拓展三：贝叶斯 Spam 过滤器 - 例 0.46

例 0.46 (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 0.46

例 0.47 (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 0.46

例 0.48 (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 0.46

例 0.49 (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 0.46

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.
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 0.46

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. 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 0.46

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. 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 0.46

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. 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 0.46

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. 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?

案例分析: 概率计算

(习题课)

回答思考题、补充例题、复盘作业

September 29, 2024

三囚徒问题: 例 0.55

例 0.55 (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 0.55

问题: 三犯人 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? 看守的回答策略为: i) 若赦免 b , 则说 c; ii) 若赦免 c , 则说 b; iii) 若赦免 a , 则以 1/2 的概率说 b 或 c ; 看守回答 a : 犯人 b 会被执行死刑. 犯人 a 兴奋不已, 认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, 犯人 a 和犯人 c 中谁的想法是正确的?

解答:

- 事件“看守说犯人 b 会被执行死刑”的“原因”有两种情况, 即事件“赦免 c , 看守说 b 会被执行死刑”或者事件“赦免 a , 看守说以 1/2 的概率说 b 或 c 会被执行死刑” .
- 用事件 A, B, C 分别表示 a, b, c 被赦免, 则 $P(A) = P(B) = P(C) = 1/3$. 用事件 D 表示看守说犯人 b 被执行死刑, 则

$$P(D|A) = 1/2 \qquad P(D|B) = 0 \qquad P(D|C) = 1$$

由全概率公式有

$$P(D) = P(A)P(D|A) + P(B)P(D|B) + P(C)P(D|C) = 1/2$$

由贝叶斯公式有

$$P(A|D) = P(A)P(D|A)/P(D) = 1/3$$

和

$$P(C|D) = P(C)P(D|C)/P(D) = 2/3$$

所以犯人 a 的想法是错误的, 犯人 c 的想法是正确的.

解答: 例 0.55

- 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 $P(C \mid D) = 2/3$.

抛投不均匀硬币: 例 0.56

例 0.56 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 0.56

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 0.56

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

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

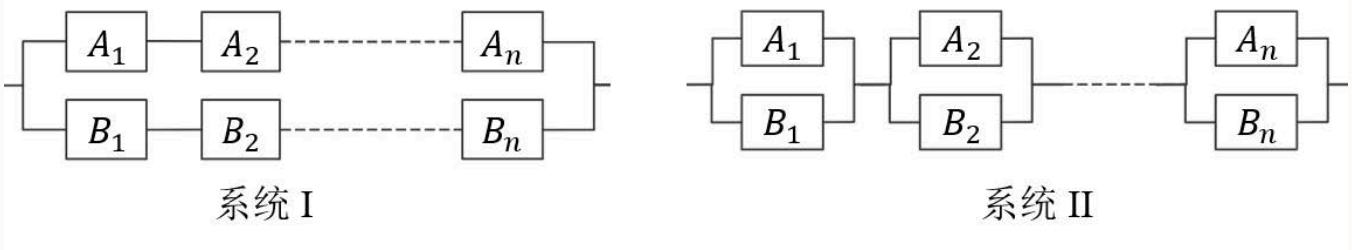
- 因此, 有

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

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

电路可靠性: 例 0.57

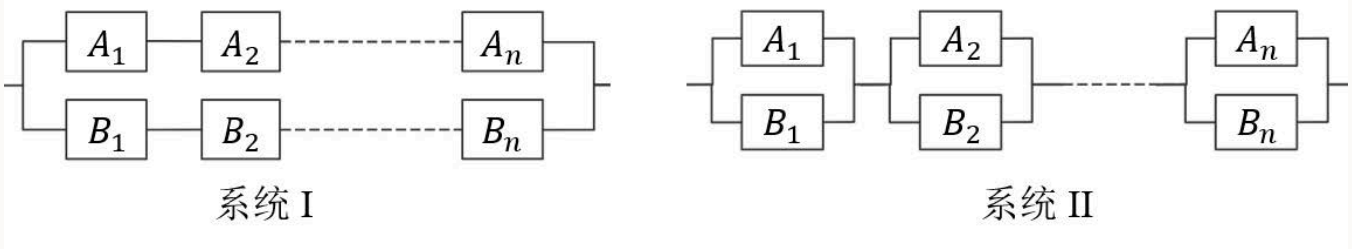
例 0.57 设构成系统的每个元件的可靠性均为 $p\,(0<p<1)$, 且各元件是否正常工作是相互独立的. 设有 $2n$ 个元件按下图所示, 两种不同连接方式构成两个不同的系统, 比较这两种系统的可靠性大小.



