贝叶斯过程

1 概率论背景

1.1 贝叶斯学派

贝叶斯学派将概率定义为对事件发生的信念程度（**Degree of Belief**），并强调利用先验知识更新信念。

- 明天下雨，70%相信 → 根据这几天都下雨，改为90%相信

1.2 频率学派

频率学派认为概率是随机事件在大量重复实验中的频率的极限值

- 扔骰子，6的概率为1/6，扔10000次大概有六分之一都是6

1.3 Compare

- 假设某疾病的发病率（先验概率）为1%，某检测方法的灵敏度（真阳性率）为99%，特异度（真阴性率）为95%。若某人检测结果为阳性，问实际患病的概率是多少？
    - 频率学派方法：（直接使用真阳性率和真阴性率）
      真阳性概率 = 99%
      假阳性概率 = 5%
      得出："检测准确率很高"的结论。忽略了疾病的基础发病率1%
        - 无法直接回答题目问题（实际患病的概率是多少$P(\text{患病}\mid\text{阳性})$）
        - 需要重新收集大量"检测阳性且患病"的样本才能计算
    - 贝叶斯学派方法：使用贝叶斯定理计算后验概率

      $$P(\text{发病}\mid\text{阳性})=\frac{P(\text{阳性}\mid\text{发病})P(\text{发病})}{P(\text{阳性})}=\frac{0.99\times 0.01}{0.99\times 0.01+0.05\times 0.99}\approx 16.3\%$$

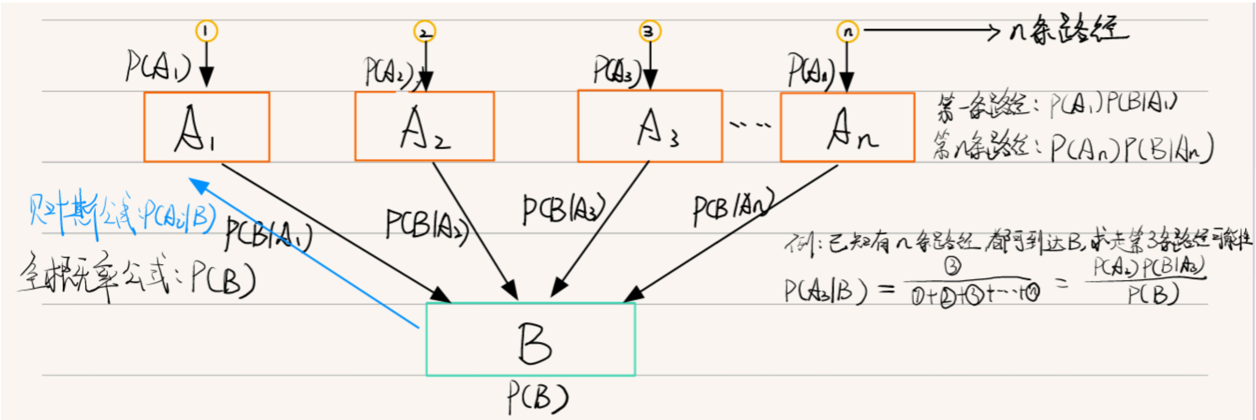
      key:贝叶斯学派通过先验（发病率）更新了概率，频率学派忽略发病率这个关键信息

