6 连续随机变量

Continuous Random Variables

PDF 积分得到边缘概率密度或概率

> 上帝不仅玩骰子，他还有时把骰子扔到人类看不见的地方。
>
> ***Not only does God definitely play dice, but He sometimes confuses us by throwing them where they can't be seen.***
>
> —— 史蒂芬·霍金 (Stephen Hawking) | 英国理论物理学家、宇宙学家 | 1942 ~ 2018

- ◀ matplotlib.pyplot.contour() 绘制平面等高线
- ◀ matplotlib.pyplot.contour3D() 绘制三维等高线
- ◀ 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() 绘制三维单色线框图
- ◀ matplotlib.pyplot.scatter() 绘制散点图
- ◀ scipy.stats.st.gaussian_kde() 高斯 KDE 函数
- ◀ seaborn.scatterplot() 绘制散点图
- ◀ statsmodels.api.nonparametric.KDEUnivariate() 一元核密度估计

6.1 一元连续随机变量

本书第 4 章区分过**离散随机变量** (discrete random variable)、**连续随机变量** (continuous random variable)。如果随机变量 X 的所有可能取值不可以逐个列举出来，而是取数轴上某一区间内的任一点的随机变量，我们就称 X 为连续随机变量。

概率密度函数：积分

本书第 4 章介绍过，离散随机变量对应的数学工具为求和Σ，连续随机变量对应积分∫。对于连续随机变量 X，如果存在非负函数 $f_X(x)$ 使得：

$$\Pr(X \in B) = \int_B f_X(x)\,\mathrm{d}x \tag{1}$$

则称函数 $f_X(x)$ 为 X 的**概率密度函数** (probability density function, PDF)。

特别地，如图 1 所示，当 B 为区间 $[a, b]$ 时，随机变量 X 的概率对应定积分：

$$\Pr(a \le X \le b) = \int_a^b f_X(x)\,\mathrm{d}x \tag{2}$$

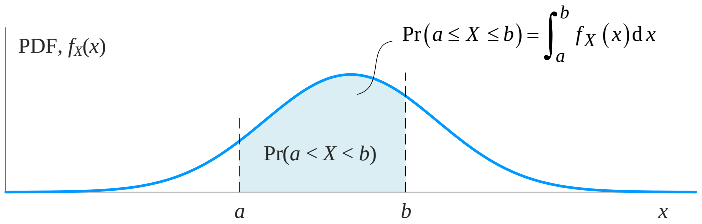

图 1. 定积分常用来计算一元连续随机变量在一定区间对应的概率

此外，本书前文提到过，PMF 和 PDF 的输入都可能是不止一个随机变量，这和多元函数一样。比如，二元离散随机变量 (X, Y) 联合概率质量函数 PMF $p_{X,Y}(x,y)$ 有两个变量，三元连续随机变量 (X_1, X_2, X_3) 的联合概率密度函数 PDF $f_{X1,X2,X3}(x_1,x_2,x_3)$ 有三个变量。

概率密度非负，面积为 1

任意概率密度函数 $f_X(x)$，必须是非负的 $f_X(x) \geqslant 0$，且满足：

$$\Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1 \tag{3}$$

上式常简写为：

$$\int_x f_X(x)\,\mathrm{d}x = 1 \tag{4}$$

如图 2 所示，从图像上来看，$f_X(x)$ 曲线和整个横轴包围区域的面积为 1。

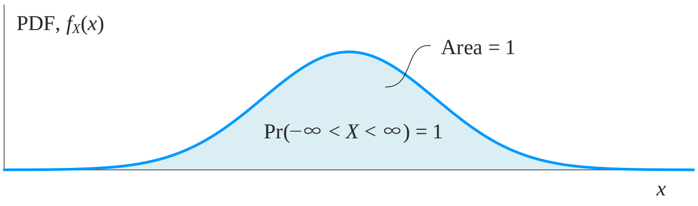

图 2. $f_X(x)$ 和横轴围成图形的面积为 1

单点集合：概率密度非负，但是概率为 0

利用数值积分方法，X 的取值范围在 $[a, a + \Delta]$ 对应的概率为：

$$\Pr(a \le X \le a+\Delta) = \int_a^{a+\Delta} f_X(x)\,\mathrm{d}x \approx f_X(a)\Delta \tag{5}$$

当 $\Delta \to 0$ 时，$\Pr(a \leqslant X \leqslant a + \Delta) \to 0$。

也就是说，对于单点集合，$X = a$ 的概率为 0：

$$\Pr(X = a) = \int_a^a f_X(x)\,\mathrm{d}x = 0 \tag{6}$$

区间端点

因此，对于连续随机变量，区间端点对概率计算不起任何作用，因此以下四个概率值等价：

$$\mathrm{P}(a \le X \le b) = \mathrm{P}(a < X \le b) = \mathrm{P}(a \le X < b) = \mathrm{P}(a < X < b) \tag{7}$$

概率密度值可以大于 1

再次强调 $f_X(x)$ 并不是概率，而是概率密度，因此 $f_X(x)$ 可大于 1。

比如，图 3 所示在 [0, 0.5] 区间上连续均匀分布的概率密度函数 $f_X(x)$。很明显，$f_X(x)$ 的最大值为 2，但是长方形的面积仍为 1：

$$\begin{aligned}\Pr(-\infty < X < \infty) &= \int_{-\infty}^{0} f_X(x)\,\mathrm{d}x + \int_0^{0.5} f_X(x)\,\mathrm{d}x + \int_{0.5}^{\infty} f_X(x)\,\mathrm{d}x \\ &= 0 + \int_0^{0.5} 2\,\mathrm{d}x + 0 \\ &= 2x\big|_0^{0.5} = 1\end{aligned} \tag{8}$$

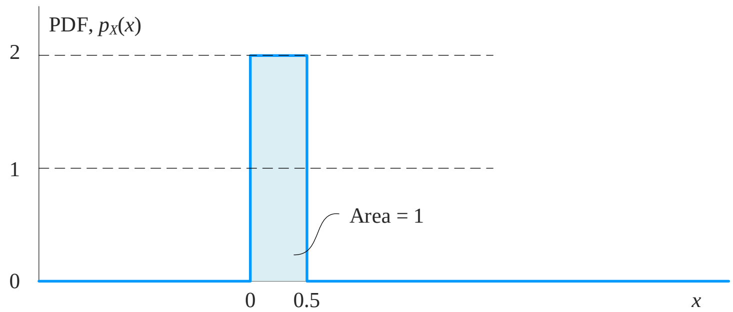

图 3. 概率密度函数 $f_X(x)$ 可以大于 1

累积分布函数

本书前文介绍，给定一元离散随机变量 X 的概率质量函数 $p_X(x)$，求解其 CDF 时，用的是累加 Σ。

以图 4 (a) 为例，对于一元连续随机变量 X，求累积分布函数 CDF $F_X(x)$ 用的是积分，也就是求面积：

$$F_X(x) = \Pr(X \le x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \tag{9}$$

图 4 (a) 中 $f_X(x)$ 图形的面积对应概率值，而图 4 (b) 中 $F_X(x)$ 的高度对应概率值。

随机变量 X 在$[a, b]$ 区间对应的概率可以用 CDF $F_X(x)$ 计算：

$$\Pr(a \le X \le b) = F_X(b) - F_X(a) \tag{10}$$

再次强调，对于一元连续随机变量，PDF 是概率密度，CDF 是概率。

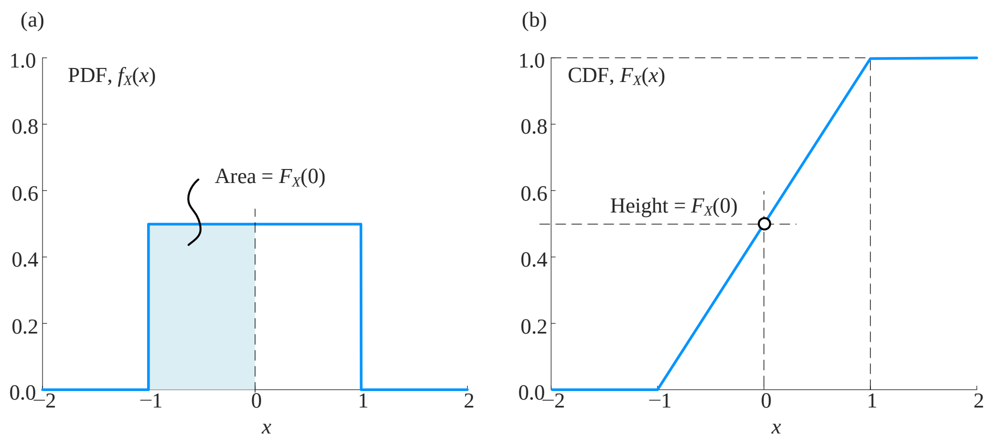

图 4. 连续均匀分布 PDF 和 CDF

6.2 期望、方差和标准差

期望值

连续随机变量 X 期望定义如下：

$$\mathrm{E}(X)=\int_{-\infty}^{\infty} x \cdot \underbrace{f_X(x)}_{\text{Weight}} \mathrm{d} x \tag{11}$$

上式也相当于加权平均。其中，$f_X(x)$ 相当于是“权重”。显然，$f_X(x)$ 非负，但是 x 取值可正可负。这也就是说，$\mathrm{E}(X)$ 可正可负。

(11) 常简写为：

$$\boxed{\mathrm{E}(X)=\int_{x} x \cdot f_X(x) \mathrm{d} x} \tag{12}$$

权重当然满足 $\int_{x} f_X(x) \mathrm{d} x=1$。

连续均匀分布

如图 5 所示，如果随机变量 X 在 $[a, b]$ 上服从**连续均匀分布** (continuous uniform distribution)，X 的概率密度函数为：

$$f_X(x)=\begin{cases}\dfrac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x<a \text{ or } x>b\end{cases} \tag{13}$$

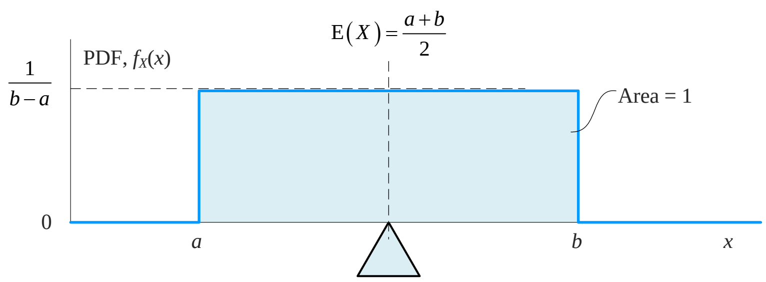

图 5. 随机变量 X 在 $[a, b]$ 上为均匀分布

X 的期望值为：

$$\mathrm{E}(X)=\int_{a}^{b} x \cdot \frac{1}{b-a} \mathrm{d} x=\frac{1}{b-a} \frac{x^2}{2}\bigg|_{a}^{b}=\frac{1}{b-a} \frac{b^2-a^2}{2}=\frac{a+b}{2} \tag{14}$$

随机变量 X 的取值在 $[a, b]$ 变化，对应的概率密度变化用 $f_X(x)$ 刻画。而求得的期望值 $\mathrm{E}(X)$ 则是一个标量，这相当于总结归纳。上式相当于找到一块均质木板的质心在长度方向上的位置。

方差

连续随机变量 X 方差的定义为：

$$\operatorname{var}(X)=\mathrm{E}\left[\left(X-\mathrm{E}(X)\right)^2\right]=\int_x\left(\underbrace{x-\mathrm{E}(X)}_{\text{Deviation}}\right)^2\cdot\underbrace{f_X(x)}_{\text{Weight}}\mathrm{d}x \tag{15}$$

同样，连续随机变量 X 的方差也满足如下计算技巧：

$$\operatorname{var}(X)=\mathrm{E}\left(\left(X-\mathrm{E}(X)\right)^2\right)=\mathrm{E}\left(X^2\right)-\left(\mathrm{E}(X)\right)^2 \tag{16}$$

其中，

$$\mathrm{E}\left(X^2\right)=\int_x x^2\cdot f_X(x)\mathrm{d}x \tag{17}$$

举个例子

对于图 5 所示均匀分布，为了方便计算 X 的方差，计算 X 平方的期望值为：

$$\mathrm{E}\left(X^2\right)=\int_a^b x^2\cdot\frac{1}{b-a}\mathrm{d}x=\frac{1}{b-a}\left.\frac{x^3}{3}\right|_a^b=\frac{1}{b-a}\frac{b^3-a^3}{3}=\frac{a^2+ab+b^2}{3} \tag{18}$$