解答: 例 0.57

问题: 设构成系统的每个元件的可靠性均为 $p\ (0 < p < 1)$, 且各元件是否正常工作是相互独立的. 设有 $2n$ 个元件按下图所示, 两种不同连接方式构成两个不同的系统, 比较这两种系统的可靠性大小.

解答:



- 对于系统 I, 它能正常工作当且仅当系统中的两条通路至少有一条正常工作, 而每条通路正常工作当且仅当它的每个元件都能正常工作; 对于系统 II, 它能正常工作当且仅当每对并联元件能够正常工作.
- 用事件 A_i 和 B_i 表示图中对应元件正常工作 $(i = 0, 1, 2, \ldots, n)$, 因此系统 I 的可靠

性为

$$\begin{aligned}
&P((A_1A_2\ldots A_n)\cup(B_1B_2\ldots B_n))\\
&=P(A_1A_2\ldots A_n)+P(B_1B_2\ldots B_n)-P(A_1A_2\ldots A_nB_1B_2\ldots B_n)\\
&=2p^n-p^{2n}=p^n(2-p^n)
\end{aligned}$$

系统 II 可靠性为

$$P\left(\bigcap_{i=1}^{n}(A_i\cup B_i)\right)=\prod_{i=1}^{n}P(A_i\cup B_i)=(2p-p^2)^n=p^n(2-p)^n$$

利用数学归纳法可证明当 $n\geq 2$ 时有 $(2-p)^n>2-p^n$ 成立, 由此可知系统 II 的可靠性更好.

多项式相等: 0.58

例 0.58 给定两个较复杂的多项式

$$F(x) = (x+2)^7(x+3)^5 + (x+1)^{100} + (x+2)(x+3) + x^{20}$$

$$G(x) = (x+3)^{100} - (x+1)^{25}(x+2)^{30} + x^{20} + (x-2)(x-3)\dots(x-100)$$

如何快速验证 $F(x) \equiv G(x)$?

解答: 例 0.58

问题: 如何快速验证 $F(x) \equiv G(x)$, 给定两个较复杂的多项式

$$F(x) = (x+2)^7(x+3)^5 + (x+1)^{100} + (x+2)(x+3) + x^{20}$$

$$G(x) = (x+3)^{100} - (x+1)^{25}(x+2)^{30} + x^{20} + (x-2)(x-3)\ldots(x-100)$$

解答:

- 若通过展开多项式合并同类项, 比较每项系数是否相同的方法来验证 $F(x) \equiv G(x)$, 则需要较高的计算时间开销.
- 设计一种利用独立随机性方法来验证 $F(x) \equiv G(x)$ 是否正确, 使得该方法验证结果为正确的概率较高, 同时降低计算时间.
  - 假设 $F(x)$ 或 $G(x)$ 的最高次项不超过 d, 考虑从集合 $[100d] = \{1, 2, \ldots, 100d\}$ 中等可能独立地随机选取 $k(< d)$ 个数 $r_1, r_2, \ldots, r_k$. 若存在 r_i 使得 $F(r_i) \neq G(r_i)$ 成立, 则返回 $F(x) \neq G(x)$, 否则返回 $F(x) \equiv G(x)$.
  - 为什么要用 "$[100d]$"?
  - 分析该方法的正确性:

1. 若多项式 $F(x) \equiv G(x)$, 则该方法得到“正确”的结果, 因为对于任意 $r_i \in [100d]$ 都有 $F(x) = G(x)$;
2. 若多项式 $F(x) \neq G(x)$ 且 $F(r_i) \neq G(r_i)$, 则该方法得到“正确”的结果, 因为存在一个 r_i 使得 $F(r_i) \neq G(r_i)$;
3. 若多项式 $F(x) \neq G(x)$ 但 $F(r_i) = G(r_i)$, 即存在 $r_i \in [100d]$ 使得 $F(r_i) = G(r_i)$ 成立, 此时 r_i 为多项式 $F(x) - G(x) = 0$ 的一个实数根. 根据代数知识可知最高次项不超过 d 的多项式 $F(x) - G(x) = 0$ 至多有 d 个实数根, 而 r_i 为 $[100d]$ 中等可能随机选取, 因此有

$$P(F(r_i) = G(r_i)) \leq \frac{d}{100d} = \frac{1}{100} .$$

进而有,

$$P\left(\bigcap_{i=1}^{k}\{F(r_i) = G(r_i)\}\right) = \prod_{i=1}^{k} P(F(r_i) = G(r_i)) \leq \frac{1}{100^k} .$$

矩阵乘法相等: 例 0.59

例 0.59 给定矩阵 $A, B, C \in \{0,1\}^{n\times n}(n \geq 10000000)$, 如何快速验证 $AB = C$.

