全概率公式 (Law of Total Probability)

September 24, 2023

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

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

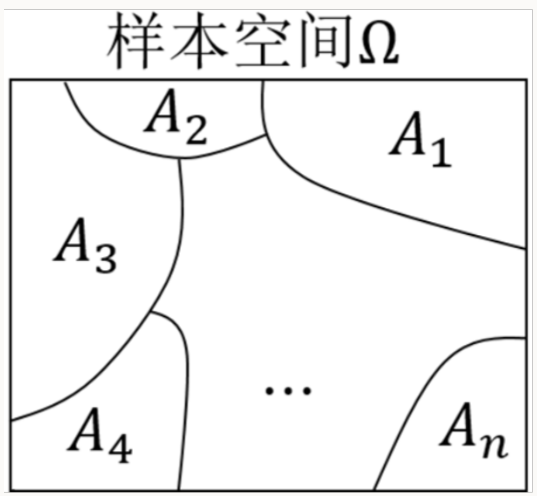
Remarks:

- 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.

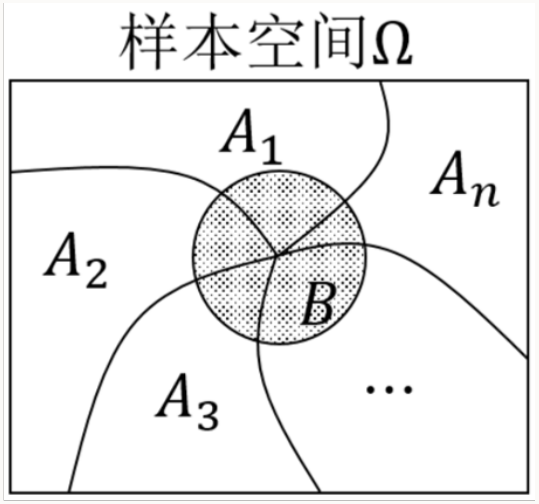
解读: 全概率公式

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 $\to P(B)$ 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

全概率公式: 例 0.36

例 0.36 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.

全概率公式: 例 0.37

例 0.37 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 $1/2$.

思考: 例 0.38

例 0.38 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.

贝叶斯公式 (Bayesian Formula)

September 24, 2023

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:

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

特别地,

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

且

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

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 $P(A_i \mid B)$.

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

贝叶斯公式：例 0.39

例 0.39 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.

例 0.36 vs 例 0.39

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,

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 $P(B \mid A_i)$.

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

贝叶斯公式的应用：例 0.40

例 0.40 (贝叶斯 Spam 过滤器) 如何确定一个电子邮件是 Spam?

- 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.
- 考察一个特定的单词 ω, 统计该单词在集合 B 和 G 中出现的次数分别为 $n_B(\omega)$ 和 $n_G(\omega)$.
- 设 S 是事件：邮件为 Spam, E 是事件：邮件内容含单词 ω. 需计算 $P(S \mid E)$.

思路: 根据贝叶斯公式, 我们需要分别估算

- Spam 邮件中含有单词 ω 的概率 $P(E \mid S)$
- 非 Spam 邮件中含有单词 ω 的概率 $P(E \mid \bar{S})$
- 比较这两者的大小

贝叶斯公式的应用：例 0.41

例 0.41 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”.

Appendix: 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

Appendix: 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 0.42

例 0.42 (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?