全概率公式 (Law of Total Probability)

September 2, 2025

全概率公式: 分割

概率论中一个重要的公式, 将一个复杂事件的概率计算分解为若干简单事件的概率计算.

定义 0.12 (分割) 若随机事件 $A_1, A_2, \ldots, A_n$ 满足

1. 任意两两事件是互不相容的 $A_i \cap A_j = \emptyset \quad i \neq j$

2. 完备性: $\Omega = \cup_{i=1}^{n} A_i$

称事件 $A_1, A_2, \ldots, A_n$ 为样本空间 Ω 的一个分割, 或称为完备事件组.

全概率公式

概率论中一个重要的公式, 将一个复杂事件的概率计算分解为若干简单事件的概率计算.

定义 0.13 (全概率公式) 若随机事件 $A_1, A_2, \ldots, A_n$ 为样本空间 Ω 的一个分割, 则对任意事件 B 有

$$P(B) = \sum_{i=1}^{n} P(BA_i) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i)$$

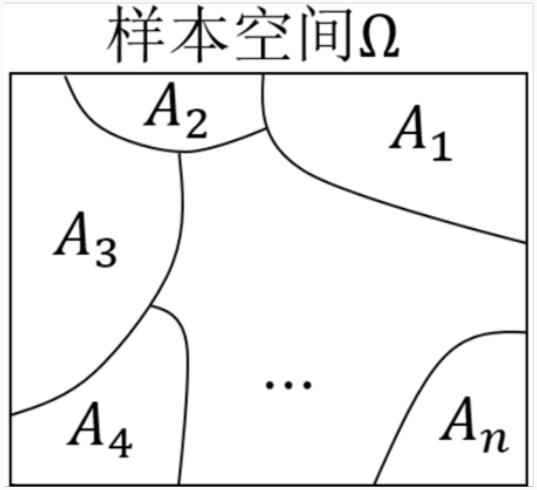
Remarks:

- 将一个复杂事件分解为若干不相容的简单事件之和, 通过分别计算简单事件的概率, 利用概率的可加性得到复杂事件的概率.

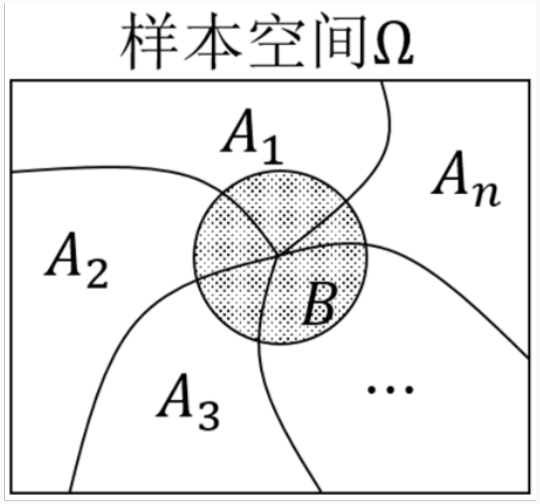
解读: 全概率公式

将事件 B 看作某一过程的结果, 将事件 $A_1, A_2, \ldots, A_n$ 看作产生该结果的若干原因 $\rightarrow P(B)$ 可计算.