根据 (16)，X 的方差为：

$$\begin{aligned}\operatorname{var}(X)&=\mathrm{E}\left(\left(X-\mathrm{E}(X)\right)^2\right)=\mathrm{E}\left(X^2\right)-\left(\mathrm{E}(X)\right)^2\\&=\frac{a^2+ab+b^2}{3}-\frac{(a+b)^2}{4}=\frac{(b-a)^2}{12}\end{aligned} \tag{19}$$

数值积分

如图 6 所示，随机变量 X 在 $[0, 1]$ 上为均匀分布。我们可以很容易通过积分得到期望值、方差。但是，并不是所有的概率密度函数都有解析式；此外，即便有解析式，也不代表我们能计算得到积分的解析解。

如图 7 所示，这就需要用到《数学要素》第 18 章介绍的**数值积分** (numerical integration)。当然，我们还可以用**蒙特卡洛模拟** (Monte Carlo simulation) 估算面积，这是本书后续要介绍的内容。

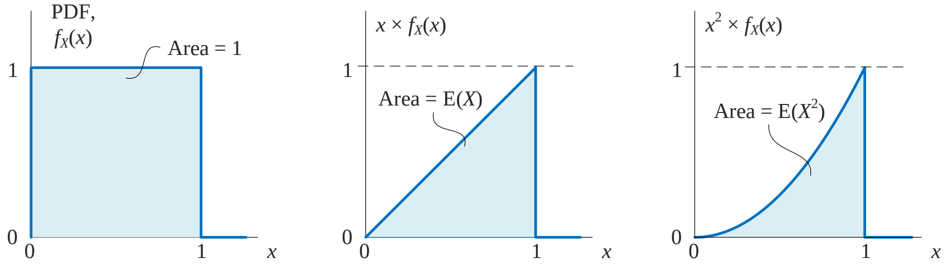

图 6. 随机变量 X 在 [0, 1] 上为均匀分布

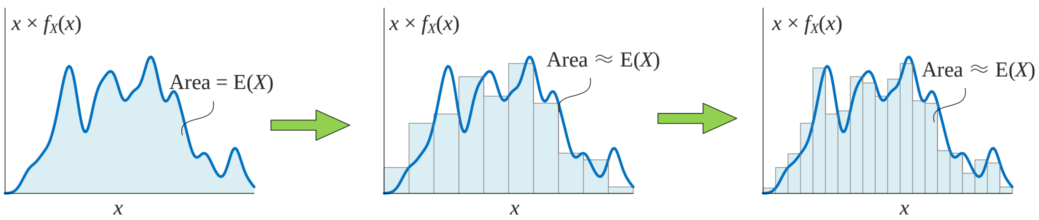

图 7. 数值积分估算期望值

6.3 二元连续随机变量

假设同一个试验中，有两个连续随机变量 X 和 Y，非负二元函数 $f_{X,Y}(x,y)$ 为 (X, Y) 的**联合概率密度函数** (joint probability density function 或 joint PDF)。

本章前文介绍，对于一元连续随机变量，积分得到的面积对应概率。而二元随机变量计算概率的工具是二重积分，从图像上来看，二重积分得到的体积对应概率。

如图 8 所示，给定积分区域 $A = \{(x, y) \mid a < x < b, c < y < d\}$，概率 $\Pr((X, Y) \in A)$ 对应的二重积分为：

$$\underbrace{\Pr\left(\left(X,Y\right)\in A\right)}_{\text{Probability}}=\int_c^d\int_a^b \underbrace{f_{X,Y}\left(x,y\right)}_{\text{Joint PDF}}\mathrm{d}\,x\,\mathrm{d}\,y \tag{20}$$

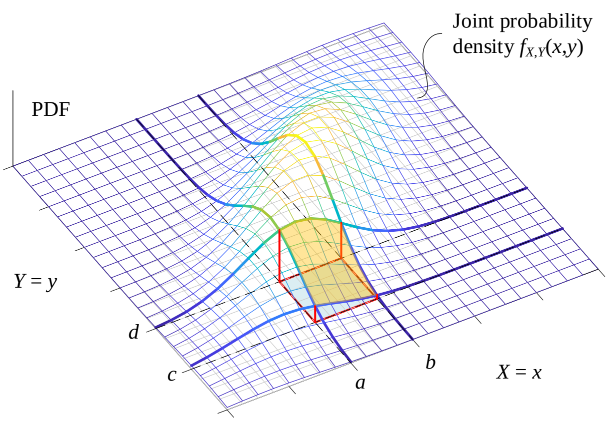

图 8. 二元 PDF $f_{X,Y}(x,y)$ 在 $A = \{(x, y) \mid a < x < b, c < y < d\}$ 二重积分

体积为 1：样本空间概率为 1

如果积分区域为整个平面，二重积分的结果为 1：

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\underbrace{f_{X,Y}(x,y)}_{\text{Joint PDF}}\mathrm{d}x\,\mathrm{d}y=1 \tag{21}$$

也就是说，图 8 中 $f_{X,Y}(x,y)$ 曲面和水平面围成几何形状的体积为 1，代表样本空间的概率为 1。这本质上也是“穷举法”。

累积概率密度 CDF

二元累积概率函数 CDF $F_{X,Y}(x,y)$ 定义为：

$$\underbrace{F_{X,Y}(x,y)}_{\text{Probability}}=\Pr(X<x,Y<y)=\int_{-\infty}^{y}\int_{-\infty}^{x}\underbrace{f_{X,Y}(s,t)}_{\text{Joint PDF}}\mathrm{d}s\,\mathrm{d}t \tag{22}$$

图 9 所示等高线为某个二元累积概率函数 $F_{X,Y}(x,y)$。图 9 还绘制了两条边缘 CDF 曲线。

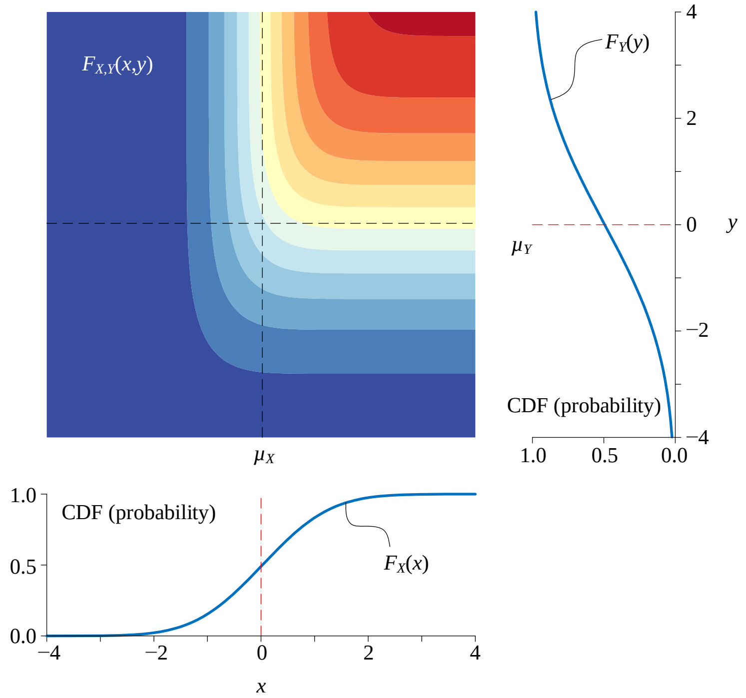

图 9. CDF 函数曲面 $F_{X,Y}(x,y)$ 平面填充等高线，边缘 CDF

6.4 边缘概率：二元 PDF 偏积分

图 10 所示为二元概率密度函数 $f_{X,Y}(x,y)$ 曲面和边缘概率曲线的关系。

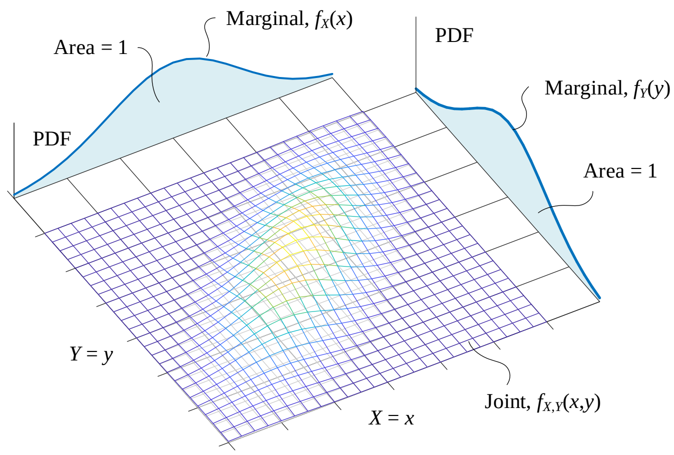

图 10. 二元联合概率密度函数曲面和边缘概率密度之间的关系

边缘概率密度函数 $f_X(x)$

如图 11 所示，连续随机变量 X 的边缘概率密度函数 $f_X(x)$ 可以通过 $f_{X,Y}(x,y)$ 对 y"偏积分"得到：

$$\underbrace{f_X(x)}_{\text{Marginal}} = \overbrace{\int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}y}^{\text{Eliminate } y} \tag{23}$$

上式，相当于消去了变量 y。

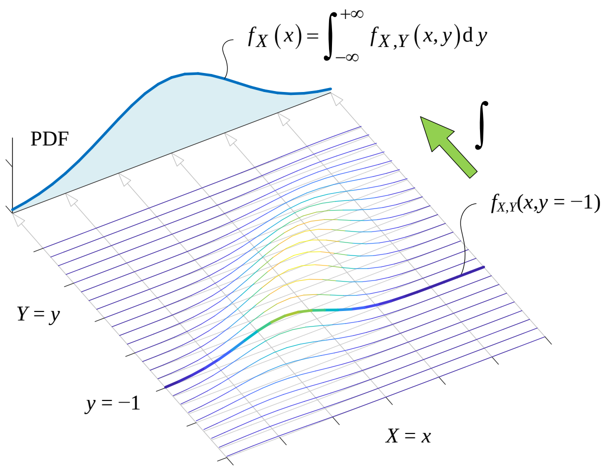

图 11. 联合概率密度 $f_{X,Y}(x,y)$ 对 y"偏积分"得到边缘概率密度 $f_X(x)$

(23) 可以简写为:

$$\underbrace{f_X(x)}_{\text{Marginal}} = \int_y \overbrace{\underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}y}^{\text{Eliminate } y} \tag{24}$$

注意，$f_X(x)$ 还是概率密度函数，而不是概率。也就是说，$f_{X,Y}(x,y)$ 二重积分得到概率，$f_{X,Y}(x,y)$ “偏积分” 得到的还是概率密度函数。

图 12 比较 $f_{X,Y}(x,y)$ 和 $f_X(x)$ 曲线。当 y 取不同值时，我们可以看到 $f_{X,Y}(x,y)$ 和 $f_X(x)$ 曲线形状不同。

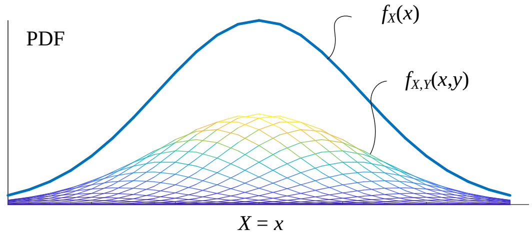

图 12. 比较联合概率密度 $f_{X,Y}(x,y)$ 和边缘概率密度 $f_X(x)$ 曲线

边缘概率密度函数 $f_Y(y)$

同理，如图 13 所示，连续随机变量 Y 的边缘分布概率密度函数 $f_Y(y)$ 可以通过 $f_{X,Y}(x,y)$ 对 x“偏积分”得到:

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \overbrace{\int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}x}^{\text{Eliminate } x} \tag{25}$$

上式相当消去了变量 x。上式也可以简写为:

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \int_x \overbrace{\underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}x}^{\text{Eliminate } x} \tag{26}$$