解答: 例 0.59

问题: 给定矩阵 $A, B, C \in \{0,1\}^{n\times n}(n \geq 10000000)$, 如何快速验证 $AB = C$.

解答:

- 若直接采用矩阵乘法计算 AB , 再与矩阵 C 进行比较, 计算复杂度开销为 $O(n^3)$
- 类似于验证多项式 $F(x) \equiv G(x)$ 的方法, 随机选取一个向量 $\bar{r} = (r_1, r_2, \ldots, r_n)^T$, 其中元素 $r_1, r_2, \ldots, r_n$ 都是从 $\{0,1\}$ 独立等可能随机选取所得, 通过验证事件

$$\text{存在一个向量 } \bar{r} \text{ 使得 } A(B\bar{r}) \neq C\bar{r}$$

  发生的概率极小来反向验证 $AB = C$.

- 这里, r 可以理解为 “函数” AB 和 C 的自变量.

解答: 例 0.59

- Freivalds (弗赖瓦尔兹) 算法

输入: 矩阵 $\mathbf{A}, \mathbf{B}, \mathbf{C}$
输出: 是/否　　%% 验证 $\mathbf{AB} \stackrel{?}{=} \mathbf{C}$

For $i = 1 : k$
　随机选择向量 $\bar{\mathbf{r}}_i = (r_{i1}, r_{i2}, \cdots, r_{in})$, 其每个元素是从 $\{0, 1\}$ 独立等可能随机采样所得
　计算向量 $\bar{\mathbf{p}}_i = \mathbf{A}(\mathbf{B})\bar{\mathbf{r}}_i - \mathbf{C}\bar{\mathbf{r}}_i$
　If $\{\bar{\mathbf{p}}_i$ 不是零向量$\}$ then
　　返回"否"
　EndIf
EndFor
返回"是".

- 计算事件"存在一个向量 $\bar{r}$ 使得 $A(B\bar{r}) \neq C\bar{r}$" 发生的概率. 设随机变量 $\bar{r_1}, \bar{r_2}, \ldots, \bar{r_k} \in \{0, 1\}^n$ 中每个元素都是从 $\{0, 1\}$ 独立等可能随机选取所得, 若 $AB \neq C$, 则有

$$P\left[\bigcap_{i=1}^{k}\{A(B\bar{r_i}) = C\bar{r_i}\}\right] \leq \frac{1}{2^k}$$

小结与发散

通过例 0.58 和例 0.59, 我们可以发现在证明一些数学的等式的时候可以通过 “设计随机试验 + 概率计算” 的方式来证明.

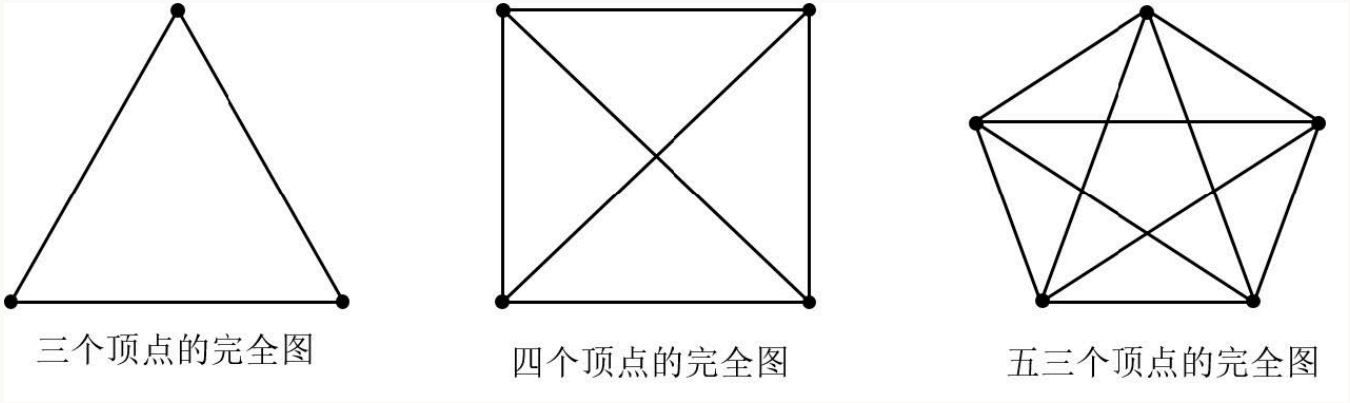
进一步, 我们可以思考如下问题要如何解决呢?

