}}=\underbrace{f_{X1,X2|Y}\left(x_1,x_2\middle|C_k\right)}_{\text{Likelihood}}\underbrace{p_Y\left(C_k\right)}_{\text{Prior}}=\underbrace{f_{X1|Y}\left(x_1\middle|C_k\right)\cdot f_{X2|Y}\left(x_2\middle|C_k\right)}_{\text{Conditional independence}}\underbrace{p_Y\left(C_k\right)}_{\text{Prior}} \tag{6}$$

再次注意，先验概率 $p_Y(C_k)$ 相当于一个缩放系数。

图 7 ~ 图 9 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。

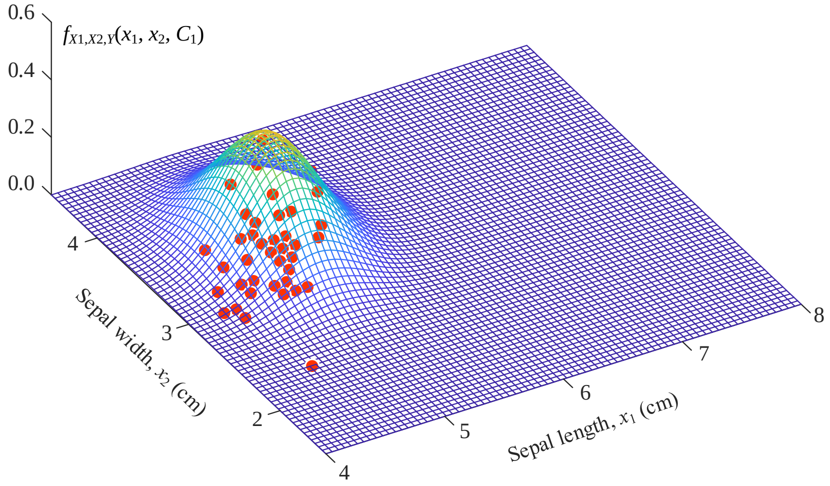

图 7. $f_{X1,X2,Y}(x_1, x_2, C_1)$ 概率密度曲面，基于高斯分布

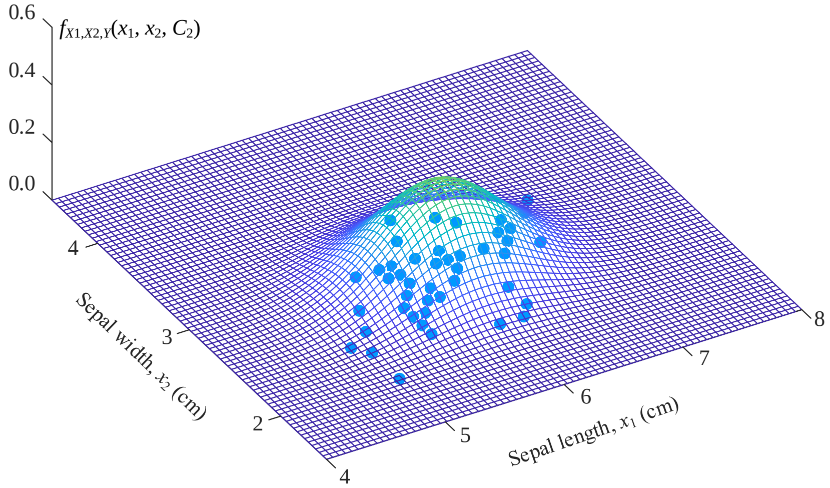

图 8. $f_{X1,X2,Y}(x_1, x_2, C_2)$ 概率密度曲面，基于高斯分布

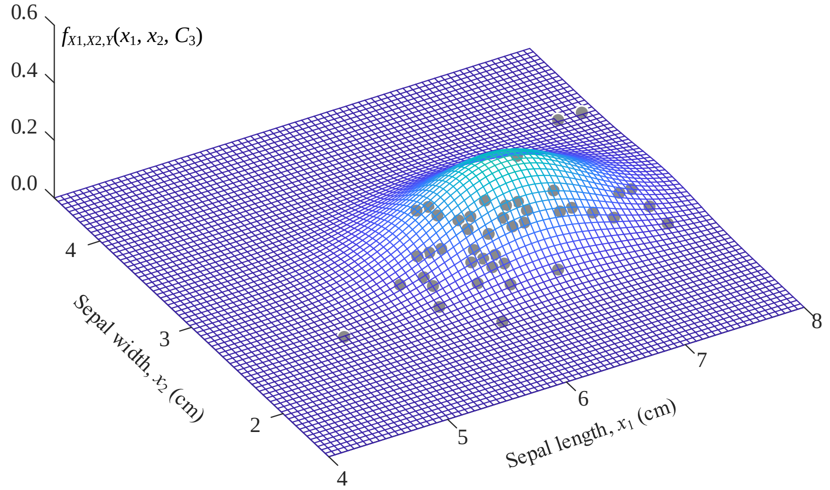

图 9. $f_{X1,X2,Y}(x_1, x_2, C_3)$ 概率密度曲面，基于高斯分布

分类依据：最大化联合概率

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：

$$\hat{y} = \arg\max_{C_k} p_Y(C_k)\prod_{j=1}^{D} f_{X_j|Y}(x_j|C_k) \tag{7}$$