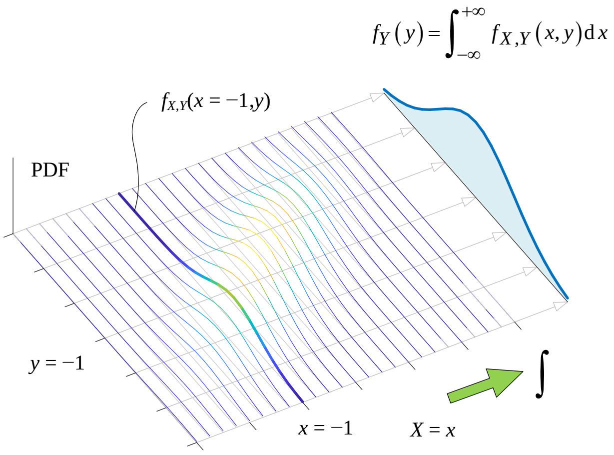

图 13. $f_{X,Y}(x,y)$ 对 x“偏积分”得到边缘分布概率密度函数 $f_Y(y)$

6.5 条件概率：引入贝叶斯定理

条件概率密度函数 $f_{X|Y}(x|y)$

设 X 和 Y 为连续随机变量，联合概率密度函数为 $f_{X,Y}(x,y)$。利用贝叶斯定理，在给定 $Y = y$ 条件下，且 $f_Y(y) > 0$，X 的条件概率密度函数 $f_{X|Y}(x|y)$ 为：

$$\underbrace{f_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{f_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{f_Y(y)}_{\text{Marginal}}} \tag{27}$$

再次强调，上式中，边缘 $f_Y(y)$ 也是概率密度。

图 14 中 $f_{X,Y}(x,y = -1)$ 曲线代表 $Y = -1$ 时联合概率密度函数。

$f_{X,Y}(x,y = -1)$ 对 x 在 $(-\infty, +\infty)$ 积分的结果为边缘概率概率密度 $f_Y(y = -1)$。也就是说，$f_{X,Y}(x,y = -1)$ 曲线面积为边缘概率密度 $f_Y(y = -1)$。

下一步，$f_{X,Y}(x,y = -1)$ 经过 $f_Y(y = -1)$ 缩放得到条件概率曲线 $f_{X|Y}(x|y = -1)$。

注意，$f_{X|Y}(x|y = -1)$ 和横轴围成图形的面积为 1，这代表 $Y = -1$ 这个新的样本空间概率为 1。

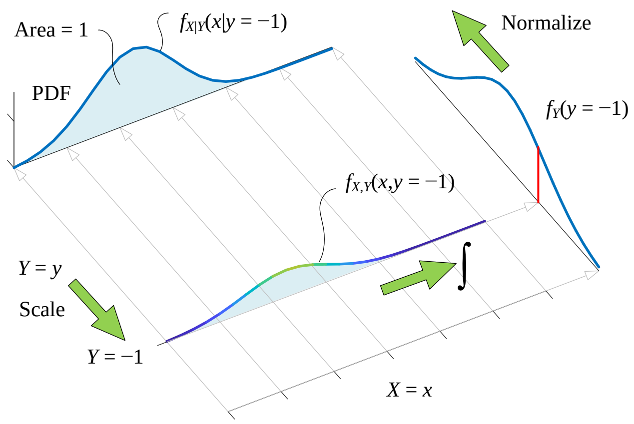

图 14. 给定 $Y=y$ 条件下且 $f_Y(y)>0$，X 的条件概率密度函数

图 15 比较 $f_X(x)$ 和 y 取不同值时条件概率密度函数 $f_{X|Y}(x|y)$ 图像。将这些曲线投影到同一个平面，得到图 16。注意，图 16 中所有曲线和横轴围成图形的面积都是 1。

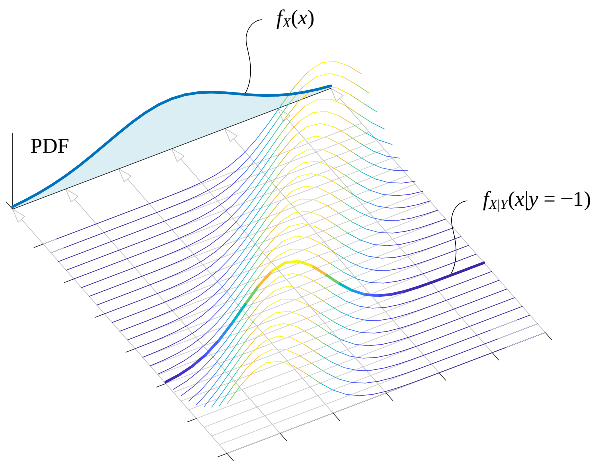

图 15. 比较边缘概率密度 $f_X(x)$ 和条件概率密度 $f_{X|Y}(x|y)$

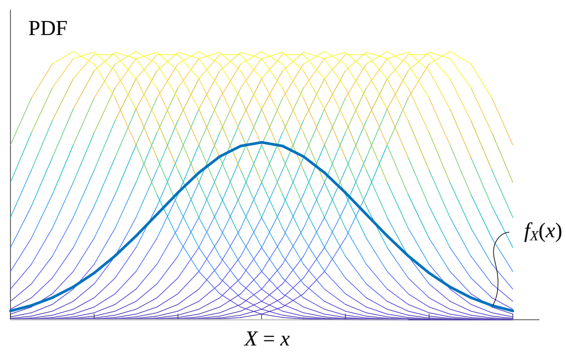

图 16. 比较边缘概率密度 $f_X(x)$ 和条件概率密度 $f_{X|Y}(x|y)$，投影在平面上

条件概率密度函数 $f_{Y|X}(y|x)$

给定 $X = x$ 条件下，且 $f_X(x) > 0$，条件概率密度函数 $f_{Y|X}(y|x)$ 可以通过下式求得：

$$\underbrace{f_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{f_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{f_X(x)}_{\text{Marginal}}} \tag{28}$$

如图 17 所示为，当 $X = -1$ 条件下，联合概率密度函数 $f_{X,Y}(x = -1,y)$ 首先对 y 在 $(-\infty, +\infty)$ 积分的结果为边缘概率密度值 $f_X(x = -1)$。下一步，$f_{X,Y}(x = -1,y)$ 经过 $f_X(x = -1)$ 缩放得到条件概率曲线 $f_{Y|X}(y|x = -1)$。

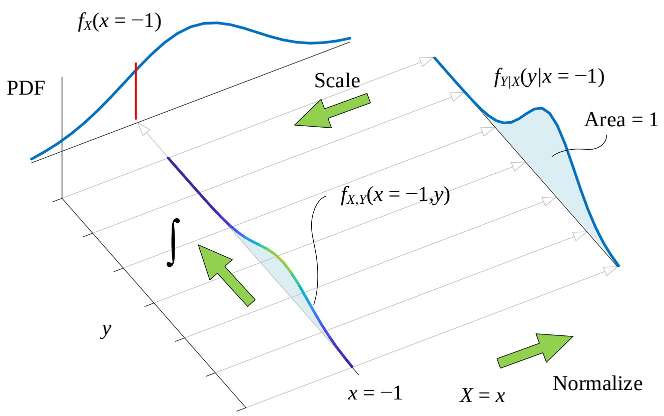

图 17. 给定 $X = x$ 条件下且 $f_X(x) > 0$，Y 的条件概率密度函数

图 18 比较 $f_Y(y)$ 和 x 取不同值时条件概率密度函数 $f_{Y|X}(y|x)$ 图像。

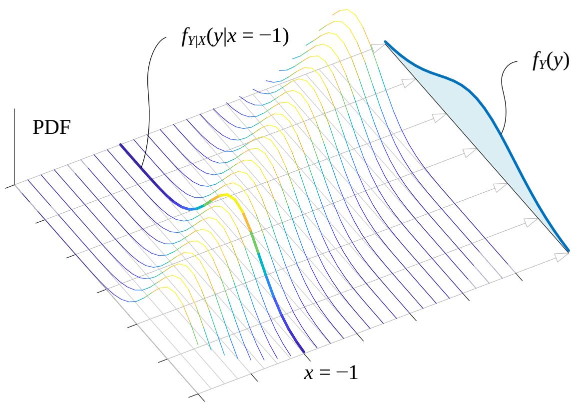

图 18. 比较边缘概率密度 $f_Y(y)$ 和条件概率密度 $f_{Y|X}(y|x)$ 图像

联合概率、边缘概率、条件概率

根据贝叶斯定理，联合概率、边缘概率、条件概率三者关系为：

$$\underbrace{f_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{f_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{f_Y(y)}_{\text{Marginal}} = \underbrace{f_{Y|X}(y|x)}_{\text{Conditional}} \underbrace{f_X(x)}_{\text{Marginal}} \tag{29}$$

在 (23) 基础上，连续随机变量 X 的边缘分布概率密度函数 $f_X(x)$ 可以通过下式获得：

$$\underbrace{f_X(x)}_{\text{Marginal}} = \int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}y = \int_{-\infty}^{+\infty} \underbrace{f_{X|Y}(x|t)}_{\text{Conditional}} \underbrace{f_Y(t)}_{\text{Marginal}} \mathrm{d}t \tag{30}$$

同理，连续随机变量 Y 的边缘分布概率密度函数 $f_Y(y)$ 可以通过下式计算得到：

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} \mathrm{d}x = \int_{-\infty}^{+\infty} \underbrace{f_{Y|X}(y|s)}_{\text{Conditional}} \underbrace{f_X(s)}_{\text{Marginal}} \mathrm{d}s \tag{31}$$

6.6 独立性：比较条件概率和边缘概率

如果连续随机变量 X 和 Y 独立，下式成立：

$$f_{X|Y}(x|y) = f_X(x) \tag{32}$$

图 19 所示为 X 和 Y 独立，条件概率密度函数 $f_{X|Y}(x|y)$ 和边缘概率密度函数 $f_X(x)$ 之间关系。我们发现条件概率 $f_{X|Y}(x|y)$ 的曲线和 Y 的取值无关。条件概率 $f_{X|Y}(x|y)$ 的曲线形状和边缘概率 $f_X(x)$ 完全一致。这和图 15 完全不同。

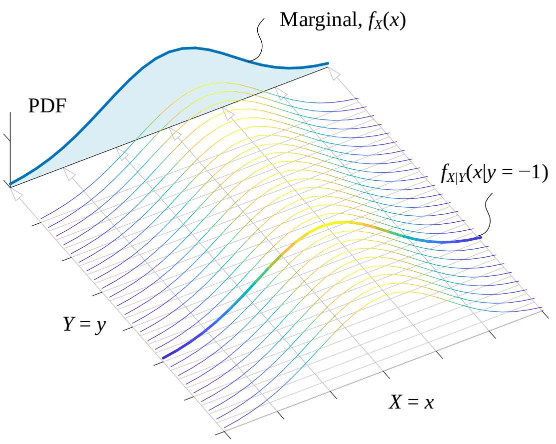

图 19. X 和 Y 独立，条件概率 $f_{X|Y}(x|y)$ 和边缘概率 $f_X(x)$ 之间关系

(32) 等价于:

$$f_{Y|X}\left(y|x\right)=f_Y\left(y\right) \tag{33}$$

图 20 所示为 X 和 Y 独立，条件概率 $f_{Y|X}(y|x)$ 和边缘概率 $f_Y(y)$ 的图像完全一致。

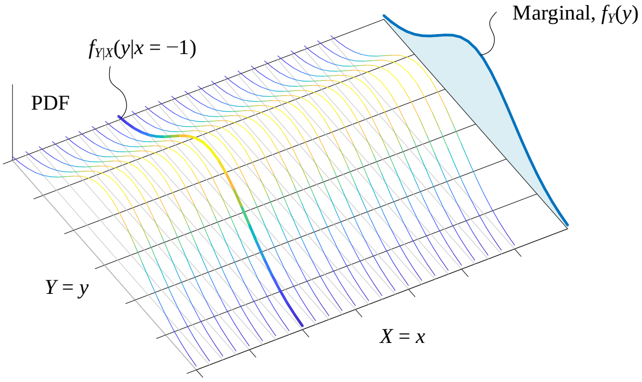

图 20. X 和 Y 独立，条件概率 $f_{Y|X}(y|x)$ 和边缘概率 $f_Y(y)$ 之间关系

独立：联合概率

对于两个连续随机变量 X 和 Y，如果两者独立，则联合概率密度函数 $f_{X,Y}(x,y)$ 为边缘概率密度函数 $f_X(x)$ 和 $f_Y(y)$ 的乘积:

$$f_{X,Y}\left(x,y\right)=f_X\left(x\right)f_Y\left(y\right) \tag{34}$$

图 21 所示为连续随机变量 X 和 Y 独立，联合概率 $f_{X,Y}(x,y)$ 曲面。图 22 所示为联合概率 $f_{X,Y}(x,y)$ 平面等高线。

