Probability and Statistics for Information Science
Lecture Notes

Hengyu Ai

Fall 2024

目录

1 概率与计数 Probability and Counting

1.1 概率模型 Probabilistic Model

定义 1.1 集合

1. 空集：$\emptyset$
2. 子集（subset）关系：$A \subseteq B$
3. 交（union）：$A \cup B$
4. 并（intersection）：$A \cap B$
5. 补（complement）：A^c
6. De Morgan 律：$(A \cup B)^c = A^c \cap B^c, (A \cap B)^c = A^c \cup B^c$

定义 1.2 样本空间（Sample Space）

一个实验的全部可能的结果的集合称为**样本空间**.

定义 1.3 事件（event）

样本空间的子集称为**事件**.
当实际结果属于一个事件时，称这个事件发生 (occured)

English	**Sets**
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S$, $A_i \cap A_j = \emptyset$ for $i \neq j$

1.2 概率与计数的朴素定义

1.2.1 概率

定义 1.4 概率（朴素定义）

假设：

1. 有限样本空间
2. 输出结果等可能

则：
令 A 为一个事件，S 为其样本空间，则 A 的概率为

$$P_{naive}(A) = \frac{|A|}{|P|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}$$

在这种情况下概率被转化为计数问题。

例 1.1 Pascal-Fermat Correspondence

Alice 和 Bob 抛硬币，三次获胜的人赢。每一轮 Alice 赌注硬币为正面 (Head)，Bob 赌注反面 (Tail)。当前比分为 $2:1$，若此时结束游戏该如何按可能的结果划分奖品？

在第五轮游戏必然结束，则剩下两局样本空间为：$S = \{HH, HT, TH, TT\}$，当且仅当投出 TT 时 Bob 获胜，则应该给 Alice $\frac{3}{4}$，给 Bob $\frac{1}{4}$

1.2.2 计数

- **Sampling**: 从集合等可能获取一个元素
- **With Replacement & without replacement**: 取出后放回/不放回
- **Ordered & Unorderd**

例 1.2 生日问题

房间内有 k 人，假设一个人生日为一年 365 天内等可能随机一天，生日互相独立，求至少两人有相同生日的概率。

相当于从集合 $\{1,2,3,\ldots,365\}$ 取出 k 个元素，with replacement。令 A 表示“存在两人以上⋯⋯“的情况，则 A^c 为”没有任何人和别人生日相同“的情况。

$$P(A^c)=\frac{|A^c|}{|S|}=\frac{\text{without replacement 的 samples}}{\text{with replacement}}=\frac{365!}{365^k}(k\leqslant 365)$$

$$P(A)=1-P(A^c)$$

定理 1.1　广义生日问题

从 n 个值中随机选 k 次，当 $k\approx 1.18\sqrt{n}$ 时，有 50% 的概率，至少有两个选取的值相同。

考虑按每个随机变量来计算 $P(A^c)$, 则第一个变量从 $[1,n]$ 选择任意，第二变量必须取到和第一个不同的, 第三个取和一、二不同的, 并且变量之间相互独立, 以此类推:

$$P(A^c)=\frac{n}{n}\times\frac{n-1}{n}\times\frac{n-2}{n}\cdots\times\frac{n-(k-1)}{n}=\prod_{i=1}^{k}\frac{n-i}{n}$$

当 $n\gg k$ 时，$\dfrac{n-i}{n}=1-\dfrac{i}{n}\approx \mathrm{e}^{-\frac{i}{n}}$，于是 $\prod\frac{n-i}{n}=\mathrm{e}^{-\frac{1+2+\cdots+k-1}{n}}=\mathrm{e}^{-\frac{k(k-1)}{2n}}\approx\mathrm{e}^{-\frac{k^2}{2n}}$
于是 $P(A)\approx 1-\mathrm{e}^{-\frac{k^2}{2n}}$
代入 $P(A)=0.5$ 得出 $k=\sqrt{n\cdot 2\ln 2}\approx 1.18\sqrt{n}$

定理 1.2　多项式定理（Multinomial Theorem）

$$(x_1+x_2+x_3+\cdots+x_r)^n=\sum_{n_1,n_2,\ldots,n_r\geqslant 0}\frac{n!}{n_1!n_2!\cdots n_r!}x_1^{n_1}x_2^{n_2}\cdots x_r^{n_r}(n_1+n_2+\cdots+n_r=n)$$

组合意义：把 n 个不同的人分成 r 组

在证明组合恒等式时，可以使用式子两侧组合意义说明等式成立。

定理 1.3　吸收/提取恒等式

$$n\binom{n-1}{k-1}=k\binom{n}{k}$$

证明

假设有如下场景：从 n 人中选出 k 个组成小队，并且这 k 人中有一人为小队长。

定理 1.3 LHS 可以视为先选出小队长（n 种方案），然后在剩下的 $n-1$ 人中选出小队剩余成员（组合数）. RHS 可以视为先从 n 人中选出组成小队的 k 人（组合数），然后在小队内部选出队长（k 种方案），可知等式成立。 □

定理 1.4 Vandermonde 卷积

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

证明

假设有如下场景：有 m 个男生和 n 个女生，从中选出 k 个人作为一组。

定理 1.4 LHS 即为直接选择，RHS 为，枚举从男生中选出几个人（j 从 0 到 k），然后选出 $\binom{m}{j}$，再从女生中选出剩下的人（$\binom{n}{k-j}$） □