因此，比较图 7 ~ 图 9 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 mehods 为 `fit(`$\boldsymbol{X}$, $\boldsymbol{y}$`)` 和 `predit(`$\boldsymbol{q}$`)`。`fit(`$\boldsymbol{X}$, $\boldsymbol{y}$`)`用来加载样本数据，`predit(`$\boldsymbol{q}$`)`用来预测查询点 $\boldsymbol{q}$ 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 10 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。

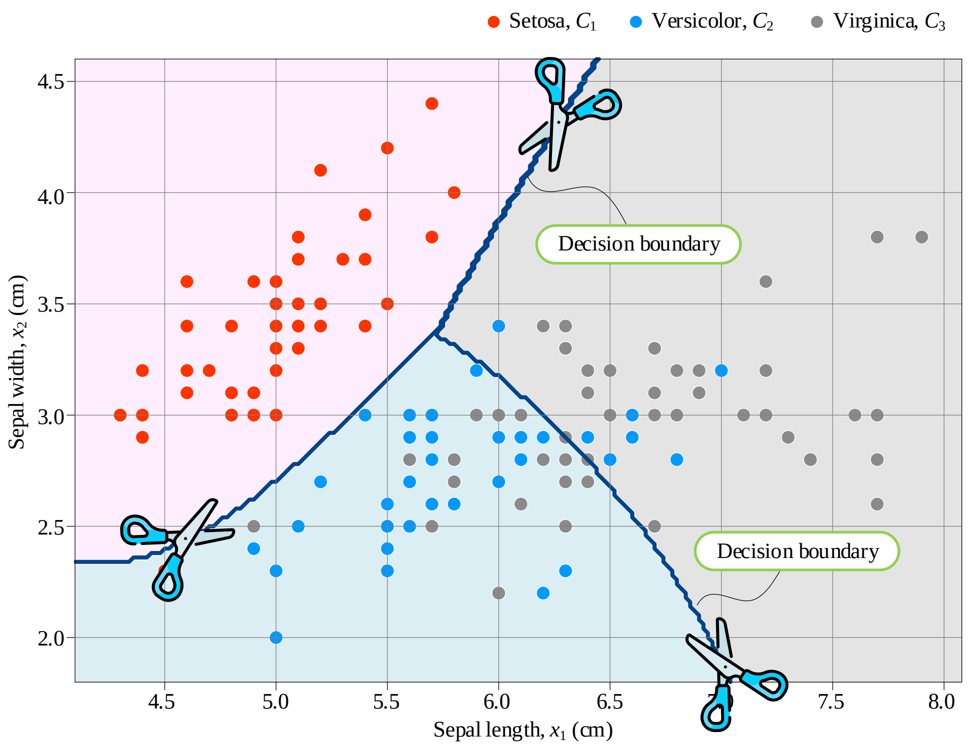

图 10. 鸢尾花分类预测，朴素贝叶斯决策边界，基于高斯分布

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 10。

5.4 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 $f(\boldsymbol{x})$ 可以这样计算：

$$\underbrace{f_{\chi}(\boldsymbol{x})}_{\text{Evidence}}=\sum_{k=1}^{K}\left\{\underbrace{p_Y(C_k)}_{\text{Prior}}\underbrace{\prod_{j=1}^{D}f_{X_j|Y}(x_j|C_k)}_{\text{Conditional independence}}\right\} \tag{8}$$