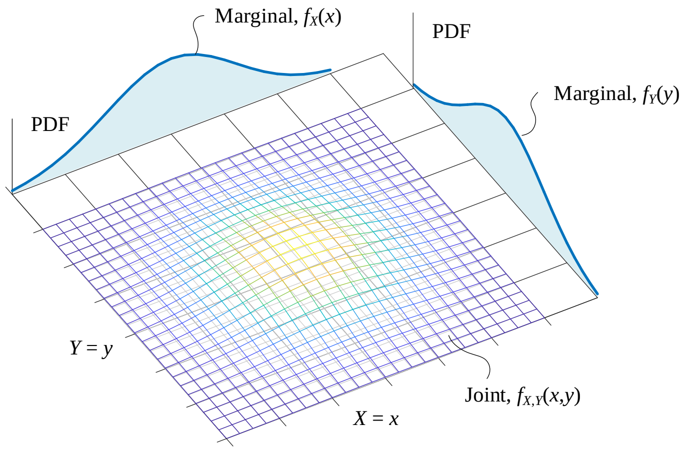

图 21. 连续随机变量 X 和 Y 独立，联合概率密度 $f_{X,Y}(x,y)$ 曲面

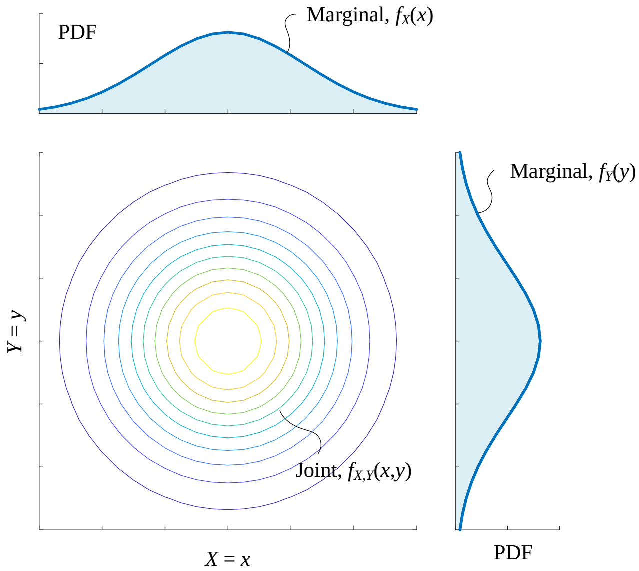

图 22. 连续随机变量 X 和 Y 独立，联合概率密度 $f_{X,Y}(x,y)$ 曲面等高线

6.7 以鸢尾花数据为例：不考虑分类标签

本章以下两节还是用鸢尾花数据集花萼长度 (X_1)、花萼宽度 (X_2)、分类标签 (Y) 为例，讲解本章前文介绍连续随机变量主要知识点。图 23 所示为不考虑分类时，鸢尾花样本数据花萼长度、花萼宽度散点图。

这两节采用和第 5 章 9、10 两节几乎一样的结构，方便大家比较阅读。

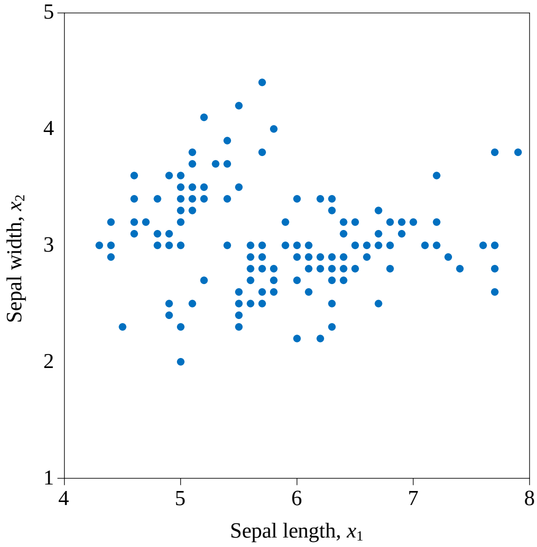

图 23. 鸢尾花数据花萼长度、花萼宽度散点图，不考虑分类

概率密度估计 → 联合概率密度函数 $f_{X1,X2}(x_1,x_2)$

基于高斯**核密度估计** (kernel density estimation, KDE)，我们可以得到如图 24 所示联合概率密度函数 $f_{X1,X2}(x_1,x_2)$。暖色系对应较大的概率密度值，也就是说鸢尾花样本分布更为密集。

再次强调，图 24 仅仅代表 $f_{X1,X2}(x_1,x_2)$ 的一种估计。即便采用相同的 KDE，使用不同的核函数、改变算法参数会导致 $f_{X1,X2}(x_1,x_2)$ 曲面形状变化。本书第 18 章将专门讲解核密度估计方法。

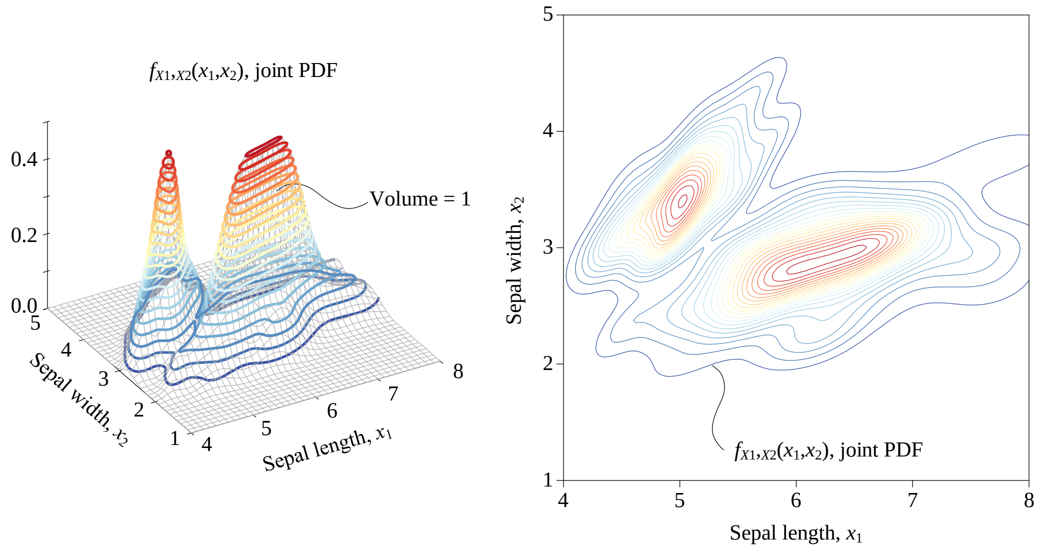

图 24. 联合概率密度函数 $f_{X1,X2}(x_1, x_2)$ 三维等高线和平面等高线，不考虑分类

举个例子，花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时，联合概率密度估计为：

$$\underbrace{f_{X1,X2}(x_1 = 6.5, x_2 = 2.0)}_{\text{Joint PDF}} \approx 0.02097 \tag{35}$$

注意，0.02097 这个数值是概率密度，不是概率。也就是说，我们不能说鸢尾花取到花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时对应的概率值为 0.02097，即便这个值某种程度上也代表可能性。

由于 $f_{X1,X2}(x_1,x_2)$ 有两个随机变量，对它二重积分可以得到概率值。二重积分就相当于“穷举法”。

采用“穷举法”，图 24 中 $f_{X1,X2}(x_1, x_2)$ 曲面和整个水平面围成的几何形体体积为 1，即：

$$\int_{x_2}\int_{x_1} f_{X1,X2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \underset{\text{Probability}}{1} \tag{36}$$

联合概率密度函数 $f_{X1,X2}(x_1,x_2)$ 的剖面线

$f_{X1,X2}(x_1, x_2)$ 本质上是个二元函数。《数学要素》第 10 章介绍过除了等高线，我们还可以使用“剖面线”分析二元函数。

如图 25 所示，当固定 x_1 取值时，$f_{X1,X2}(x_1 = c, x_2)$ 代表一条曲线。将一系列类似曲线投影到竖直平面得到图 25 (b)。图 25 (b)，这些直线和整个水平轴围成的面积就是边际概率 $f_{X1}(x_1 = c)$。而计算面积的数学工具就是“偏积分”。

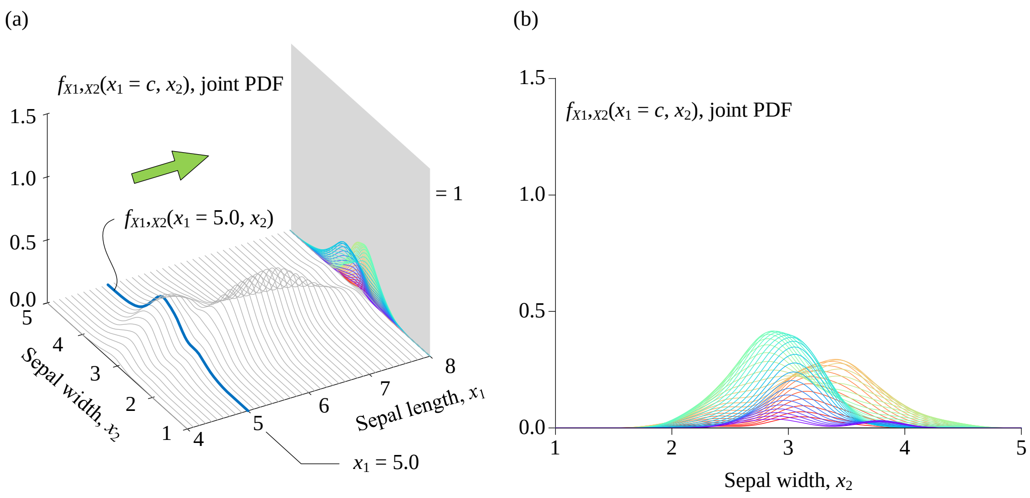

图 25. 固定 x_1 时，概率密度函数 $f_{X1,X2}(x_1,x_2)$ 随 x_2 变化

图 26 所示为固定 x_2 时，概率密度函数 $f_{X1,X2}(x_1,x_2)$ 随 x_1 变化。图 25 (b) 中直线和整个水平轴围成的面积对应边际概率 $f_{X2}(x_2 = c)$。

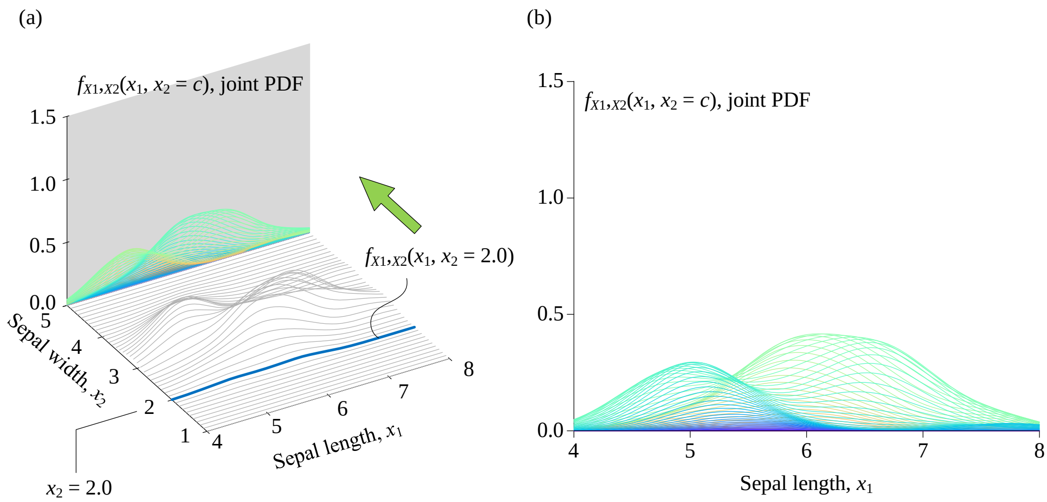

图 26. 固定 x_2 时，概率密度函数 $f_{X1,X2}(x_1,x_2)$ 随 x_1 变化

花萼长度边缘 PDF $f_{X1}(x_1)$：偏积分

图 27 所示为求解花萼长度边缘概率密度函数 $f_{X1}(x_1)$ 的过程：

$$\underbrace{f_{X1}(x_1)}_{\text{Marginal}} = \int_{x_2} \underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Joint}} \mathrm{d}\, x_2 \tag{37}$$

举个例子，当花萼长度 (X_1) 取值为 5.0 时，对应的边缘概率 $f_{X1}(5.0)$ 可以通过如下偏积分得到：

$$f_{X1}(x_1 = 5.0) = \int_{x_2} f_{X1,X2}(x_1 = 5.0, x_2) \mathrm{d}\, x_2 \tag{38}$$

图 27 中彩色阴影面积对应边缘概率，即 $f_{X1}(x_1)$ 曲线特定一点的高度。再次强调，$f_{X1}(x_1)$ 本身也是概率密度，不是概率值。$f_{X1}(x_1)$ 再积分可以得到概率。