1. 每一种原因 A_i 已知, 即 $P(A_i)$ 已知
2. 每一种原因 A_i 对结果 B 的影响已知, 即 $P(B \mid A_i)$ 已知

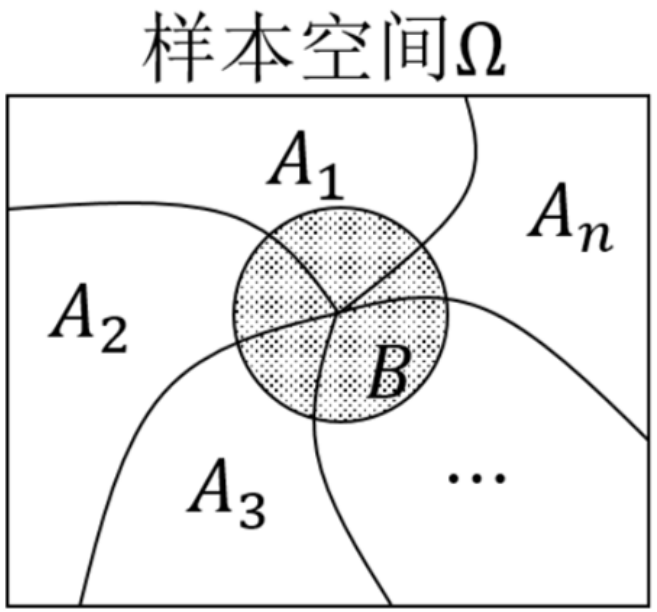


分割



全概率公式

解读: 全概率公式



全概率公式

考虑事件 $A_1, A_2, \ldots, A_n$ 对事件 B 的影响，则有

$$P(B) = \sum_{i=1}^{n} P(BA_i) \quad \text{且} \quad P(BA_i) = P(A_i)P(B \mid A_i)$$

全概率公式: 例 0.41

例 0.41 小明参加一次竞赛, 目前排名不理想, 分析其原因:

- 方法不够新颖的概率为 50%, 通过设计新方法后取得理想排名的概率为 50%,
- 程度代码有误的概率为 30%, 通过纠正代码后取得理想排名的概率为 60%,
- 数据不充分的概率为 20%, 通过采集更多数据后取得理想排名的概率为 80%,

小明可以任意选择 (组合) 策略, 求小明最后取得理想排名的概率.

解答：例 0.41

解答:

- 用 B 表示小明最后取得理想排名的事件, 用 A_1, A_2, A_3 分别表示方法不够新颖, 程度代码有误, 数据不充分这三个事件, 根据题意有

$$P(A_1) = 50\%, P(A_2) = 30\%, P(A_3) = 20\%$$

$$P(B|A_1) = 50\%, P(B|A_2) = 60\%, P(B|A_3) = 80\%.$$

- 小明最后取得理想排名的概率

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 59\%.$$

全概率公式: 例 0.42

例 0.42 随意抛 n 次均匀的硬币, 证明: 正面朝上的次数是偶数 (或奇数) 的概率为 $1/2$.

解答：例 0.42

题目: 随意抛 n 次均匀的硬币, 证明: 正面朝上的次数是偶数 (或奇数) 的概率为 $1/2$.

解答:

- 用事件 A 表示前 $n-1$ 次抛硬币正面朝上的次数为偶数, 其对立事件 $\bar{A}$ 表示 $n-1$ 次抛硬币正面朝上的次数为奇数, 事件 B 表示前 n 次抛硬币正面朝上的次数为偶数. 于是有

$$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) = \frac{P(A)}{2} + \frac{P(\bar{A})}{2} = \frac{1}{2}.$$

- 直接计算概率. 若正面朝上的次数时偶数, 则随意抛 n 次硬币中正面朝上的次数为偶数分别有 $\{0, 2, 4, \ldots, 2k\}(2k \leq n)$, 根据概率公式直接计算有

$$\sum_{0\leq k\leq n/2} \binom{n}{2k} \left(\frac{1}{2}\right)^{2k} \left(\frac{1}{2}\right)^{n-2k} = \frac{1}{2^n} \sum_{0\leq k\leq n/2} \binom{n}{2k} = \frac{1}{2}, \quad \text{其中} \sum_{0\leq k\leq n/2} \binom{n}{2k} = 2^{n-1}$$

全概率公式 vs 条件概率: 抛投不均匀硬币例 0.43

例 0.43 设一个箱子中有 $k+1$ 枚不均匀的硬币, 投掷第 i 枚硬币时正面向上的概率为 i/k $(i=0,1,2,\ldots,k)$. 现从箱子中任意取出一枚硬币, 并任意重复投掷多次, 若前 n 次正面向上, 求第 $n+1$ 次正面向上的概率.

解答: 例 0.43

问题: 设一个箱子中有 $k+1$ 枚不均匀的硬币, 投掷第 i 枚硬币时正面向上的概率为 i/k $(i=0,1,2,\ldots,k)$. 现从箱子中任意取出一枚硬币, 并任意重复投掷多次, 若前 n 次正面向上, 求第 $n+1$ 次正面向上的概率.

解答: 本题中, 事件 “前 n 次正面向上” 和 “第 $n+1$ 次正面向上” 在 “从箱子中任意取出一枚硬币反复抛掷多次” 发生的情况下是条件独立的, 即同一枚硬币抛掷的结果是独立事件.

用 A 表示第 $n+1$ 次投掷正面向上的事件, 用 B 表示前 n 次投掷正面向上的事件, 用 C_i 表示从箱子中取出第 i 枚硬币的事件 $(i=0,1,2,\ldots,k)$. 求 $P(A \mid B)$.