- 证明不等式 $2ab \leq a^2 + b^2$
- 比较 n^k 和 $(k+1)^{n+1}$ 的大小?
- F 和 G 是两个不规则的图形, 比较这两者的面积大小?

这套方法有何利弊?

完全图着色 (边着色): 例 0.60

例 0.60 设平面上有 n 个顶点, 其中任意三个顶点不在同一条直线上, 用 $n(n-1)/2$ 条边将这些顶点连接起来的图称为 n 个顶点的完全图. 例如三个、四个、五个顶点的完全图如下所示.



三个顶点的完全图　　四个顶点的完全图　　五三个顶点的完全图

现将图中的每条边都分别染成红色或蓝色, 给定两正整数 $n \geq 10$ 和 $k > n/2$, 是否存在一种染色方法, 使得图上任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色?

解答: 例 0.60

问题: 如上所述.

解答:

- 若通过穷举的方法, 则计算的开销较大
- 可以利用概率的方法证明至少存在一种染色方法使得任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色
  - 假设每条边都等可能独立地被染成红色或蓝色, 即每条边为红色或蓝色的概率均为 $1/2$. 从 n 个不同的顶点中选出 k 个顶点有 $\binom{n}{k}$ 种不同的选法, 分别对应于 $\binom{n}{k}$ 个包含 k 个顶点的子集, 这里将的子集分别标号为 $1, 2, \ldots, \binom{n}{k}$.
  - 用 E_i 表示第 i 个子集中 $k(k-1)/2$ 条边染成相同颜色的事件, 根据题意可得

$$P(E_i) = 2(1/2)^{k(k-1)/2} \qquad i = 1, 2, \ldots, \binom{n}{k}$$

  - 若存在 k 个顶点, 其对应的 $k(k-1)/2$ 条边是同一种颜色的事件可表示为 $\bigcup_{i=1}^{\binom{n}{k}} E_i$.

根据布尔不等式有

$$P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq \sum_{i=1}^{\binom{n}{k}} P(E_i) = \binom{n}{k}(1/2)^{k(k-1)/2-1}$$

当 $n \geq 10$ 和 $k > n/2$ 时有 $P\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq 1$, 因此事件“完全图中任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色”的概率大于零. 这意味着至少存在一种染色方法使得任意 k 个顶点相对应的 $k(k-1)/2$ 条边不是同一颜色.

隐私问题的调查 0.61

例 0.61 每个人都有一些隐私或秘密, 相关信息不希望被外人知晓. 对于具有社会普遍性的隐私问题, 需要对相关问题进行一些必要的调查. 需要设计一种调查方案, 使被调查者既愿意作出真实回答、又较好地保护个人隐私. 经过多年研究与实践, 心理学家和统计学家设计了一种巧妙的方案. 核心是如下两个问题:

[问题 A]: 你的生日是否在 7 月 1 日之前? [问题 B]: 你是否有抑郁倾向?

同时再准备一个箱子, 里面装有 m 个白球和 n 个红球, 被调查者随机抽取一球, 若抽中白色回答问题 A, 否则回答问题 B. 无论抽中哪个问题都只需回答“是”或“否”, 并将答案放入一个密封箱中 (假设在保护隐私的情况下, 学生诚实回答问题). 上述过程在一无人的房间内进行, 以保障被调查者的隐私.

若有 $N(N \geq 500)$ 位学生参加调查, 请估计出具有抑郁倾向的学生比例.

解答: 例 0.61

问题: 如上所述.

