Introduction

定义（α -代数、可测集与可测空间）

- 设 E 是一个基本事件空间，考虑其幂集 2^E，并让 $\mathfrak{F}$ 是一个 Ω 的子集集合。$\mathfrak{F}$ 中的元素被称为随机事件（random events）。
- 如果 $\mathfrak{F}$ 满足以下性质，它被称为一个 σ-代数：
  1. $E \in \mathfrak{F}$（基本事件空间 E 属于 $\mathfrak{F}$）。
  2. 如果 $A, B \in \mathfrak{F}$，则 $A - B \in \mathfrak{F}$（差集仍在 $\mathfrak{F}$ 中）。
  3. 如果 $A_1, A_2, \ldots \in \mathfrak{F}$，则

$$\bigcup_{i=1}^{\infty} A_i \in \mathfrak{F} \tag{1}$$

且

$$\bigcap_{i=1}^{\infty} A_i \in \mathfrak{F} \tag{2}$$

（可数并集和交集仍在 $\mathfrak{F}$ 中）。
- 这意味着空集 $\emptyset$ 也属于 $\mathfrak{F}$。如果 E 是可数的，则 2^E 是一个 σ-代数。
- 如果 $\mathfrak{F}$ 是一个 σ-代数，其元素被称为可测集，$(E, \mathfrak{F})$ 被称为可测空间（或 Borel 空间）。

解释：σ-代数、可测集与可测空间

σ-代数（sigma-algebra）、可测集和可测空间是概率论中的基础概念，用来帮助我们定义和处理随机事件。想象一下，你有一个事件空间（比如抛硬币可能出现正面或反面），σ-代数就像是一个“规则集”，告诉你哪些事件集合是可以被合理测量（比如计算概率）的。

1. **事件空间和集合**：
   - 我们有一个基本的事件空间 E，里面包含所有可能的事件（比如抛硬币的结果：正面、反面）。
   - 我们可以从这些事件中挑选一些子集（比如“正面”或“反面”），这些子集就是“随机事件”。
2. **σ-代数的规则**：
   - σ-代数 $\mathfrak{F}$ 是一个集合的集合，它必须满足以下简单规则：
     - 整个事件空间 E 必须在 $\mathfrak{F}$ 里。
     - 如果两个事件 A 和 B 在 $\mathfrak{F}$ 里，那么它们的差集（比如 A 减去 B）也必须在 $\mathfrak{F}$ 里。
     - 如果有无数个事件 $A_1, A_2, A_3, \ldots$ 都在 $\mathfrak{F}$ 里，那么它们的可数并集（所有事件的集合）以及可数交集（所有事件都满足的集合）也必须在 $\mathfrak{F}$ 里。
   - 这些规则确保 $\mathfrak{F}$ 足够“完整”，包括空集 $\emptyset$（什么也没有的事件）。
3. **可测集和可测空间**：

- 如果一个集合属于 σ-代数 $\mathfrak{F}$，它就被称为“可测集”，意思是可以被赋予概率（比如“正面出现”的概率是 0.5）。
- 事件空间 E 和它的 σ-代数 $\mathfrak{F}$ 一起被称为“可测空间”或“Borel 空间”，这是概率论的基础框架。

简单来说，σ-代数就像一个“合法事件清单”，确保我们能合理地定义和计算随机事件的概率。

测度与概率测度

这些概念是构建概率空间的基础，帮我们理解如何给事件分配概率。

1. **测度的定义**：
   - 假设我们有一个可测空间 $(E, \mathfrak{F})$（之前提到的σ-代数和事件空间的组合）。
   - 一个“测度” P 是一个非负的函数，它从可测集 $\mathfrak{F}$ 映射到非负实数 $[0, +\infty]$，并且满足以下规则：
     - 空集 $\emptyset$ 的测度为 0，也就是 $P(\emptyset) = 0$。
     - 对于任何一组互不相交（两两没有重叠）的可测集 $A_1, A_2, A_3, \ldots$（比如抛硬币的正面和反面完全分开），它们的总测度等于每个集合测度的和：

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)。 \tag{3}$$

       这种性质叫做“可数可加性”或“σ-可加性”。
2. **概率测度的定义**：
   - 如果一个测度 P 满足 $P(E) = 1$（整个事件空间 E 的测度为 1），那么它被称为“概率测度”。
   - 对于概率测度来说，$P(E) = 1$ 是必须的，因为概率总和应该等于 1（比如一个确定会发生的事件概率是 100%）。
   - 满足这些条件的 $(E, \mathfrak{F}, P)$ 被称为“概率空间”，这是概率论的核心框架。

简单来说，测度就像一个“测量工具”，告诉我们一个事件集合的大小（或概率），而概率测度特别适用于概率问题，确保所有可能事件的总概率是 1 。

总和法则（Sum Rule）

“总和法则”（Sum Rule）是计算概率的重要工具。它帮助我们理解事件之间的关系，尤其是当事件互为补集或与其他事件结合时。

1. **事件和补集的关系**：
   - 假设 A 是一个事件，它的补集是“非 A”（记为 $\neg A$）。整个事件空间 E 可以被分成 A 和 $\neg A$ 的组合，也就是 $A + \neg A = E$（这里“+”表示并集）。
   - 因为 E 是所有可能事件的总和，它的概率 $P(E) = 1$（因为某件事一定会发生）。
   - 因此，A 和 $\neg A$ 的概率总和是 1：