方法一：条件概率公式拆解 $P(A \mid B)$，事件 A 和 B 独立

- 因为 $P(A \mid B) = P(AB)/P(B)$, 且

$$P(AB) = \sum_{i=0}^{k} P(C_i)P(AB \mid C_i) = \sum_{i=0}^{k} P(C_i)P(A \mid C_i)P(B \mid C_i) = \frac{1}{k+1}\sum_{i=0}^{k} \frac{i^{n+1}}{k^{n+1}}$$

以及全概率公式

$$P(B) = \sum_{i=0}^{k} P(C_i)P(B \mid C_i) = \frac{1}{k+1} \sum_{i=0}^{k} \frac{i^n}{k^n}$$

由此可知

$$P(A \mid B) = \frac{\sum_{i=0}^{k} (i/k)^{n+1}}{\sum_{i=0}^{k} (i/k)^n}$$

- 另外, 当 k 非常大或 $k \to \infty$ 时可利用积分近似

$$\frac{1}{k} \sum_{i=1}^{k} (i/k)^n \approx \int_0^1 x^n dx = \frac{1}{n+1} \qquad \text{和} \qquad \frac{1}{k} \sum_{i=1}^{k} (i/k)^{n+1} \approx \int_0^1 x^{n+1} dx = \frac{1}{n+2}$$

此时有 $P(A|B) \approx (n+1)/(n+2)$.

解答: 例 0.43

方法二：全概率公式拆解 $P(A \mid B)$

- 用 A 表示第 $n+1$ 次投掷正面向上的事件, 用 B 表示前 n 次投掷正面向上的事件, 用 C_i 表示从箱子中取出第 i 枚硬币的事件 $(i = 0, 1, 2, \ldots, k)$.
- 因为

$$P(A \mid B) = \sum_i P(A \mid BC_i)P(C_i \mid B) \neq \sum_i P(A \mid BC_i)P(C_i)$$

其中,

$$P(A \mid BC_i) = \frac{P(AB \mid C_i)}{P(B \mid C_i)}$$

和

$$P(C_i \mid B) = \frac{P(C_i)P(B \mid C_i)}{P(B)}$$

- 因此, 有

$$P(A \mid B) = \sum_i \frac{P(AB \mid C_i)P(C_i)}{P(B)}$$

只需要计算: $P(AB \mid C_i)$, $P(C_i)$, 和 $P(B)$.

思考: 例 0.44

例 0.44 假设有 n 个箱子, 每个箱子里有 a 只白球和 b 只红球, 现从第一个箱子取出一个球放入第二个箱子, 第二个箱子取出一个球放入第三个箱子, 依次类推, 求从最后一个箱子取出一球是红球的概率.

解答：例 0.44

题目: 假设有 n 个箱子, 每个箱子里有 a 只白球和 b 只红球, 现从第一个箱子取出一个球放入第二个箱子, 第二个箱子取出一个球放入第三个箱子, 依次类推, 求从最后一个箱子取出一球是红球的概率.

解答:

- 用 A_i 表示第 i 个箱子取出红球的事件 $(i \in [n])$, 则 $\bar{A}_i$ 表示第 i 个箱子取出白球的事件, 于是有

$$P(A_1) = b/(a+b) \qquad \text{和} \qquad P(\bar{A}_1) = a/(a+b).$$

- 根据全概率公式有

$$\begin{aligned} P(A_2) &= P(A_1)P(A_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) \\ &= \frac{b}{a+b} \times \frac{b+1}{a+b+1} + \frac{a}{a+b} \times \frac{b}{a+b+1} = \frac{b}{a+b}. \end{aligned}$$

  由此可知 $P(\bar{A}_2) = a/(a+b)$. 依次类推重复上述过程 $n-1$ 次, 最后一个箱子中取出一球是红球的概率为 $b/(a+b)$.

思考: 例 0.45

例 0.45 有两个箱子，1 号箱有 n_1 个白球，m_1 个红球，2 号箱有 n_2 个白球，m_2 个红球.

- 先从 1 号箱任取一只球放到 2 号箱中，再从 2 号箱中任取一球. 问求得白球的概率？
- 先从 1 号箱任取两只球放到 2 号箱中，再从 2 号箱中任取一球. 问求得白球的概率？

解答：例 0.45

题目: 有两个箱子，1 号箱有 n_1 个白球，m_1 个红球，2 号箱有 n_2 个白球，m_2 个红球.