解答:

- 事件 “答卷选择 ‘是’ ” 的 “原因” 有两种情况, 即事件 “学生抽到红球时, 选择 ‘是’ ” 或者事件 “学生抽到白球时, 选择 ‘是’ ” .
- 设有 N_y 张答卷选择 “是” , 一个学生有抑郁倾向的概率为 p ; 不妨假设一个学生的生日在 7 月 1 日之前的概率为 $1/2$, 根据全概率公式有

$$\begin{aligned} &P(\text{一个学生回答 ‘是’}) \\ =&P(\text{一个学生回答 ‘是’}|\text{ 红球})P(\text{红球}) + P(\text{一个学生回答 ‘是’}|\text{ 白球})P(\text{白球}) \end{aligned}$$

由此可得

$$\frac{N_y}{N} \approx \frac{m}{m+n} \times \frac{1}{2} + \frac{n}{m+n} \times p$$

进一步估计出具有抑郁倾向的学生比例 $p \approx (m+n)N_y/nN - m/2n$.

生育问题 0.62

例 0.62 关于生育的新闻

- 民间机构“育娲人口研究智库”官方微信号本周发布一篇该智库专家团队署名文章，呼吁大力发放生育补贴。这篇文章称，当前中国面临内需不振和生育率低迷两大难题。大力发放生育补贴，短期刺激消费、扩大内需，长期提振生育率、提升社会活力，一举多得。
- 中国新出生人口自 2017 年以来连续六年下滑，在 2022 年首次跌破千万，降至 956 万。中国人口进入负增长时代。
- 由任泽平、梁建章等 5 位人口经济专家联署的这篇文章写道，“当前中国经济，孩子是最好的投资”。目前基建投资趋于饱和、制造业产能过剩、房地产供给总量平衡、教育出现过度竞争，但是在孩子数量上的投资是不够的。
- 撰文的专家认为，要提高生育率，就需要减轻育龄家庭的生育成本。

而发放育儿补贴，是降低生育成本最主要、最直接的方式。近年来，为了鼓励生育，中国不少省份和城市推出生育或育儿补贴的相关政策，但似乎收效甚微，而且很多政策并未落实。

现有补贴消息“部分资金还会向两孩或两孩以上家庭，提供每孩每月约 800 元的补贴 (不包括第一个孩子)；向三孩家庭提供每孩每月 1600 元的补贴（不包括第二个、第三个孩子)”。试：建模该案例，并分析该政策的合理性。

解答: 例 0.62

首先进行建模，即事件 A_i 表示某个家庭生第 i 胎的概率，$i = 0, 1, 2, 3$. 注意，这里 A_3 笼统记录大于等于 3 孩的. 则有

$$P(A_0) + P(A_1) + P(A_2) + P(A_3) = 1$$

- 所有先验概率都可以通过统计计数近似得到，比如根据卫健委《2022 年我国卫生健康事业发展统计公报》2022 年我国出生人口 956 万人，其中 1 孩有 440 万，2 孩有 363 万，3 孩有 143 万。所以可以大致计算得到

  $$P(A_1 \mid A_1A_2A_3) = 46.1\% , \quad P(A_2 \mid A_1A_2A_3) = 38.9\% , \quad P(A_3 \mid A_1A_2A_3) = 15.0\% .$$

- 我们另外得知，生一孩者才有可能生二孩，则根据全概率公式有

  $$P(A_{k+1}) = P(A_k)P(A_{k+1} \mid A_k) + P(\bar{A}_k)P(A_{k+1} \mid \bar{A}_k) , \quad k = 0, 1, 2$$

  根据不生一孩者不可能生二孩，有

  $$P(A_{k+1}) = P(A_k)P(A_{k+1} \mid A_k) , \quad k = 0, 1, 2$$

- 展开上述“传递”公式

$$P(A_1) = P(A_0)P(A_1 \mid A_0)$$
$$P(A_2) = P(A_1)P(A_2 \mid A_1) = P(A_0)P(A_1 \mid A_0)P(A_2 \mid A_1)$$
$$P(A_3) = P(A_2)P(A_3 \mid A_2) = P(A_0)P(A_1 \mid A_0)P(A_2 \mid A_1)P(A_3 \mid A_2)$$