定理 1.5 Bose-Einstein 计数定理

组合意义：可重复选择，顺序无影响，从 n 中选 k

形式描述：$x_1 + x_2 + \cdots + x_n = k, x_i \in \mathbb{N}, i = 0, 1, 2, \ldots, n$，这个不定方程的解的个数为

$$\binom{n+k-1}{n-1}$$

类似的，当解要求必须为正时，可以按照**隔板法**求解

例 1.3

求如下方程解的个数：

$$x_1 + x_2 + x_3 + x_4 = 88, x_1 \geq 3, x_2 \geq 5, x_3 \geq 8, x_4 \geq 10$$

令 $y_1 = x_1 - 3, y_2 = x_2 - 5, y_3 = x_3 - 8, y_4 = x_4 - 10$，转化为 定理 1.5

选择 n 个对象中的 k 个对象，可能的方式数量：

	顺序有关	顺序无关
允许放回	n^k	$\binom{n+k-1}{k}$
不允许放回	$n(n-1)\cdots(n-k+1)$	$\binom{n}{k}$

1.3 其他非公理化的概率定义

1.3.1 无限样本空间中的概率

按照测度理论的几何度量定义。

例 1.4

随机变量 $x \in [0,3]$, 求 x 到 0 比到 1 更近的概率.

一维的几何度量为线段长度：

$$P(A) = \frac{\text{Length of } [0,0.5]}{\text{Length of } [0,3]}$$

例 1.5

随机变量 $x \in [1,2]$, 求 $P(x=1.5)$

将 $[1,2]$ 分为 $[1,1+\frac{1}{2n+1}),[1+\frac{1}{2n+1},1+\frac{2}{2n+1}),\cdots$
可知对任何一个区间，$P(x \in [l,r)) = \dfrac{1}{2n+1}$

$$\forall n \geq 1, 0 \geq P(x=1.5) \leq P(x \in A_n) = \frac{1}{2n+1}$$

令 $n \to \infty$, 可知 $P(x=1.5)=0$

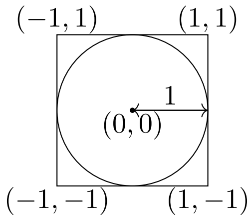
在上面的例子中，出现了概率为 0 的事件，但是那个事件并不是不可能时间。

一个不可能事件概率为 0，概率为 0 的事件为**事件测度为 0**，不代表事件不会发生。

更高维度也有相似的测度定义的概率：

例 1.6

在一个正方形中均匀任取一点，求这个点在正方形内接圆的概率。



二维的测度使用面积衡量。

$$P(A) = \frac{S_{\text{circle}}}{S_{\text{square}}} = \frac{\pi}{4}$$

1.3.2 统计定义的概率

概率是重复一个实验很多次后得到的频率。

大数定理，Monte Carlo 方法

1.4 公理化的概率定义

定义 1.5 公理化的概率

给定样本空间 S，构成事件类的 S 子集满足如下公理

1. S 是一个事件
2. 对于任何事件 A，其补集是事件
3. 可列的数个事件的并是事件

概率函数 $P: A \to \mathbb{R}$ 满足如下公理

1. $P(\emptyset) = 0, P(S) = 1$
2. 对于互不相交的事件 $A_1, A_2, \dots$

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$$

根据上述概率的公理，可以导出一些概率的性质：

- $P(A^c) = 1 - P(A)$
- 若 $A \subseteq B$，则 $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

定理 1.6 容斥原理

对于任何 n 个事件 $A_1, A_2, \dots, A_n$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

例 1.7 De Montmort 匹配问题

从均匀洗牌的 n 张卡的卡堆内，卡按照 $1 \sim n$ 标号，逐张翻开卡牌，若有任何一次满足翻开第 k 张时，那张牌标号为 k 则获胜。

令 A_i 表示第 i 张开的牌标号为 i 的事件，则 $P(\text{win}) = P(A_1 \cap A_2 \cap \cdots \cap A_n)$

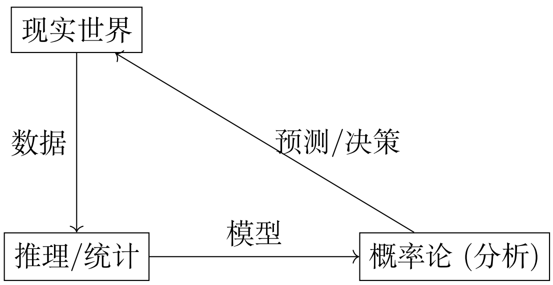
$P(A_i) = \dfrac{1}{n}$

根据对称性，$P(A_i \cup A_j \cup A_k \cdots \cup A_p)$ 对于相同的 p 都相同

$$P(A_i \cup A_j \cup A_k \cdots \cup A_p) =$$

用于分析具有不确定结果的现象的框架:

- 保持一致推理的规则
- 用于预测和决策



2 条件概率

内容总结

3

随机变量与分布

内容总结

4 期望

内容总结

5

连续随机变量

内容总结

6 联合分布

内容总结

7

Monte Carlo 方法

内容总结

8 条件期望

内容总结

9

统计推断

内容总结

10

Markov 链

内容总结