解答:

- (1) 先从 1 号箱任取一只球放到 2 号箱中，再从 2 号箱中任取一球. 问求得白球的概率？
- A = " 1 号箱取到白球"，B="2 号箱取到白球"，则 $B = \bar{A}B + AB$ (记号)，进而有

$$P(B) = P(\bar{A}B) + P(AB) = P(\bar{A})P(B|\bar{A}) + P(A)P(B|A)$$
$$= \frac{m_1}{n_1+m_1}\frac{n_2}{n_2+m_2+1} + \frac{n_1}{n_1+m_1}\frac{n_2+1}{n_2+m_2+1}$$

- (2) 先从 1 号箱任取两只球放到 2 号箱中，再从 2 号箱中任取一球. 问求得白球的概率？
- A_i = “ 1 号箱恰好取到 i 个白球” ($i = 0, 1, 2$)，B=“2 号箱取到白球”，则

$$\begin{aligned} P(B) &= P(A_0B) + P(A_1B) + P(A_2B) \\ &= P(A_0)P(B|A_0) + P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \\ &= \frac{C_{m_1}^2}{C_{n_1+m_1}^2}\frac{n_2}{n_2+m_2+2} + \frac{C_{m_1}^1 C_{n_1}^1}{C_{n_1+m_1}^2}\frac{n_2+1}{n_2+m_2+2} + \frac{C_{n_1}^2}{C_{n_1+m_1}^2}\frac{n_2+2}{n_2+m_2+2} \end{aligned}$$

贝叶斯公式 (Bayesian Formula)

September 2, 2025

贝叶斯公式

概率论中另一个重要的公式, 研究在一种结果已发生的情况下事何种原因导致了该结果.

定义 0.14 (贝叶斯公式) 若随机事件 $A_1, A_2, \ldots, A_n$ 为样本空间 Ω 的一个分割, 事件 B 满足 $P(B) > 0$. 则有:

$$P(A_i \mid B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(B)}$$

特别地, 分子结合乘法公式

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(B)}$$

且分母结合全概率公式

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(\bar{A})P(B \mid \bar{A})}$$

解读: 贝叶斯公式

将事件 B 看作结果, 将事件 A_i 看作产生结果的原因之一. 如果

- 该原因发生的概率 $P(A_i)$ 已知
- 原因 A_i 对结果 B 的影响已知, 即概率 $P(B \mid A_i)$ 已知

则可求事件 B 由某种原因引起的概率 $P(A_i \mid B)$.

$$P(A_i \mid B) = \frac{P(A_iB)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(B)}$$

贝叶斯公式：例 0.46

例 0.46 小明参加一次竞赛, 目前排名不理想, 分析其原因:

- 方法不够新颖的概率为 50%, 通过设计新方法后取得理想排名的概率为 50%,
- 程度代码有误的概率为 30%, 通过纠正代码后取得理想排名的概率为 60%,
- 数据不充分的概率为 20%, 通过采集更多数据后取得理想排名的概率为 80%,

因为时间有限, 小明只能选择三种策略 (设计新方法、纠正代码、采集更多数据) 中一种, 想要取得理想排名, 小明应该选择哪一种方案.

解答：例 0.46

解答:

- 用 B 表示小明最后取得理想排名的事件, 用 A_1, A_2, A_3 分别表示方法不够新颖, 程度代码有误, 数据不充分这三个事件, 根据题意有

$$P(A_1) = 50\%, \qquad P(A_2) = 30\%, \qquad P(A_3) = 20\%$$

$$P(B|A_1) = 50\%, \qquad P(B|A_2) = 60\%, \qquad P(B|A_3) = 80\%$$

- 小明最后取得理想排名的概率

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = 59\%.$$

- 根据贝叶斯公式有

$$P(A_1|B) = P(A_1)P(B|A_1)/P(B) = 25/59, P(A_2|B) = 18/59, P(A_3|B) = 16/59.$$

因此小明应该选择设计新方法来获得理想排名的概率更高.