上一章提到，上式本身是一种概率密度估算方法，具体如图 11 所示。

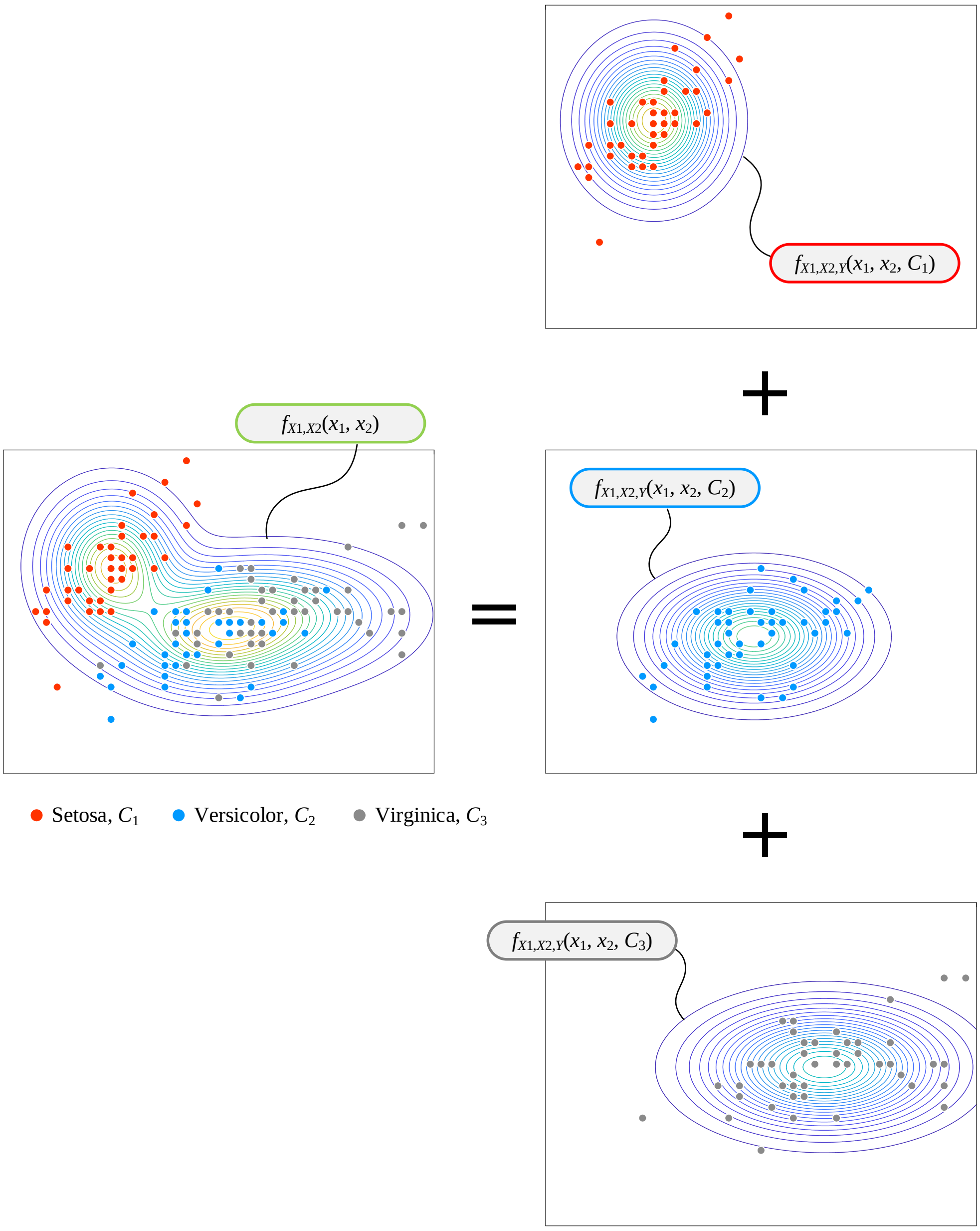

图 11. 估算证据因子概率密度，基于高斯分布

图 12 所示为估算得到的二元概率密度曲面 $f_{X1,X2}(x_1, x_2)$。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。