如图 27 (b) 所示，$f_{X1}(x_1)$ 曲线和整个横轴围成图形的面积为 1。大家可以试着用数值积分计算期望值 $\mathrm{E}(X_1)$。

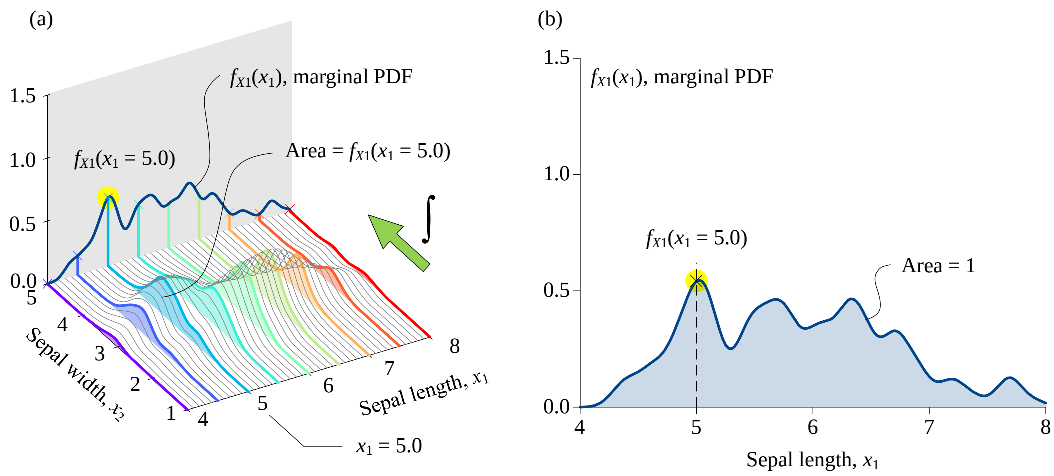

图 27. 偏积分求解边缘概率 $f_{X1}(x_1)$

花萼长度边缘 PDF $f_{X2}(x_2)$：偏求和

图 28 所示为求解花萼宽度边缘概率密度函数的过程：

$$\underbrace{f_{X2}(x_2)}_{\text{Marginal}} = \int_{x_1} \underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Joint}} \mathrm{d}\, x_1 \tag{39}$$

举个例子，当花萼宽度 (X_2) 取值为 2.0 时，对应的边缘概率密度 $f_{X2}(2.0)$ 可以通过如下偏积分得到：

$$f_{X2}\left(x_2=2.0\right)=\int_{x_1} f_{X1,X2}\left(x_1,x_2=2.0\right)\mathrm{d}\,x_1 \tag{40}$$

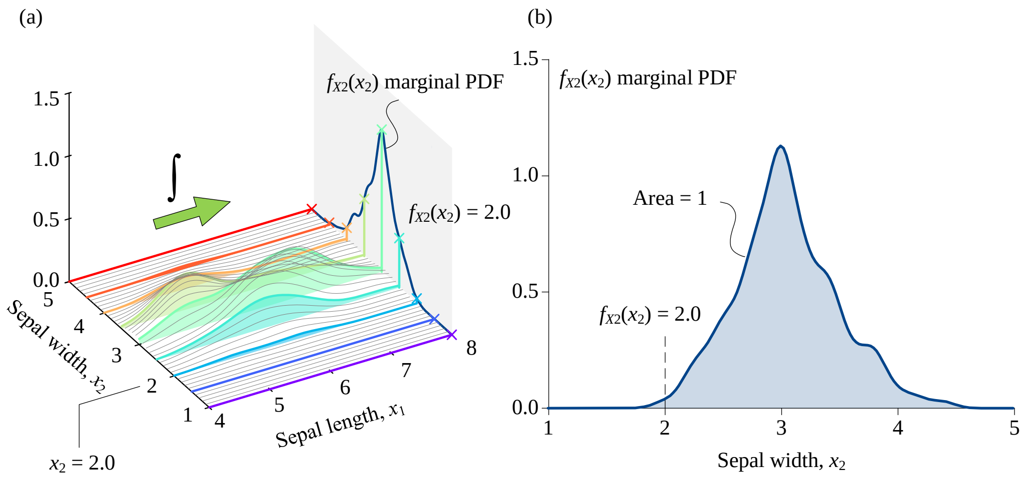

图 28. 偏积分求解边缘概率 $f_{X2}(x_2)$

联合 PDF vs 边缘 PDF

图 29 所示为联合 PDF 和边缘 PDF 之间关系。图中联合概率密度函数 $f_{X1,X2}(x_1,x_2)$ 采用高斯 KDE 估计得到。$f_{X1,X2}(x_1,x_2)$ 比较精准地捕捉到了鸢尾花样本数据的分布特征。

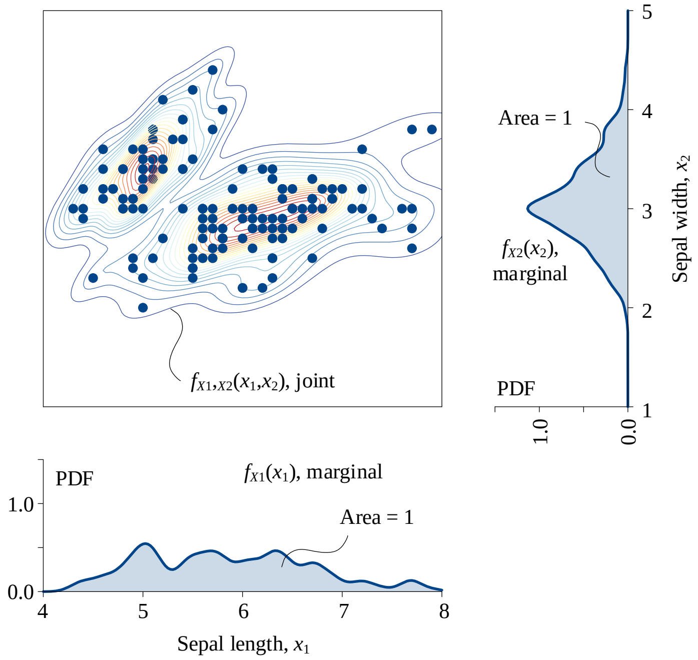

图 29. 联合 PDF 和边缘 PDF 之间关系

假设独立

如果假设 X_1 和 X_2 独立，联合概率密度 $f_{X1,X2}(x_1,x_2)$ 可通过下式计算得到：

$$
f_{X1,X2}(x_1,x_2) = f_{X1}(x_1) \cdot f_{X2}(x_2) \tag{41}
$$

图 30 所示为假设 X_1 和 X_2 独立时 $f_{X1,X2}(x_1,x_2)$ 的平面等高线和边缘 PDF 之间关系。

比较鸢尾花样本数据分布和假设 X_1 和 X_2 独立时估算得到的 $f_{X1,X2}(x_1,x_2)$ 等高线，很遗憾地发现这个联合概率密度函数 $f_{X1,X2}(x_1,x_2)$ 没有反映样本数据分布。

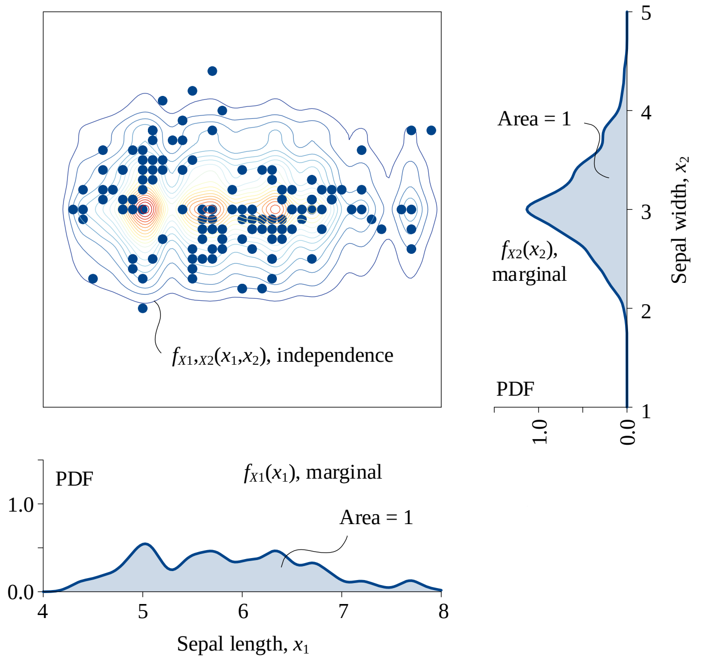

图 30. 联合概率，假设 X_1 和 X_2 独立

给定花萼长度，花萼宽度的条件 PDF $f_{X2 \mid X1}(x_2 \mid x_1)$

如图 31 所示，利用贝叶斯定理，条件概率密度 $f_{X2 \mid X1}(x_2 \mid x_1)$ 可以通过下式计算：

$$
\underbrace{f_{X2|X1}(x_2 | x_1)}_{\text{Conditional}} = \frac{\overbrace{f_{X1,X2}(x_1,x_2)}^{\text{Joint}}}{\underbrace{f_{X1}(x_1)}_{\text{Marginal}}} \tag{42}
$$

注意，上式中 $f_{X1}(x_1) > 0$。上式分母中的边缘概率 $f_{X1}(x_1)$ 起到归一化作用。如图 31 (b) 所示，经过归一化的条件概率曲线围成的面积变为 1，即：

$$
\int_{x_2} \underbrace{f_{X2|X1}(x_2 | x_1)}_{\text{Conditional}} \mathrm{d}x_2 = \int_{x_2} \frac{\overbrace{f_{X1,X2}(x_1,x_2)}^{\text{Joint}}}{\underbrace{f_{X1}(x_1)}_{\text{Marginal}}} \mathrm{d}x_2 = \frac{\int_{x_2} f_{X1,X2}(x_1,x_2)\,\mathrm{d}x_2}{f_{X1}(x_1)} = \frac{f_{X1}(x_1)}{f_{X1}(x_1)} = 1 \tag{43}
$$

将不同位置的条件 PDF $f_{X2 \mid X1}(x_2 \mid x_1)$ 曲线投影到平面得到图 32。图 32 (b) 中每条曲线和横轴围成面积都是 1。请大家仔细比较图 25 和图 32。此外，$f_{X2 \mid X1}(x_2 \mid x_1)$ 本身也是一个二元函数。图 33 所示为 $f_{X2 \mid X1}(x_2 \mid x_1)$ 三维等高线和平面等高线。