例 0.41 vs 例 0.46

描述: 小明参加一次竞赛, 目前排名不理想, 分析其原因:

- 方法不够新颖的概率为 50%, 通过设计新方法后取得理想排名的概率为 50%,
- 程度代码有误的概率为 30%, 通过纠正代码后取得理想排名的概率为 60%,
- 数据不充分的概率为 20%, 通过采集更多数据后取得理想排名的概率为 80%,

问题：

(1) 因为时间有限, 小明只能选择三种策略 (设计新方法、纠正代码、采集更多数据) 中一种,

(2) 小明可以任意选择 (组合) 策略（留作作业；可以使用一种，两种，或者三种方法）

求小明最后取得理想排名的概率.

贝叶斯公式：towards 机器学习

贝叶斯公式在机器学习中提供一种概率论框架下实施决策的基本方法:

- 求 后验 (posterior) 概率, 即事件 B 发生后各个原因 A_i 的概率
- $P(A_i)$ 表示原因 A_i 发生的 先验 (prior) 概率
- $P(B \mid A)$ 表示事件结果 B 相对于原因 A 的条件概率 (conditional probability), 或称之为 似然 (likelihood)
- $P(B)$ 是用于归一化的 证据 (evidence) 因子

根据例题, 利用 $P(A_i \mid B)$ 来决策的时候, 与证据因子 $P(B)$ 无关. 因此, 将估计 $P(A_i \mid B)$ 的问题转化为如何基于样本集来估计先验 $P(A_i)$ 和似然 $P(B \mid A_i)$.

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)} \overset{\text{converting}}{\Rightarrow} \text{后验概率} = \frac{\text{先验概率} \times \text{似然}}{\text{证据因子}}$$

贝叶斯公式：例 0.47

例 0.47 (贝叶斯公式找原因) 一个仓库中有 10 个同规格的产品，其中的 5、3、2 个分别为甲、乙、丙生产的. 已知甲、乙、丙生产的次品率分别为 $\frac{1}{10}$、$\frac{1}{15}$、$\frac{1}{20}$. 现在任取一个产品进行质检，检到为正品的概率是多少？如果检测的产品是正品，该产品是甲生产的概率为多少？

解答：例 0.47

解答:

- 设 A = “检测产品为正品”，B_1 = “该产品是甲生产的”，依次定义 B_2 和 B_3.
- 检测为正品的概率为

$$P(A) = P(B_1)P(A \mid B_1) + P(B_2)P(A \mid B_2) + P(B_3)P(A \mid B_3) = 0.92\ .$$

- 如果检测的产品是正品，该产品是甲生产的概率为

$$P(B_1 \mid A) = \frac{P(B_1)P(A \mid B_1)}{P(A)} = 0.4891\ .$$

拓展一：先验概率与主观概率

- 贝叶斯公式最存在争议之处: 先验概率 $P(A_i)$ 的选取
- 很多实际应用中根据以往的数据得先验, 符合概率的频率解释, 但需要以往大量的历史数据, 在实际应用中通常难以满足
- 很多应用中先验概率可能由某一种主观的方式给出, 例如对未来宏观经济形势、或对某人诚信度
- 主观概率: 将概率解释为信任程度、明显带有主观性

贝叶斯公式的应用：例 0.48

例 0.48 寓言故事狼来了：一个小孩每天到山上放羊, 山里有狼出没, 第一天他在山上喊“狼来了！狼来了!”, 山下的村民们闻声便去打狼, 到了山上发现没有狼; 第二天仍是如此; 第三天狼真来了, 可无论小孩怎么喊叫, 也没有人来救他. 假设

- 村民们对这个小孩的印象一般, 认为小孩说谎话和说真话的概率相同, 均为 $1/2$
- 小孩说谎话——喊狼来了时, 狼真来的概率为 $1/3$
- 小孩说真话——喊狼来了时, 狼真来的概率为 $3/4$

若第一天、第二天上山均没有发现狼, 请分析这两天中村民们的心理活动——“对小孩说谎与否的认识”.

