

贝叶斯统计基础

沈明宏

2024 年 7 月 30 日

目录

1 概率论基础	2
2 概率分布	4
3 贝叶斯定理	5
3.1 案例一：核酸检测	6

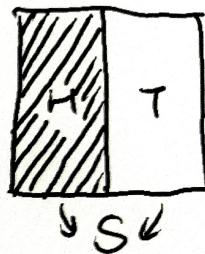
1 概率论基础

假设一共投了 S 次硬币，其中硬币朝上的次数为 H ，则硬币朝上的概率 $P(H)$ 为

$$P(H) = \frac{H}{S} \quad (1)$$

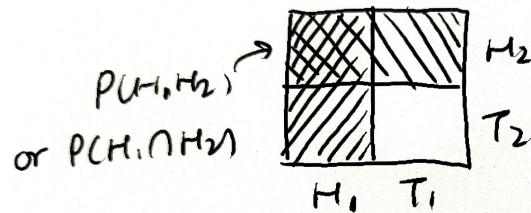
显然有：

$$0 \leq P(H) \leq 1 \quad (2)$$



假设我们让两个人分别投硬币，第一个人投硬币的次数为 S_1 ，第二个人投硬币的次数为 S_2 ，则两个人投硬币朝上的次数分别为 H_1 和 H_2 ，则两个人投硬币朝上的概率 $P(H_1)$ 和 $P(H_2)$ 分别为：

$$\begin{aligned} P(H_1) &= \frac{H_1}{S_1} \\ P(H_2) &= \frac{H_2}{S_2} \end{aligned} \quad (3)$$



我们可以计算两个人各投一枚硬币，都朝上的概率 $P(H_1 H_2)$ 为：

$$\begin{aligned} P(H_1 H_2) &= P(H_1 \cap H_2) \\ &= \frac{H_1 H_2}{S_1 S_2} \end{aligned} \quad (4)$$

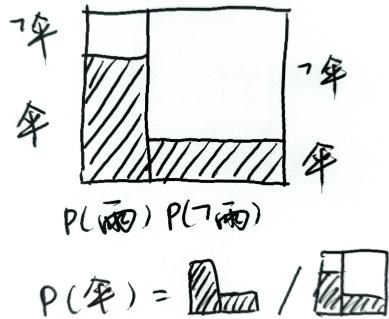
同理，我们可以计算下雨的概率 $P(\text{雨})$ 、带伞的概率 $P(\text{伞})$ 等。

$$\begin{aligned} P(\text{雨}) &= \frac{\text{下雨天数}}{\text{总天数}} \\ P(\text{伞}) &= \frac{\text{带伞人数}}{\text{总人数}} \end{aligned} \quad (5)$$

显然，有：

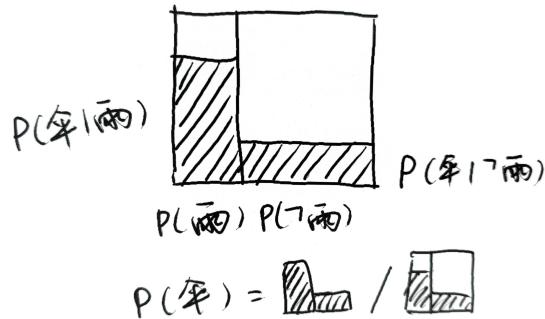
$$\begin{aligned} P(\text{雨}) + P(\text{不下雨}) &= 1 \\ P(\text{伞}) + P(\text{不带伞}) &= 1 \end{aligned} \quad (6)$$

注意，这里和上面 $P(H)$ 不同，这里的 $P(\text{雨})$ 和 $P(\text{伞})$ 是相关的。



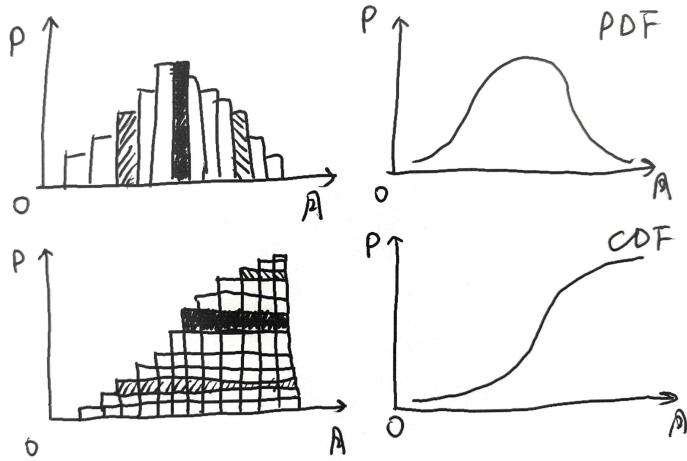
这时，为了计算两根阴影柱子的高度，我们需要引入“条件概率”。比方说，我们想知道下雨的情况下带伞的概率 $P(\text{伞}|\text{雨})$ ，则有：

$$\begin{aligned} P(\text{伞}|\text{雨}) &= \frac{P(\text{雨} \cap \text{伞})}{P(\text{雨})} \\ P(\text{雨}|\text{伞}) &= \frac{P(\text{雨} \cap \text{伞})}{P(\text{伞})} \end{aligned} \quad (7)$$