2 贝叶斯定理（贝叶斯过程的基础）

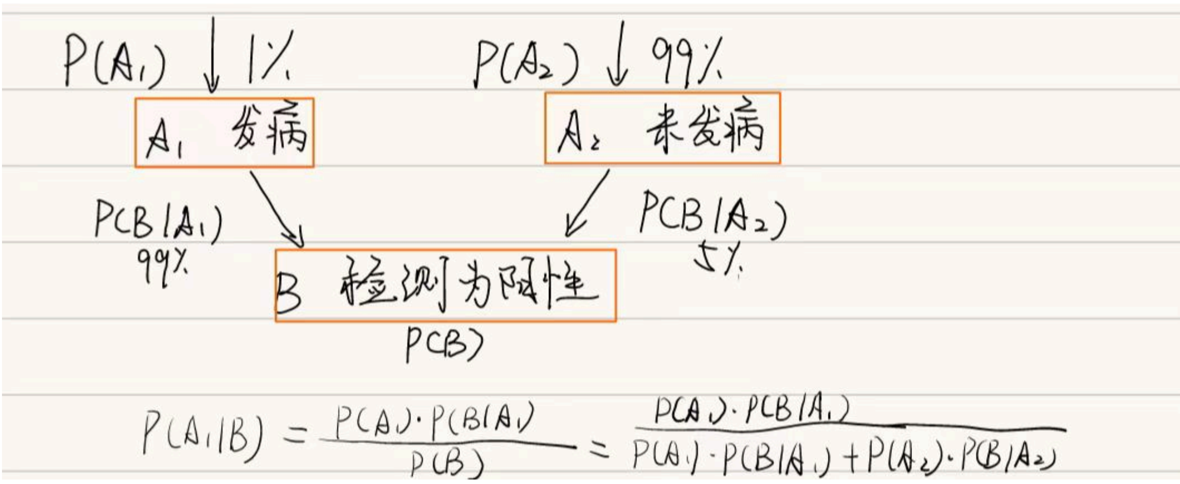
贝叶斯定理公式：

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=0}^{n} P(B|A_j)P(A_j)}$$

- 模型图：
  贝叶斯：已知是结果是B求某一条路径到B的概率
  全概率：求全部到B的概率



  - 例：



$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=0}^{n} P(B|A_j)P(A_j)}$$

- $P(A_i|B)$ - （后验概率）在观察到事件B时，事件A_i的概率
- $P(A_i)$ - （先验概率）在未观察到事件B时，事件A_i的概率
- $P(B)$ - （边缘概率）所有可能A的情况下，B发生的总概率也就是求和/积分
- $P(B|A_i)$ - （似然度）事件A_i已经发生的条件下，事件B发生的概率
  量化事件A_i与事件B之间的相关性
- 变形：

$$P(A_i|B) = P(A_i) * \frac{P(B|A_i)}{P(B)}$$

调节因子：

$$\frac{P(B|A_i)}{P(B)}$$

后验概率 = 先验概率 × 调节因子

$\text{先验概率} \xrightarrow{\text{新的数据或证据}} \text{后验概率}$

3 贝叶斯推断

- 贝叶斯推断是一个推理框架，通过数据更新对未知量的信念
    - 处理的对象是参数，关键工具是贝叶斯公式
- 贝叶斯过程是对整个函数进行建模
    - 在贝叶斯推断中对“数值”进行建模，而贝叶斯过程是对“函数形状”本身建模
      贝叶斯过程 = 把贝叶斯推断应用于函数空间