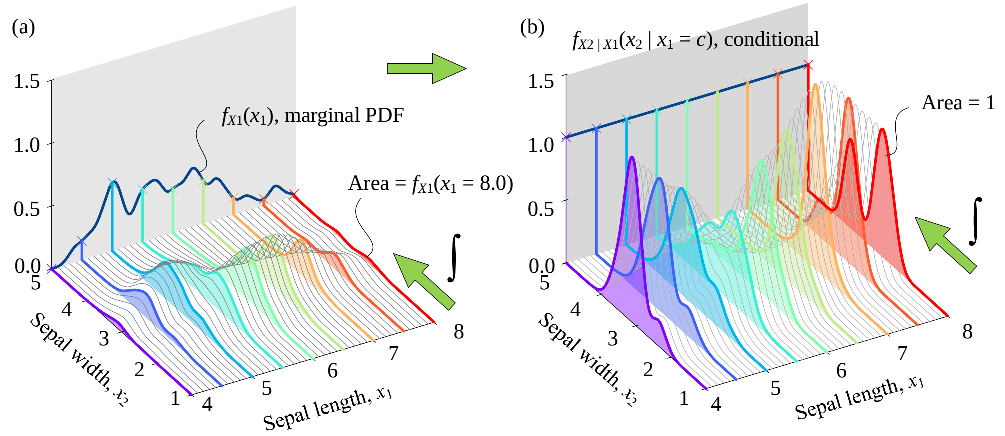

图 31. 计算条件概率 $f_{X2|X1}(x_2 | x_1)$ 原理

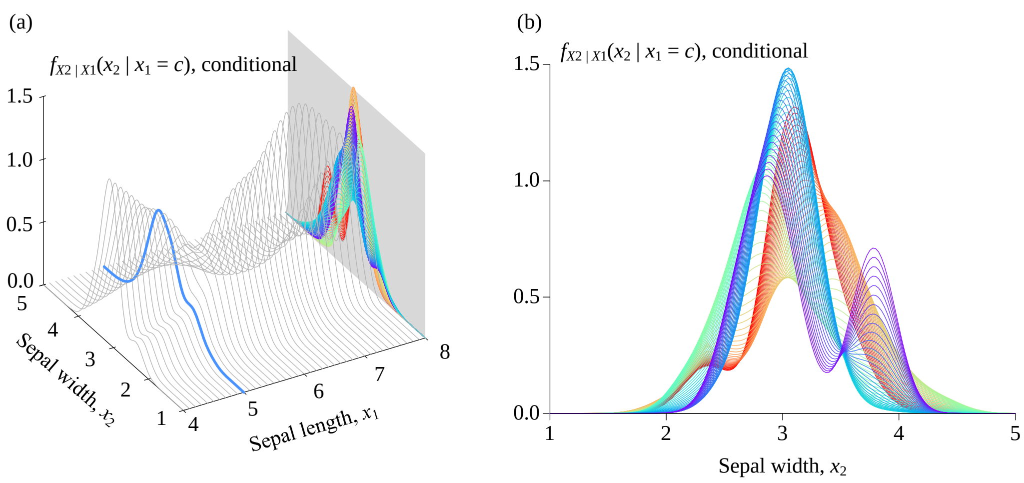

图 32. $f_{X2|X1}(x_2 | x_1)$ 曲线投影到平面

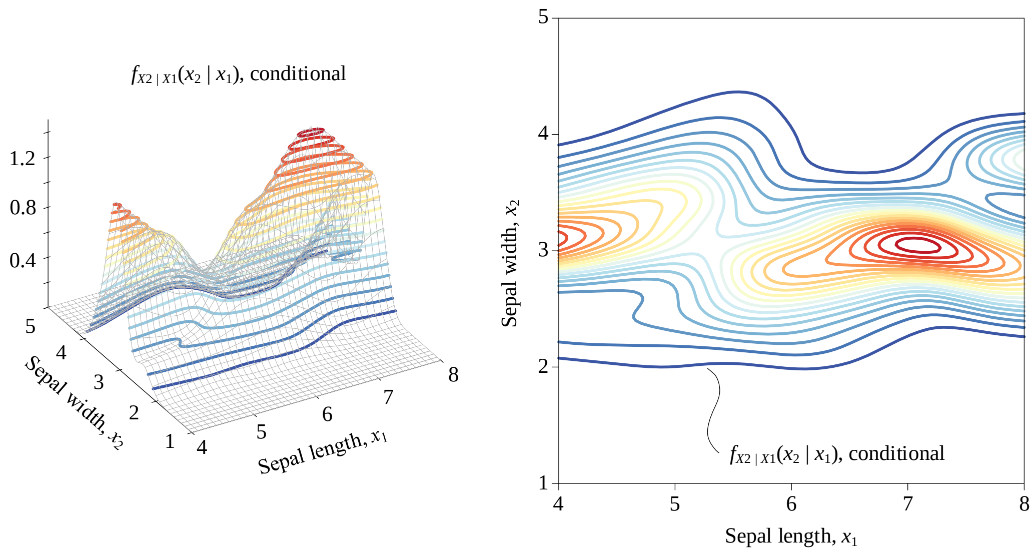

图 33. $f_{X2|X1}(x_2 | x_1)$ 条件概率密度三维等高线和平面等高线，不考虑分类

给定花萼宽度，花萼长度的条件概率密度函数 $f_{X1|X2}(x_1|x_2)$

如图 34 所示，同样利用贝叶斯定理，条件 PDF $f_{X1|X2}(x_1|x_2)$ 可以通过下式计算：

$$\underbrace{f_{X1|X2}(x_1|x_2)}_{\text{Conditional}} = \frac{\overbrace{f_{X1,X2}(x_1,x_2)}^{\text{Joint}}}{\underbrace{f_{X2}(x_2)}_{\text{Marginal}}} \tag{44}$$

注意，上式中 $f_{X2}(x_2) > 0$。类似前文，上式中分母中 $f_{X2}(x_2)$ 同样起到归一化作用。如图 34 (b) 所示，经过归一化 $f_{X1|X2}(x_1|x_2)$ 面积变为 1，即：

$$\int_{x_1} \underbrace{f_{X1|X2}(x_1|x_2)}_{\text{Conditional}} \mathrm{d}x_1 = \int_{x_1} \frac{\overbrace{f_{X1,X2}(x_1,x_2)}^{\text{Joint}}}{\underbrace{f_{X2}(x_2)}_{\text{Marginal}}} \mathrm{d}x_1 = \frac{\int_{x_1} f_{X1,X2}(x_1,x_2)\mathrm{d}x_1}{f_{X2}(x_2)} = \frac{f_{X2}(x_2)}{f_{X2}(x_2)} = 1 \tag{45}$$

将不同位置的条件概率密度 $f_{X1|X2}(x_1|x_2)$ 曲线投影到平面得到图 35。图 35 (b) 中每条曲线和横轴围成面积都是 1。也请大家仔细比较图 26 和图 35。

$f_{X1|X2}(x_1|x_2)$ 同样也是一个二元函数，如图 36 所示的 $f_{X1|X2}(x_1|x_2)$ 三维等高线和平面等高线。

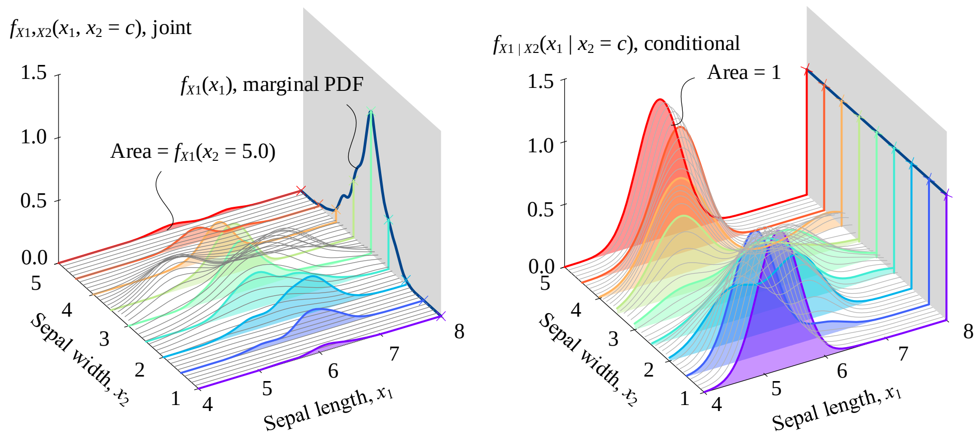

图 34. 计算条件概率 $f_{X1|X2}(x_1|x_2)$ 原理

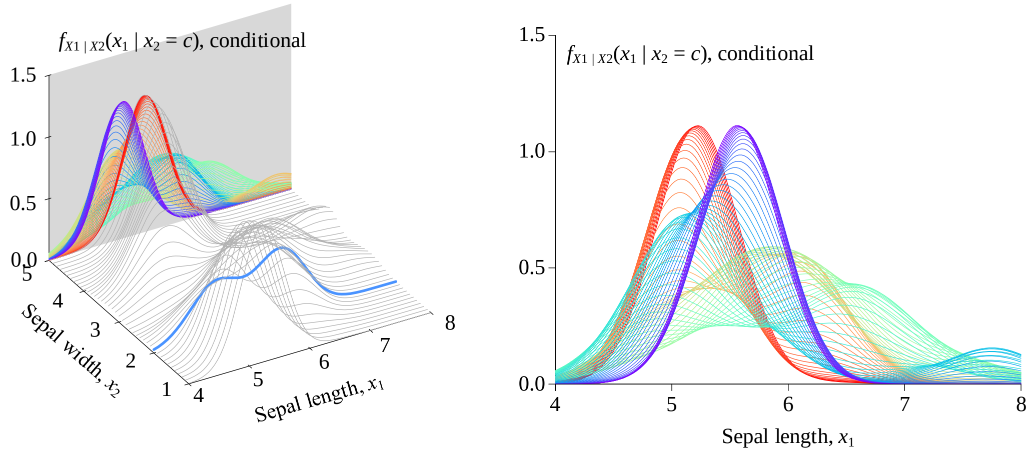

图 35. $f_{X1\,|\,X2}(x_1\,|\,x_2)$ 曲线投影到平面

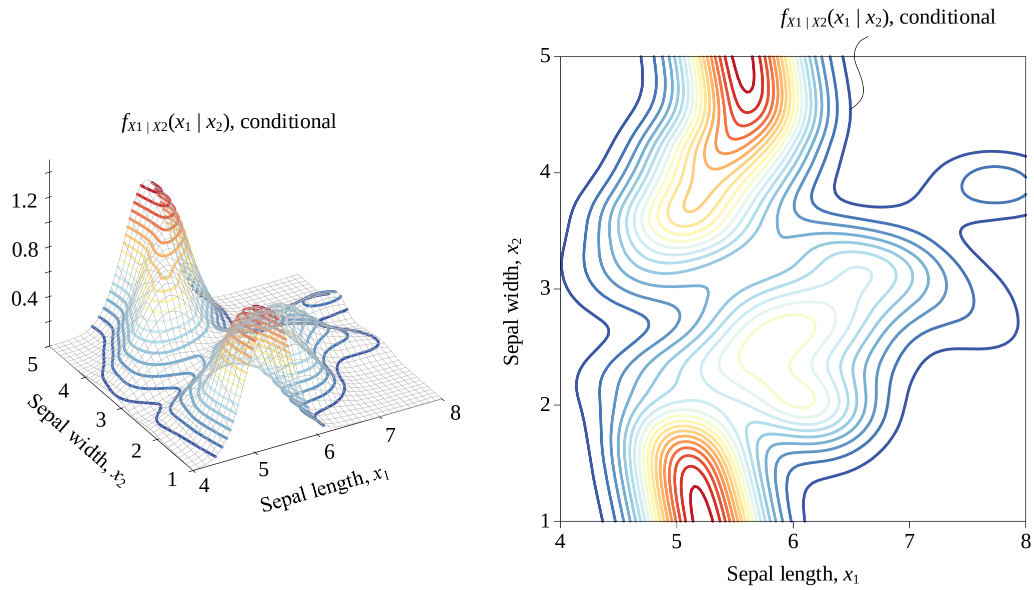

图 36. $f_{X1\,|\,X2}(x_1\,|\,x_2)$ 条件概率密度三维等高线和平面等高线，不考虑分类

6.8 以鸢尾花数据为例：考虑分类标签

本节将以鸢尾花标签为条件讨论条件概率。图 37 所示为考虑分类标签的鸢尾花数据散点图。

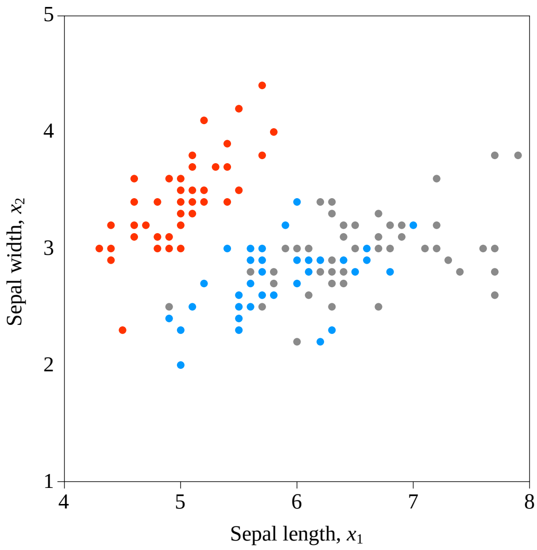

图 37. 鸢尾花数据花萼长度、花萼宽度散点图，考虑分类

给定分类标签 $Y = C_1$ (setosa)

图 38 所示为给定分类标签 $Y = C_1$ (setosa) 条件下，条件概率 $f_{X1,X2\mid Y}(x_1, x_2 \mid y = C_1)$ 平面等高线和条件边缘概率密度曲线。

$f_{X1,X2\mid Y}(x_1, x_2 \mid y = C_1)$ 曲面和整个水平面围成体积为 1，也就是说：

$$\int_{x_2}\int_{x_1} \underbrace{f_{X1,X2\mid Y}\left(x_1, x_2 \mid C_1\right)}_{\text{Conditional PDF}} \mathrm{d}\, x_1 \,\mathrm{d}\, x_2 = \underbrace{1}_{\text{Probability}} \tag{46}$$

用 KDE 估算 $f_{X1,X2\mid Y}(x_1, x_2 \mid y = C_1)$ 时，我们仅仅考虑标签为 C_1 的数据。同理，估算条件边缘概率曲线 $f_{X1\mid Y}(x_1 \mid y = C_1)$、$f_{X2\mid Y}(x_2 \mid y = C_1)$ 时，我们也不考虑其他标签数据。