解答：例 0.48

解答:

- 用 B_1, B_2 分别表示第一天和第二天狼来了的事件. 用 A_1 表示小孩第一天说谎的事件, 用 A_2 表示小孩第一天没有狼的情况下第二天仍说谎的事件, 根据题意可知

$$P(A_1) = P(\bar{A_1}) = \frac{1}{2}, \qquad P(B_1|A_1) = \frac{1}{3}, \qquad P(B_1|\bar{A_1}) = \frac{3}{4}$$

$$P(B_2|A_2) = \frac{1}{3}, \qquad P(B_2|\bar{A_2}) = \frac{3}{4}$$

- 根据贝叶斯公式可知, 村民第一天上山打狼但是没有发现狼时, 认为小孩说谎的概率发生了变化,

$$P(A_2) = P(A_1|\bar{B_1}) = \frac{P(\bar{B_1}|A_1)P(A_1)}{P(\bar{B_1}|A_1)P(A_1) + P(\bar{B_1}|\bar{A_1})P(\bar{A_1})} = \frac{8}{11} \approx 0.7273.$$

- 村民第天上山打狼但是没有发现狼时, 认为小孩说谎的概率又发生了变化,

$$P(A_2|\bar{B_2}) = \frac{P(\bar{B_2}|A_2)P(A_2)}{P(\bar{B_2}|A_2)P(A_2) + P(\bar{B_2}|\bar{A_2})P(\bar{A_2})} = \frac{64}{73} \approx 0.8767.$$

村民认为小孩说谎的概率从 $50\% \to 72.73\% \to 87.67\%$, 不再相信狼来了的谎言.

拓展二：全概率公式 VS 贝叶斯公式

- **Recall:** 将事件 $A_1, A_2, \ldots, A_n$ 看作事件 B 发生的原因, 而事件 B 是伴随着原因 $A_1, A_2, \ldots, A_n$ 而发生的结果.
- 应用条件是相同的:
  - 事件 $A_1, A_2, \ldots, A_n$ 为样本空间 Ω 的一个分割
- 解决的问题不同:
  - 若知道各种原因 $P(A_i)$、在该原因下事件 B 发生的概率 $P(B \mid A_i)$, 此时利用全概率公式计算概率 $P(B)$.
  - 若知道各种原因 $P(A_i)$、在该原因下事件 B 发生的概率 $P(B \mid A_i)$, 若结果事件 B 已经发生, 利用贝叶斯公式探讨是由某原因 A_i 导致该结果的概率 $P(A_i \mid B)$.

贝叶斯公式：例 0.49

例 0.49 (三囚徒问题) 犯人 a, b, c 均被判为死刑, 法官随机赦免其中一人, 看守知道谁被赦免但不会说. 犯人 a 问看守: b 和 c 谁会被执行死刑? 看守的策略:

1. 若赦免 b, 则说 c
2. 若赦免 c, 则说 b
3. 若赦免 a, 则以 $1/2$ 的概率说 b 或 c

看守回答犯人 a: 犯人 b 会被执行死刑. 犯人 a 兴奋不已, 因为自己生存的概率为 $1/2$. 犯人 a 将此事告诉犯人 c. c 同样高兴, 因为他觉得自己的生存几率为 $2/3$.

那么谁才是正确的呢?

解答：例 0.49

- 用事件 A, B, C 分别表示犯人 a, b, c 被赦免, 因为法官随机赦免, 所以

$$P(A) = P(B) = P(C) = 1/3\,.$$

- 用事件 D 表示看守人说犯人 b 被执行死刑, 根据看守的策略, 有

$$P(D \mid A) = 1/2\,, \quad P(D \mid B) = 0\,, \quad P(D \mid C) = 1\,.$$

- 我们要求解的是 “哪个犯人才是导致 D 事件发生的最大原因”, 即求 max,

$$P(A \mid D) = \frac{P(A)P(D \mid A)}{P(D)}\,, \quad P(C \mid D) = \frac{P(C)P(D \mid C)}{P(C)}$$

- 根据上式我们还需要知道 $P(D)$, 可以根据全概率公式求得

$$P(D) = P(A)P(D \mid A) + P(B)P(D \mid B) + P(C)P(D \mid C) = 1/2\,.$$

- 最后得到, $P(A \mid D) = 1/3$ and $P(C \mid D) = 2/3$.

拓展三：贝叶斯 Spam 过滤器 - 例 0.50

例 0.50 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.

贝叶斯公式的应用：例 0.50

例 0.51 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. (目标导向)
3. 设 S 是事件 "邮件为 Spam", E 是事件 "邮件内容含单词 ω". 需计算 $P(S \mid E)$ 或者比较 $P(S \mid E)$ 和 $P(\bar{S} \mid E)$ 的大小.

贝叶斯公式的应用：例 0.50

例 0.52 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. (目标导向)
3. 设 S 是事件 "邮件为 Spam", E 是事件 "邮件内容含单词 ω". 需计算 $P(S \mid E)$ 或者比较 $P(S \mid E)$ 和 $P(\bar{S} \mid E)$ 的大小.