2 概率分布

刚才的事件都是离散的，但是在实际中，我们通常会遇到连续的情况。这时，我们就需要引入概率密度函数（PDF）和概率分布函数（CDF）。



概率密度函数 $f(x)$ 是一个函数，满足：

$$\begin{aligned} f(x) &\geq 0 \\ \int_{-\infty}^{+\infty} f(x) dx &= 1 \end{aligned} \tag{8}$$

概率分布函数 $F(x)$ 是一个函数，满足：

$$F(x) = \int_{-\infty}^x f(x) dx \tag{9}$$

概率分布函数是概率密度函数的积分，所以有：

$$F(b) - F(a) = P(a \leq x \leq b) = \int_a^b f(x) dx \tag{10}$$

反之，如果我们知道概率分布函数 $F(x)$ ，则可以通过求导得到概率密度函数 $f(x)$ ：

$$f(x) = \frac{dF(x)}{dx} \tag{11}$$

3 贝叶斯定理

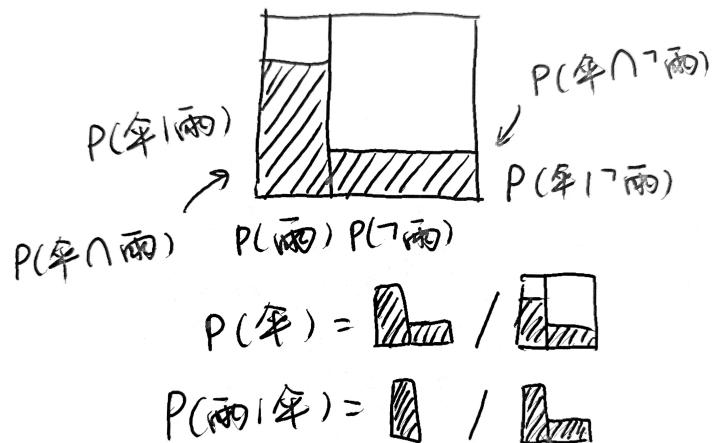
$P(\text{雨}|\text{伞})$ 是个非常奇怪的概率，什么叫“带伞的时候下雨的概率”？

但其实，这才是我们科研中在做的。我们通常是只能收集到“带伞”的数据，然后根据这个数据来推断下雨的概率。

这时，我们就需要用到贝叶斯定理。

根据上面的方程，我们可以得到：

$$\begin{aligned} P(\text{雨} \cap \text{伞}) &= P(\text{雨}|\text{伞}) \times P(\text{伞}) \\ &= P(\text{伞}|\text{雨}) \times P(\text{雨}) \end{aligned} \quad (12)$$



也就是说：

$$\begin{aligned} P(\text{雨}|\text{伞}) \times P(\text{伞}) &= P(\text{伞}|\text{雨}) \times P(\text{雨}) \\ P(\text{雨}|\text{伞}) &= \frac{P(\text{伞}|\text{雨}) \times P(\text{雨})}{P(\text{伞})} \end{aligned} \quad (13)$$

这可以抽象为：

$$P(\text{参数}|\text{数据}) = \frac{P(\text{数据}|\text{参数}) \times P(\text{参数})}{P(\text{数据})} \quad (14)$$

这是因为 $P(\text{数据})$ 是一个常数，所以在算概率分布的时候可以不用管。

$$P(\text{参数}|\text{数据}) \propto P(\text{数据}|\text{参数}) \times P(\text{参数}) \quad (15)$$

如果用 D 表示数据, θ 表示参数, 则有:

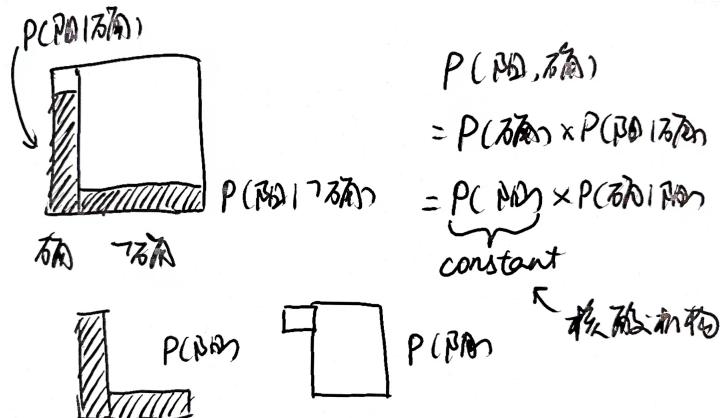
$$P(\theta|D) \propto P(D|\theta) \times P(\theta) \quad (16)$$

这就是贝叶斯定理的核心。

- 先验概率 (Prior): 我们对参数的初始估计 $P(\theta)$
- 似然函数 (Likelihood): 数据在参数下的概率 $P(D|\theta)$
- 后验概率 (Posterior): 我们对参数的最终估计 $P(\theta|D)$

3.1 案例一：核酸检测

张三去做核酸检测。已知新冠的确诊率为 0.0001, 核酸阳性的确诊率为 0.99, 核酸阴性的确诊率为 0.01。假设张三的检测结果为阳性, 求张三确诊的概率。



$$\begin{aligned} P(\text{确诊|阳性}) &= \frac{P(\text{阳性|确诊}) \times P(\text{确诊})}{P(\text{阳性})} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \quad (17) \\ &= 0.0098 \end{aligned}$$