图 38 中，$f_{X1\mid Y}(x_1 \mid y = C_1)$、$f_{X2\mid Y}(x_2 \mid y = C_1)$ 分别和 x_1、x_2 围成的面积也是 1，即：

$$\begin{aligned} \int_{x_1} \underbrace{f_{X1\mid Y}\left(x_1 \mid C_1\right)}_{\text{Conditional PDF}} \mathrm{d}\, x_1 &= \underbrace{1}_{\text{Probability}} \\ \int_{x_2} \underbrace{f_{X2\mid Y}\left(x_2 \mid C_1\right)}_{\text{Conditional PDF}} \mathrm{d}\, x_2 &= \underbrace{1}_{\text{Probability}} \end{aligned} \tag{47}$$

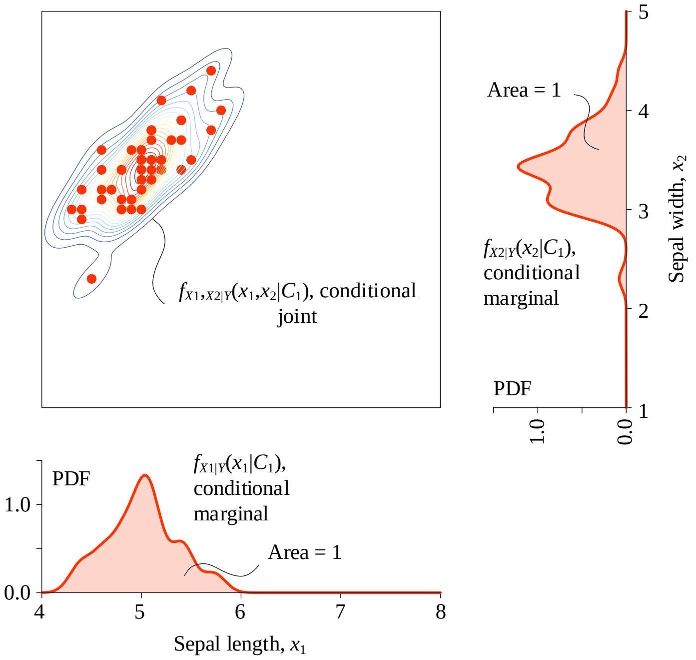

图 38. 条件概率 $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_1)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_1$ (setosa)

给定分类标签 $Y = C_2$ (versicolor)

图 39 所示为，给定分类标签 $Y = C_2$ (versicolor)，条件概率 $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_2)$ 平面等高线和条件边缘概率密度曲线。请大家自行分析这幅图。

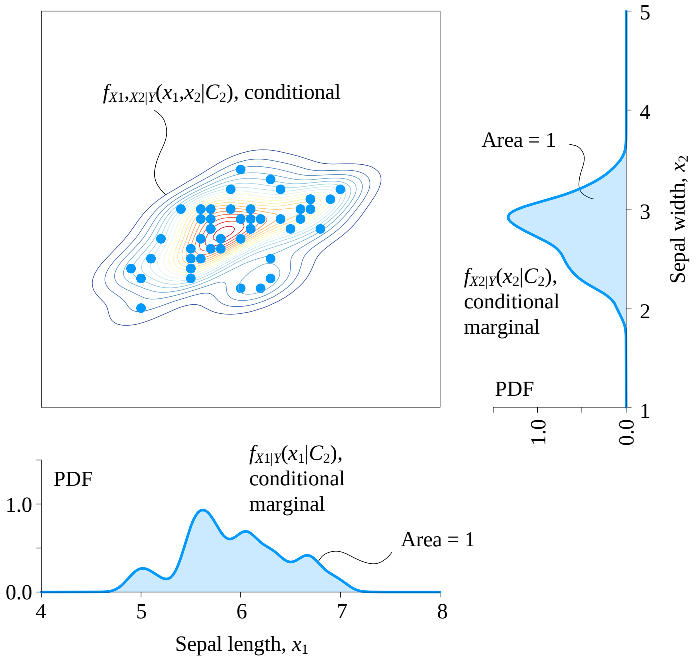

图 39. 条件 PDF $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_2)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_2$ (versicolor)

给定分类标签 $Y = C_3$ (virginica)

图 40 所示为，给定分类标签 $Y = C_3$ (virginica)，条件概率 $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_3)$ 平面等高线和条件边缘概率密度曲线。也请大家自行分析这幅图。

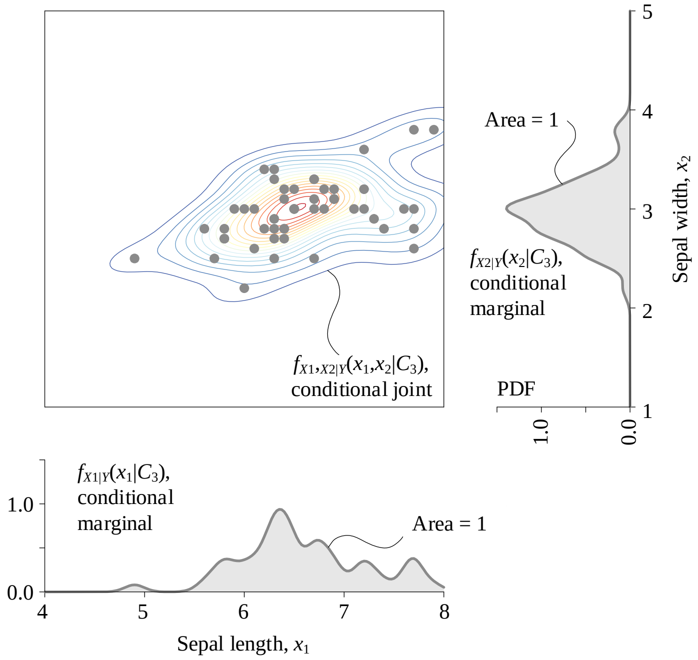

图 40. 条件 PDF $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,y = C_3)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_3$ (virginica)

全概率定理：穷举法

如图 41 所示，利用全概率定理，三幅条件概率等高线叠加可以得到联合概率密度，即：

$$
\begin{aligned}
f_{X1,X2}(x_1,x_2) = &f_{X1,X2|Y}(x_1,x_2|y=C_1)\,p_Y(C_1)+\\
&f_{X1,X2|Y}(x_1,x_2|y=C_2)\,p_Y(C_2)+\\
&f_{X1,X2|Y}(x_1,x_2|y=C_3)\,p_Y(C_3)
\end{aligned}
\tag{48}
$$

此外，请大家思考 $f_{X1}(x_1)$、$f_{X1\,|\,Y}(x_1\,|\,y = C_1)$、$f_{X1\,|\,Y}(x_1\,|\,y = C_2)$、$f_{X1\,|\,Y}(x_1\,|\,y = C_3)$ 四者关系。

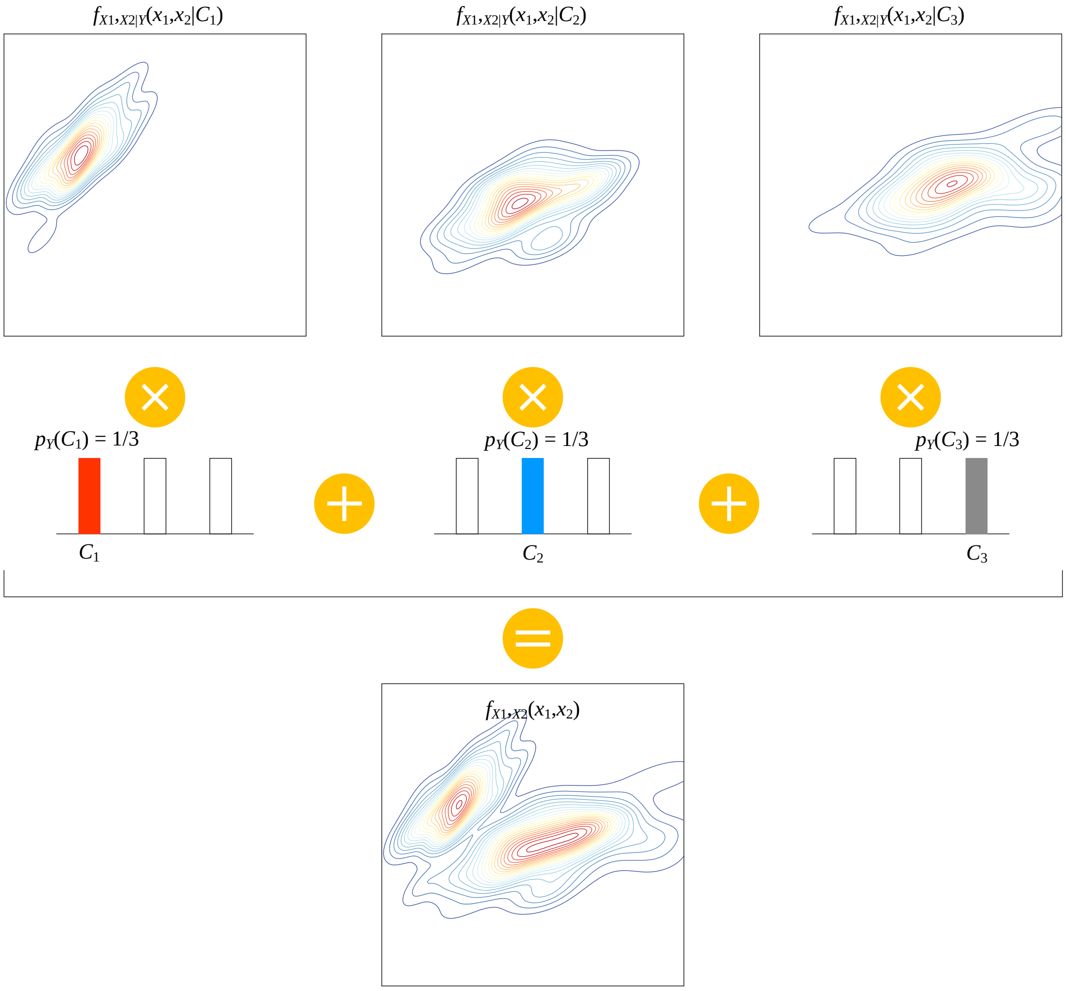

图 41. 利用全概率定理，计算 $f_{X1,X2}(x_1, x_2)$

给定 X_1 和 X_2，Y 的条件概率：后验概率

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

$$\overbrace{f_{Y|X1,X2}\left(C_k\,|x_1,x_2\right)}^{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1,x_2,C_k\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1,x_2\right)}_{\text{Evidence}}} \tag{49}$$

从分类角度来看，这相当于已知某个样本鸢尾花花萼长度和花萼宽度，该样本对应不同分类的概率。请大家修改代码自行绘制不同的后验概率 PDF 曲面。

本书后续将从这个角度探讨若何判定鸢尾花分类。

假设条件独立

如图 42 所示，如果假设条件独立，$f_{X1,X2|Y}(x_1, x_2 | y = C_1)$ 可以通过下式计算得到：

$$\underbrace{f_{X1,X2|Y}\left(x_1,x_2 \mid y=C_1\right)}_{\text{Conditional joint}} = \underbrace{f_{X1|Y}\left(x_1 \mid y=C_1\right)}_{\text{Conditional marginal}} \cdot \underbrace{f_{X2|Y}\left(x_2 \mid y=C_1\right)}_{\text{Conditional marginal}} \tag{50}$$

同理我们可以计算得到 $f_{X1,X2|Y}(x_1, x_2 | y = C_2)$、$f_{X1,X2|Y}(x_1, x_2 | y = C_3)$，具体如图 43、图 44 所示。