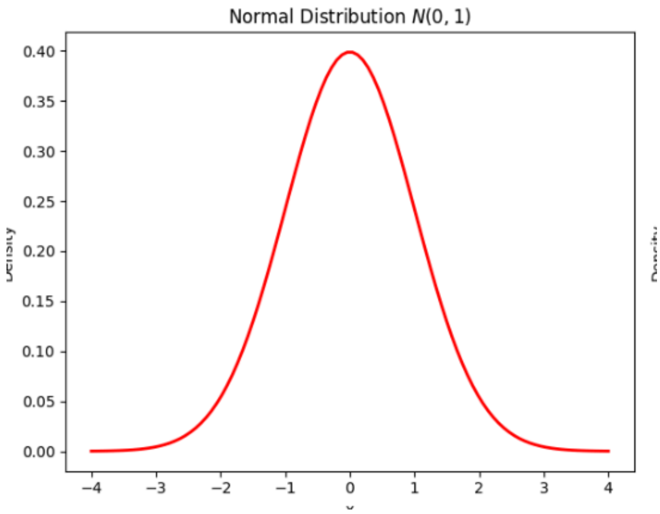
贝叶斯过程：先验→似然→后验

3.1 核心步骤

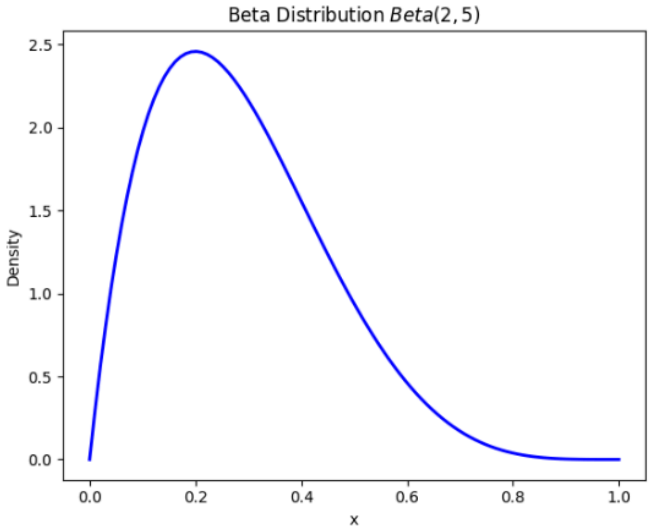
3.1.1 定义先验分布

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

- 目的：反映在看到任何数据之前，我们对参数 θ 的主观信念或知识。$P(\theta)$
- 常用形式：
    - 正态分布 $\mathcal{N}(\mu, \sigma^2)$：适用于连续参数



    - Beta分布$\mathrm{Beta}(\alpha, \beta)$：适用于$[0, 1]$之间的概率建模



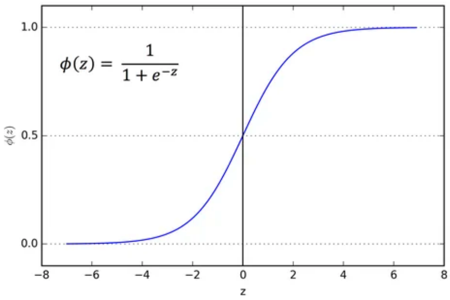
3.1.2 构建似然函数

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

- 目的：建模“在参数 θ 给定时，数据 D 出现的概率”。$P(D \mid \theta)$
- 常见构建方式：
  - 回归问题：假设误差为高斯噪声 → $y_i = f(x_i) + \varepsilon, \varepsilon \sim N(0, \sigma 2)$
    例：预测房价、温度等
    模型的输出y_i是对某个函数$f(x_i)$的预测，再加上一点误差
    - $f(x_i)$：模型对输入x_i的预测（比如线性回归：$f(x) = wx + b$）
    - ε：高斯噪声，表示真实值与预测值之间的不可避免的随机误差
  - 分类问题：伯努利分布或多项式分布
    预测一个离散的类别
    - 模型输出：一个概率$p \in [0, 1]$
    - 假设标签分布：伯努利分布（0-1分布）（事件发生的概率为p，不发生的概率为1-p）