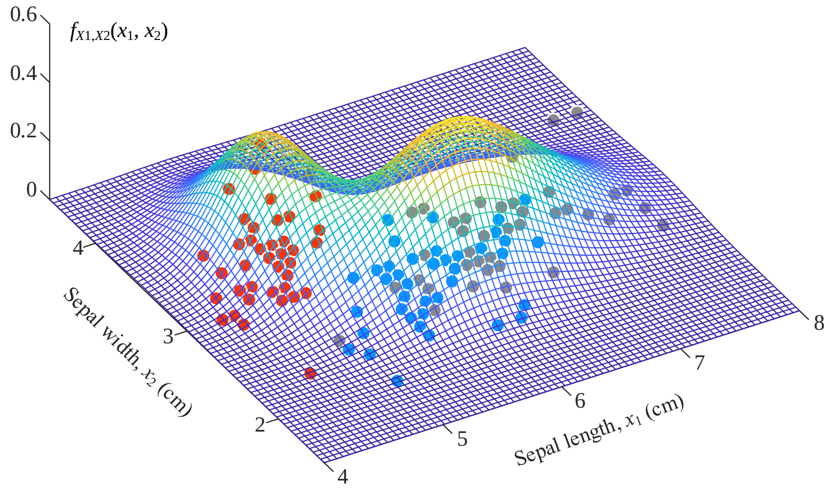

图 12. 估算得到的二元概率密度曲面，特征条件独立，基于高斯分布

5.5 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：

$$\underbrace{f_{Y|\chi}\left(C_k \mid \boldsymbol{x}\right)}_{\text{Posterior}} = \frac{\overbrace{f_{\chi|Y}\left(\boldsymbol{x} \mid C_k\right)}^{\text{Likelihood}} \overbrace{p_Y\left(C_k\right)}^{\text{Prior}}}{\underbrace{f_\chi\left(\boldsymbol{x}\right)}_{\text{Evidence}}} \tag{9}$$

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：

$$\underbrace{f_{Y|\chi}\left(C_k \mid \boldsymbol{x}\right)}_{\text{Posterior}} = \frac{\overbrace{\prod_{j=1}^{D} f_{X_j|Y}\left(x_j \mid C_k\right)}^{\text{Likelihood}} \overbrace{p_Y\left(C_k\right)}^{\text{Prior}}}{\underbrace{f_\chi\left(\boldsymbol{x}\right)}_{\text{Evidence}}} \tag{10}$$

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：

$$\hat{y} = \underset{C_k}{\arg\max}\ f_{Y|\chi}\left(C_k \mid \boldsymbol{x}\right) \tag{11}$$

图 13 ~ 图 15 所示为 $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$、$f_{Y|X1,X2}(C_2 \mid x_1, x_2)$ 和 $f_{Y|X1,X2}(C_3 \mid x_1, x_2)$ 三个后验概率曲面。