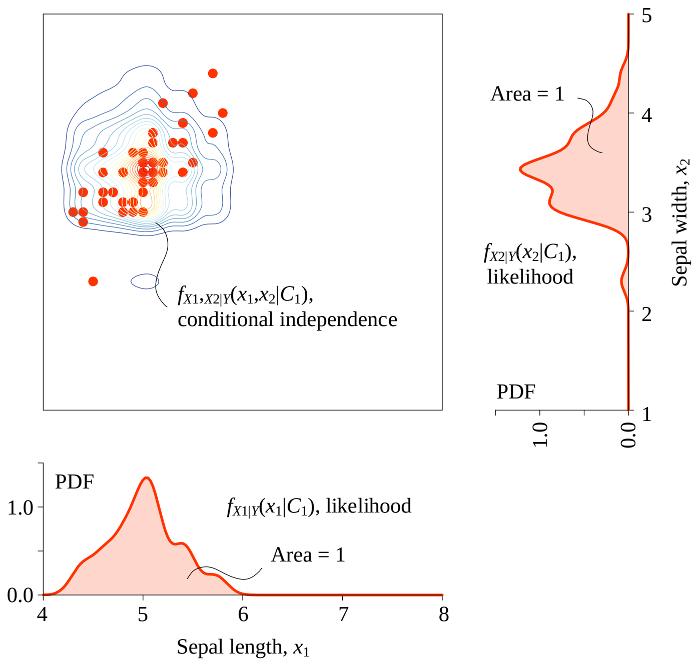

图 42. 给定 $Y = C_1$，X_1 和 X_2 条件独立，估算条件概率 $f_{X1,X2|Y}(x_1, x_2 | y = C_1)$

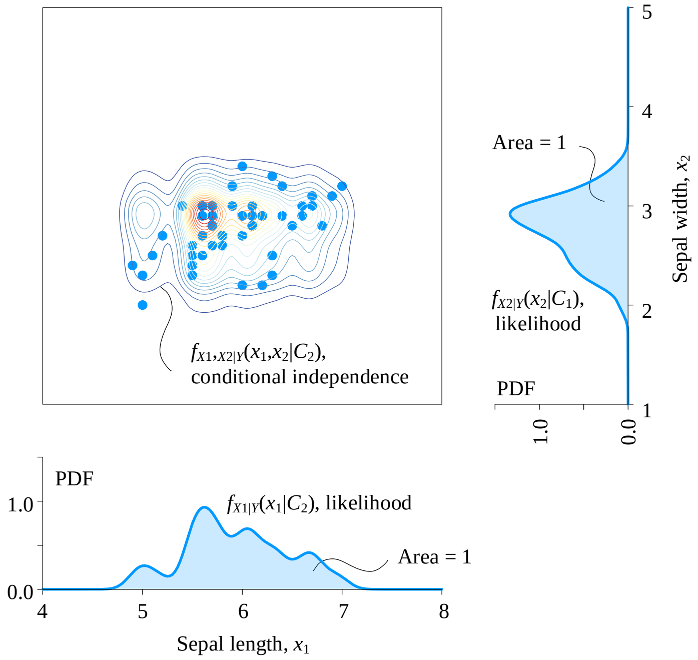

图 43. 给定 $Y = C_2$，X_1 和 X_2 条件独立，估算条件概率 $f_{X1,X2\mid Y}(x_1, x_2 \mid y = C_2)$

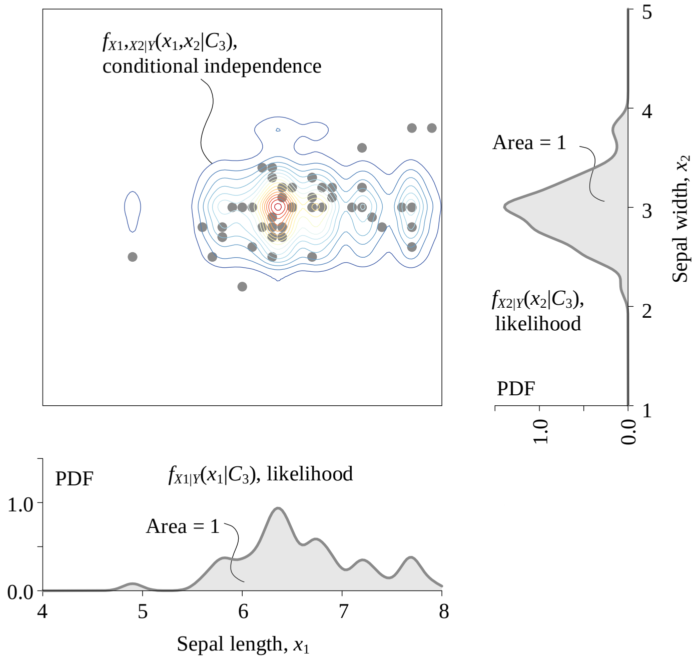

图 44. 给定 $Y = C_3$，X_1 和 X_2 条件独立，估算条件概率 $f_{X1,X2\mid Y}(x_1, x_2 \mid y = C_3)$

如图 45 所示，并利用全概率定理，我们也可以估算 $f_{X1,X2}(x_1, x_2)$：

$$
\begin{aligned}
f_{X1,X2}\left(x_1,x_2\right) = & f_{X1,X2|Y}\left(x_1,x_2\left|y=C_1\right.\right)p_Y\left(C_1\right)+ \\
& f_{X1,X2|Y}\left(x_1,x_2\left|y=C_2\right.\right)p_Y\left(C_2\right)+ \\
& f_{X1,X2|Y}\left(x_1,x_2\left|y=C_3\right.\right)p_Y\left(C_3\right) \\
= & f_{X1|Y}\left(x_1\left|y=C_1\right.\right)f_{X2|Y}\left(x_2\left|y=C_1\right.\right)p_Y\left(C_1\right)+ \\
& f_{X1|Y}\left(x_1\left|y=C_2\right.\right)f_{X2|Y}\left(x_2\left|y=C_2\right.\right)p_Y\left(C_2\right)+ \\
& f_{X1|Y}\left(x_1\left|y=C_3\right.\right)f_{X2|Y}\left(x_2\left|y=C_3\right.\right)p_Y\left(C_3\right)+
\end{aligned}
\tag{51}
$$

这是**朴素贝叶斯分类器** (Naive Bayes classifier) 的重要技术细节之一。本系列丛书《机器学习》一册将讲解朴素贝叶斯分类器。

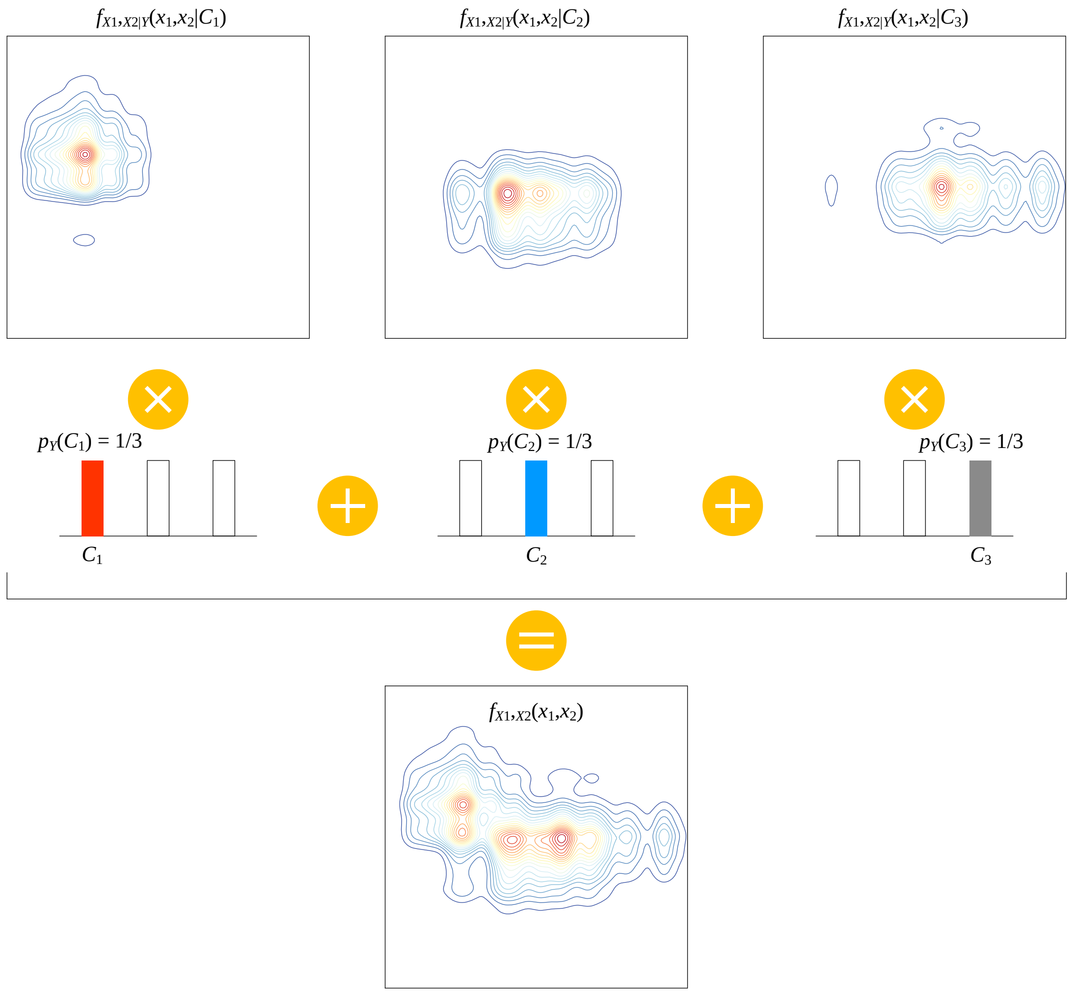

图 45. 利用全概率定理，估算 $f_{X1,X2}(x_1, x_2)$，假设条件独立

Bk5_Ch07_01.py 绘制本章大部分图像。

表 1. 比较离散和连续随机变量

	离散随机变量	连续随机变量
随机变量	取值可以一一列举出来，有限个或可数无穷个，比如 {0, 1}，{非负整数}	取值不可以一一列举出来，比如闭区间 [0, 1] 或 {非负实数}
一元随机变量概率质量/密度函数	概率质量函数 PMF，$p_X(x)$ PMF 本身就是概率值 $0 \le p_X(x) \le 1$ 计算工具：Σ	概率密度函数 PDF，$f_X(x)$ PDF 本身为概率密度 $0 \le f_X(x)$ 注意 $f_X(x)$ 可以大于 1 计算工具：∫
归一化	$\sum_x p_X(x) = 1$	$\int_x f_X(x)\,\mathrm{d}x = 1$
概率质量/密度函数图像	火柴梗图	曲线
计算概率 CDF	求和 $F_X(x) = \Pr(X \le x) = \sum_{t \le x} p_X(t)$	积分 $F_X(x) = \Pr(X \le x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$
期望	$\mathrm{E}(X) = \sum_x x \cdot p_X(x)$	$\mathrm{E}(X) = \int_x x \cdot f_X(x)\,\mathrm{d}x$
方差	$\mathrm{var}(X) = \sum_x (x - \mathrm{E}(X))^2 p_X(x)$	$\mathrm{var}(X) = \int_x (x - \mathrm{E}(X))^2 \cdot f_X(x)\,\mathrm{d}x$
常见分布	离散均匀分布，伯努利分布，二项分布，多项分布，泊松分布，几何分布，超几何分布	连续均匀分布，高斯分布，逻辑分布，学生 *t*-分布，对数正态分布，指数分布，卡方分布，Beta 分布
二元随机变量联合概率	概率质量函数 PMF，$p_{X,Y}(x,y)$	概率密度函数 PDF，$f_{X,Y}(x,y)$
边缘概率 求和法则	$p_{X,Y}(x,y)$ 偏求和结果为边缘 PMF $p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$	$f_{X,Y}(x,y)$ 偏积分结果为边缘 PDF $f_X(x) = \int_y f_{X,Y}(x,y)\,\mathrm{d}y$ $f_Y(y) = \int_x f_{X,Y}(x,y)\,\mathrm{d}x$
条件概率 $p_Y(y) > 0,\ p_X(x) > 0$ $f_Y(y) > 0,\ f_X(x) > 0$	$p_{X\|Y}(x\|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$ $p_{Y\|X}(y\|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)}$	$f_{Y\|X}(y\|x) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$ $f_{X\|Y}(x\|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$
条件概率归一化	$\sum_x p_{X\|Y}(x\|y) = 1$ $\sum_y p_{Y\|X}(y\|x) = 1$	$\int_x f_{X\|Y}(x\|y)\,\mathrm{d}x = 1$ $\int_y f_{Y\|X}(y\|x)\,\mathrm{d}x = 1$

随机变量独立	$p_{X\|Y}(x\|y)=p_X(x)$ $p_{Y\|X}(y\|x)=p_Y(y)$	$f_{X\|Y}(x\|y)=f_X(x)$ $f_{Y\|X}(y\|x)=f_Y(y)$
随机变量独立条件下，联合概率	$p_{X,Y}(x,y)=p_X(x)\,p_Y(y)$	$f_{X,Y}(x,y)=f_X(x)\,f_Y(y)$
随机变量条件独立，条件联合概率	$p_{X_1,X_2\|Y}(x_1,x_2\|y)=p_{X_1\|Y}(x_1\|y)\cdot p_{X2\|Y}(x_2\|y)$	$f_{X_1,X_2\|Y}(x_1,x_2\|y)=f_{X_1\|Y}(x_1\|y)\cdot f_{X2\|Y}(x_2\|y)$