$$y_i \sim Bernoulli(p_i), p_i = \sigma(f(x_i))$$

  - σ是sigmoid函数，保证输出在$[0, 1]$之间



  - 如果 $y_i = 1$（预测是正类），则发生概率为 p_i；否则发生概率为 $1 - p_i$

3.1.3 计算后验分布

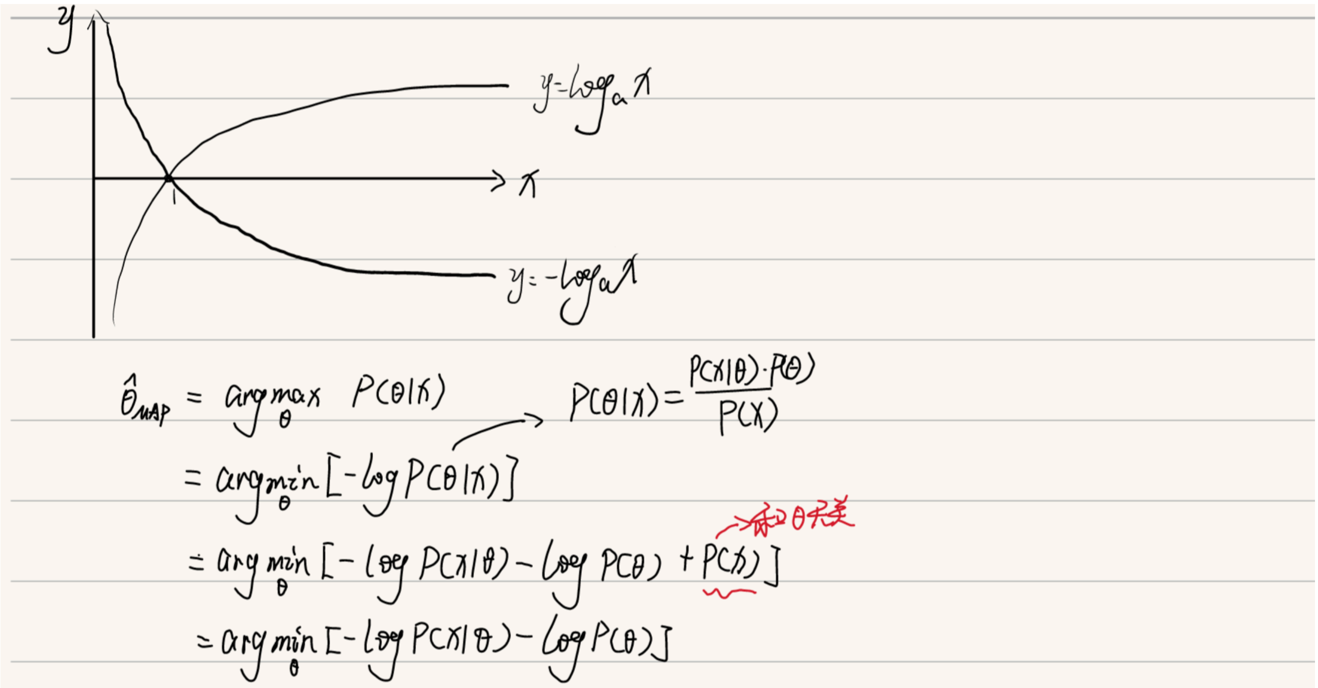
- 目的：结合先验 + 数据，通过贝叶斯公式更新对参数的信念。
- $$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$
- 实际应用中高维积分$P(D) = \int P(D|\theta)P(\theta)d\theta$很难得到精确的结果
  - 解决办法：MCMC（马尔科夫链蒙特卡洛）：用采样逼近积分……

4 最大后验估计（MAP）

在对事物建模时，用 θ 表示模型的参数，请注意，解决问题的**本质就是求 θ**

MAP是贝叶斯学派常用的估计方法

假设数据 $x_1, x_2, \ldots, x_n$ 是i.i.d.（独立同分布）的一组抽样，$X = (x_1, x_2, \ldots, x_n)$。那么MAP对$\theta$的估计方法可以如下推导：



MAP 的定义为：

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta \mid X)$$

将最大化转为最小化负对数形式：

$$= \arg\min_{\theta} [-\log P(\theta \mid X)]$$

代入贝叶斯公式：

$$P(\theta \mid X) = \frac{P(X \mid \theta) P(\theta)}{P(X)}$$

因此：

$$= \arg\min_{\theta} -\log\left(\frac{P(X \mid \theta) P(\theta)}{P(X)}\right)$$

拆分对数项：

$$= \arg\min_{\theta} \left[-\log P(X \mid \theta) - \log P(\theta) + \log P(X)\right]$$

由于 $\log P(X)$ 与 θ 无关，可省略：

$$= \arg\min_{\theta} \left[-\log P(X \mid \theta) - \log P(\theta)\right]$$

5 高斯过程

高斯过程是贝叶斯过程的一个具体体现
example：使用高斯过程对函数进行回归预测