思路: 根据条件概率公式和全概率公式

$$P(S \mid E) = \frac{P(SE)}{P(E)} = \frac{P(S)P(E \mid S)}{P(E)} = \frac{P(S)P(E \mid S)}{P(S)P(E \mid S) + P(\bar{S})P(E \mid \bar{S})}$$

我们需要分别估算：Spam 邮件中含有单词 ω 的概率 $P(E \mid S)$ 和非 Spam 邮件中含有单词 ω 的概率 $P(E \mid \bar{S})$.

贝叶斯公式的应用：例 0.50

例 0.53 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. (目标导向)
3. 设 S 是事件 "邮件为 Spam", E 是事件 "邮件内容含单词 ω". 需计算 $P(S \mid E)$ 或者比较 $P(S \mid E)$ 和 $P(\bar{S} \mid E)$ 的大小.
4. 我们需要分别估算：Spam 邮件中含有单词 ω 的概率 $P(E \mid S)$ 和非 Spam 邮件中含有单词 ω 的概率 $P(E \mid \bar{S})$.
5. 统计该单词 ω 在集合 B 和 G 中出现的频率分别为 $p_B(\omega)$ 和 $p_G(\omega)$. 认为：$P(E \mid S) = p_B(\omega)$ 和 $P(E \mid \bar{S}) = p_G(\omega)$.

贝叶斯公式的应用：例 0.50

例 0.54 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. 如何选定 "有效的" 单词 ω 呢？(原因导向)

贝叶斯公式的应用：例 0.50

例 0.55 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. 如何选定 "有效的" 单词 ω 呢？(原因导向)
3. 直觉上应该选择：Spam 邮件中含有单词 ω 的概率 $P(E \mid S)$ 高，而非 Spam 邮件中含有单词 ω 的概率 $P(E \mid \bar{S})$ 低的单词 ω. (但是这两个概率值不是互补的，可能一同大小)

贝叶斯公式的应用：例 0.50

例 0.56 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 考察一个特定的单词 ω, 计算 "如果一封邮件中含有单词 ω，则该邮件是 Spam" 的概率. 如何选定 "有效的" 单词 ω 呢？(原因导向)
3. 直觉上应该选择：Spam 邮件中含有单词 ω 的概率 $P(E \mid S)$ 高，而非 Spam 邮件中含有单词 ω 的概率 $P(E \mid \bar{S})$ 低的单词 ω.
4. 回到条件概率公式 $P(S \mid E) = P(S)P(E \mid S)/P(E)$，其中，$P(E)$ 是集合 B 和 G 中包含单词 ω 的邮件的频率，$P(S)$ 是集合 B 和 G 中 Spam 邮件的频率，$P(E \mid S)$ 被认为单词 ω 在集合 B 中出现的频率 $p_B(\omega)$.

贝叶斯公式的应用：例 0.50

例 0.57 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 认为检测 "某单词在邮件中出现" 就可以识别 Spam 邮件. 一封邮件所涉及的词库包括 $\omega_1, \ldots, \omega_n$. 为了选择对识别 Spam 邮件 "有效的" 单词 ω，我们需要计算 (贝叶斯公式)

$$P(E_i \mid S) = \frac{P(E_i)P(S \mid E_i)}{P(S)}$$

其中，E_i 表示邮件中包含单词 ω_i. 然后比较 $P(E_i \mid S)$ 的大小关系. (真正的原因导向)

贝叶斯公式的应用：例 0.50

例 0.58 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

1. 假设我们有一个垃圾邮件的集合 B 和一个不是垃圾的邮件集合 G. 利用贝叶斯公式来预测一个新的电子邮件是 Spam 的概率.
2. 认为检测 "某单词在邮件中出现" 就可以识别 Spam 邮件. 一封邮件所涉及的词库包括 $\omega_1, \ldots, \omega_n$. 为了选择对识别 Spam 邮件 "有效的" 单词 ω，我们需要计算 (贝叶斯公式)

$$P(E_i \mid S) = \frac{P(E_i)P(S \mid E_i)}{P(S)}$$

   其中，E_i 表示邮件中包含单词 ω_i. 然后比较 $P(E_i \mid S)$ 的大小关系. (真正的原因导向)
3. any case?