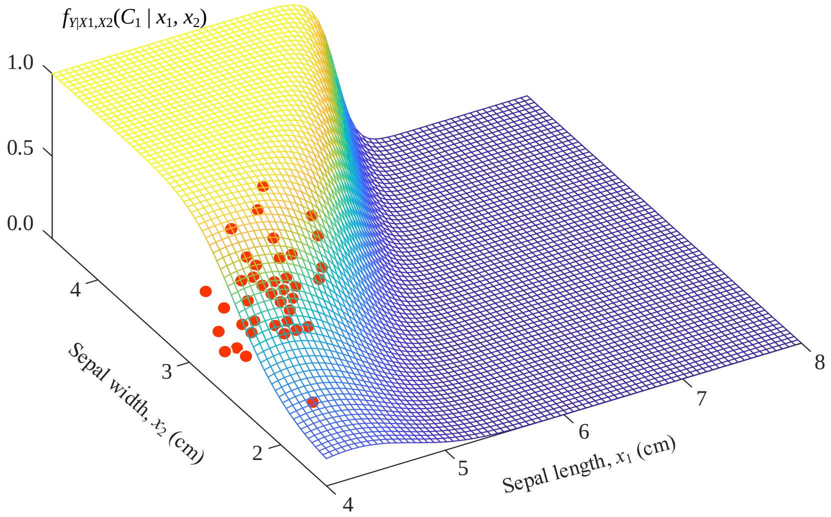

图 13. $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ 后验概率曲面，基于高斯分布

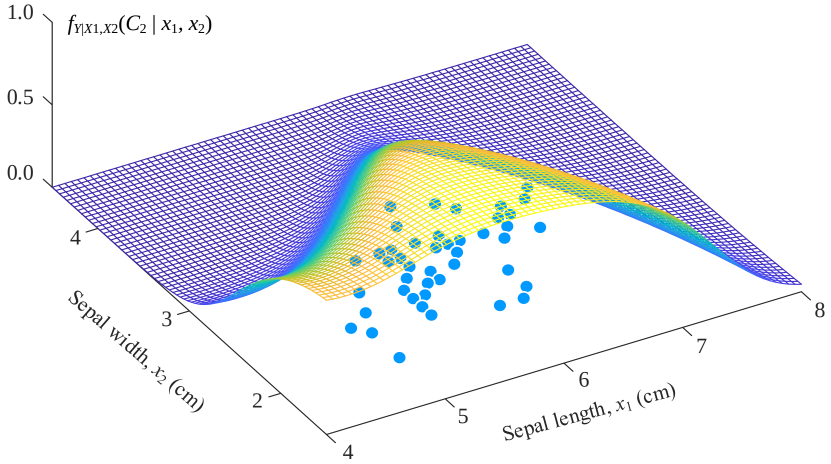

图 14. $f_{Y|X1,X2}(C_2 \mid x_1, x_2)$ 后验概率曲面，基于高斯分布

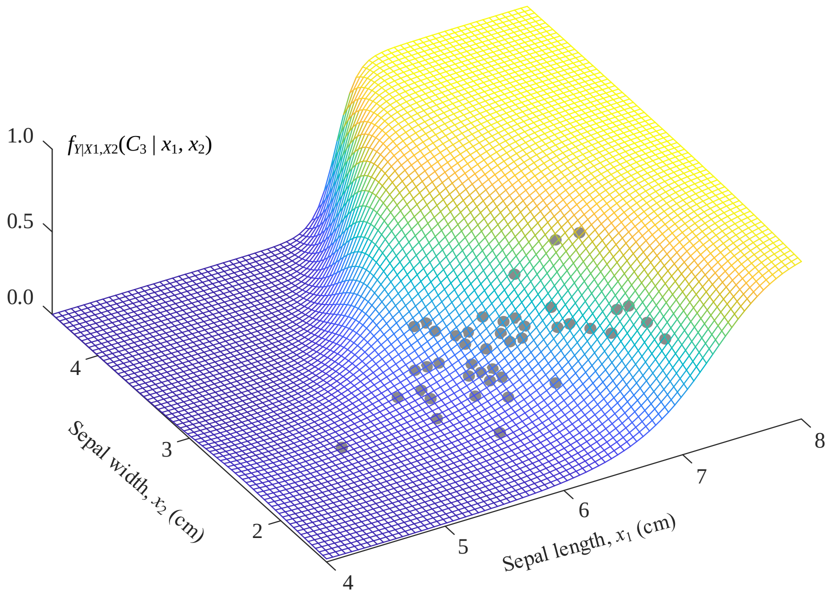

图 15. $f_{Y|X1,X2}(C_3 \mid x_1, x_2)$ 后验概率曲面，基于高斯分布

对于鸢尾花三分类问题，如图 16 所示，比较 $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$、$f_{Y|X1,X2}(C_2 \mid x_1, x_2)$ 和 $f_{Y|X1,X2}(C_3 \mid x_1, x_2)$ 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。

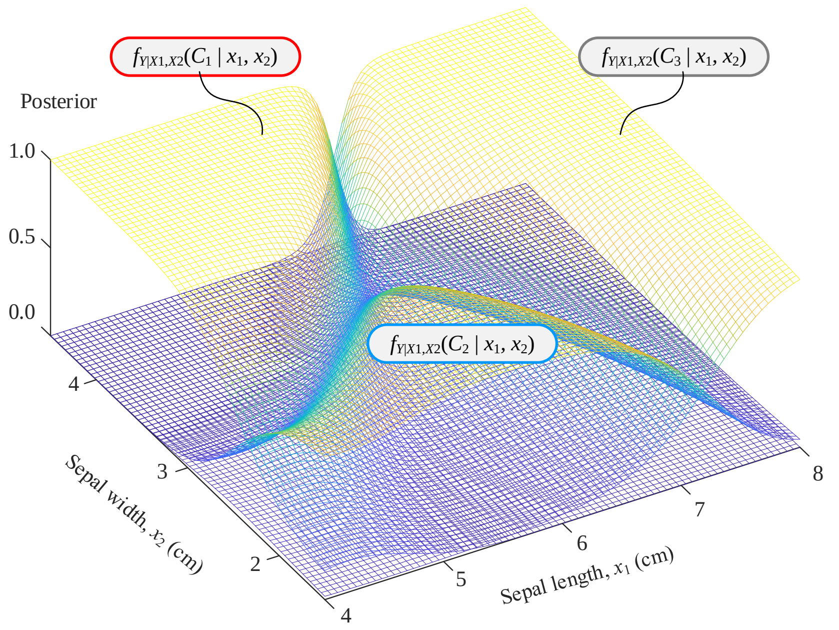

图 16. 比较三个后验概率曲面，基于高斯分布

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。