- 根据该式，我们会知道
  - 补贴消息“部分资金还会向两孩或两孩以上家庭，提供每孩每月约 800 元的补贴 (不包括第一个孩子)；向三孩家庭提供每孩每月 1600 元的补贴（不包括第二个、第三个孩子）”是在试图增加 $P(A_2 \mid A_1)$ 和 $P(A_3 \mid A_2)$. 而我们知道，只增加条件概率部分 $P(A_{k+1} \mid A_k)$ 是不够的，还要增加先验概率部分 $P(A_k)$.
  - 注意到 $P(A_0)$ 是不可能知道的，所以 $P(A_1)$，$P(A_2)$，和 $P(A_3)$ 会变成关于 $P(A_0)$ 的函数（连乘关系）. 因此只增加 $P(A_2 \mid A_1)$ 和 $P(A_3 \mid A_2)$ 也是不够的，仍需要增加 $P(A_1 \mid A_0)$. 而且根据上式来看，提升 $P(A_1 \mid A_0)$ 会同步提升 $P(A_1)$，$P(A_2)$ 和 $P(A_3)$. 因此，刺激方向不对.
  - 注意，该刺激能否或者何种程度提升生育率，不在此建模的考虑范围内.
  - 还有别的建模方式吗？

休谟问题 0.63

例 0.63 [休谟问题] 下一步会发生什么？此刻做什么会确保下一步成功的把握性更大？是人类社会学最核心的问题之一 [即永远逃逸的未来性才是哲学的根本问题]。休谟提出了两个最深刻的怀疑论问题：

1. 关于因果或预测未来
2. 事实是否推导价值

两个问题都触及人类知识和生活的基础，是最严重而难以逾越的难题，至今没有完美解法。到目前来讲，解决方式大致有两种 [from 赵汀阳，中国社会科学院大学哲学院教授]

- 在哲学层面的先验论，比如康德方案、胡塞尔理论等. 这里都需要一个先验或者先天结构 (a priori)，通过哲学理论建立的、捕捉到了永恒绝对或普遍必然或先验的原理.

- 先验是否可以通过经验化/数据化的方法来建立？或者休谟问题是否可以通过经验化/数据化的方法来验证？

休谟问题 0.63

休谟问题是否可以通过经验化/数据化的方法来验证？

- 事实是否可以推导价值。事实 A 是客观的，价值 B 存在一定的主观性，那么事实是否可以推导价值即计算问题

$$P(B) = P(A)P(B \mid A)$$

- 当前的人工智能技术可以视为一种经验化的方法，用来近似/估计 priori. 经验化的方法短板在于：不能保证在无穷经验中的普遍有效性，即要确保无误需要见过所有的 cases (可能世界) 才行。比如：自然语言里会不断出现来自生活的新奇词汇，于是语言有“先验义务”去解释所有奇怪词汇。假如某些词汇终究无法解释，就必须考虑两种可能性：（1）或者这种语言不是一种成熟语言，反思能力不足，无法将某些词汇转换为一组可识别的描述；（2）或者那些词汇本身不合理，根本无法转换为任何描述，也就无法被解释。再比如：时变环境。
- 估计贝叶斯方法论是当前包括在大模型在内的人工智能的核心技术，至今还没有发现比贝叶斯概率论更好的方法去选择一个相对可信的未来。把经验知识理解为

一个无穷开放的演化过程，即不断以新经验去修正旧知识的无穷迭代过程，按照贝叶斯的概念来说，就是不断以“后验概率”去修正“先验概率”的无穷过程。这种经验论方法的巧妙之处在于，既然人类没有遍历无穷多可能世界的全知能力，也没有对无穷经验普遍必然有效的先验知识，那么，以一种无穷性（认知的无穷迭代过程）去应对另一种无穷性（世界的无穷可能性）就是最优方法。

- 贝叶斯方法论所蕴含的哲学理解有几点：
  1. 经验知识的有效性在于此种知识对于未来的有效性，这里的有效性是该知识对预测未来是否有效. 换言之，只能说经验知识具有不同程度的有效性，但是“有效性”是一个在概率上可衡量的概念.
  2. 未来是无限开放的概念，意味着无穷多可能性.
  3. 建构经验知识是无限连续的过程，意味着关于世界的真理不可能有一个定论.
  4. 经验知识的无限修正过程表现为相关可能性的无限收敛过程，即由无穷多可能性不断收敛为逼近必然性的不可完成过程.
- 休谟理论证实/证伪的重要意义：
  1. 通过因果干预可以影响未来
  2. 人类价值可以用客观事实来衡量，且可以用自然科学方法研究.