$$P(A) + P(\neg A) = P(E) = 1, \tag{4}$$

     所以 $P(A) = 1 - P(\neg A)$。这告诉我们，一个事件的概率可以通过它的补集概率计算出来。
2. **总和法则（Sum Rule）**：

- 现在考虑事件 A 和另一个事件 B。我们可以用 A 和 B 的交集（记为 $A \cap B$）来表示 A。
- A 可以被分成两部分：A 与 B 的交集（$A \cap B$）和 A 与 B 的补集（$\neg B$）的交集（$A \cap \neg B$）。
- 因此，A 的概率可以表示为：

$$P(A) = P(A, B) + P(A, \neg B), \tag{5}$$

  其中 $P(A, B)$ 是 A 和 B 同时发生的联合概率（也就是 $P(A \cap B)$）。
- 这个公式被称为“总和法则”，它帮助我们根据 A 与 B 的关系计算 A 的概率。

简单来说，总和法则让我们可以用一个事件与其他事件的联合概率来计算它的总概率，尤其是当我们知道某些事件的发生与否时。这在概率计算和机器学习中非常有用。

条件概率（Conditional Probability）

“条件概率”（Conditional Probability）是概率论中一个关键概念，帮助我们理解在已知某个事件发生的情况下，另一个事件发生的可能性。

1. **条件概率的定义**：
   - 假设事件 A 的概率 $P(A) > 0$（也就是说，A 不是不可能发生）。
   - 条件概率 $P(B|A)$ 是“在 A 发生的情况下，B 发生的概率”，它的计算公式是：

$$P(B|A) = \frac{P(A, B)}{P(A)}, \tag{6}$$

     其中 $P(A, B)$ 是 A 和 B 同时发生的联合概率（也就是 $P(A \cap B)$），$P(A)$ 是事件 A 的概率。
   - 这个公式告诉我们，条件概率是通过 A 和 B 的联合概率除以 A 的概率得出的。
2. **条件概率的性质**：
   - 我们可以很容易证明，条件概率 $P(B|A) \geq 0$（概率不会是负数）。
   - 在已知 A 的情况下，整个事件空间 E 的条件概率 $P(E|A) = 1$（因为 A 发生时，某件事一定会发生）。
   - 如果两个事件 B 和 C 互不相交（$B \cap C = \emptyset$），那么它们的条件概率之和是：

$$P(B + C|A) = P(B|A) + P(C|A)。 \tag{7}$$

   - 因此，对于固定的 A，$(E, \mathfrak{F}, P(\cdot|A))$ 形成一个概率空间。这意味着条件概率满足概率的基本规则。
3. **联合概率和条件概率的关系**：
   - 从条件概率的定义可以直接得出联合概率的公式：

$$P(A, B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)。 \tag{8}$$

   - 这表明，A 和 B 同时发生的概率可以通过任意一个条件概率乘以对应的边缘概率计算。

简单来说，条件概率让我们可以在已知某些信息（比如 A 已经发生）的情况下，计算其他事件（比如 B）的发生概率。

多事件条件下的条件概率公式

基本定义

在概率论中，对于多个条件事件 $A_1, A_2, \ldots, A_n$ 和事件 B，条件概率 $P(B|A_1, A_2, \ldots, A_n)$ 表示在已知事件 $A_1, A_2, \ldots, A_n$ 发生的情况下，事件 B 发生的概率。其定义为：

$$P(B|A_1, A_2, \ldots, A_n) = \frac{P(B, A_1, A_2, \ldots, A_n|\text{其他条件})}{P(A_1, A_2, \ldots, A_n|\text{其他条件})} \tag{9}$$

解释：这个公式表示 B 在给定多个条件 $A_1, A_2, \ldots, A_n$ 发生时的概率。分子和分母都是在某些其他条件下的联合条件概率，用于计算 B 在特定条件下发生的概率。

具体实例

假设我们有三个变量 B、E 和 A，想计算 $P(B = 0|E = 1, A = 1)$。根据条件概率的定义，可以写成：

$$P(B = 0|E = 1, A = 1) = \frac{P(B = 0, E = 1|A = 1)}{P(E = 1|A = 1)} \tag{10}$$

解释：在这个例子中，$B = 0$ 是目标事件，$E = 1$ 和 $A = 1$ 是条件。分子 $P(B = 0, E = 1|A = 1)$ 是 $B = 0$ 和 $E = 1$ 在 $A = 1$ 条件下的联合条件概率；分母 $P(E = 1|A = 1)$ 是 $E = 1$ 在 $A = 1$ 条件下的条件概率。这个公式通过条件概率的比值计算 $B = 0$ 在 $E = 1$ 和 $A = 1$ 条件下的概率。

注意事项

- 为了使公式定义有效，需要确保分母 $P(A_1, A_2, \ldots, A_n|\text{其他条件}) > 0$，否则条件概率未定义。
- 如果多个事件之间存在独立性或条件独立性（例如 $A_i \perp A_j|C$），公式可能简化为更简单的形式，但需要根据具体问题确定。

解释：分母不能为 0 是概率计算的基本要求。如果某些事件在特定条件下独立，联合条件概率可以分解为乘积，从而简化计算，但这取决于具体分布。

推广应用

这个公式可以扩展到更多变量和条件。例如，对于 n 个条件事件 $C_1, C_2, \ldots, C_n$ 和目标事件 D，在给定其他条件 X 下的条件概率为：

$$P(D|C_1, C_2, \ldots, C_n, X) = \frac{P(D, C_1, C_2, \ldots, C_n|X)}{P(C_1, C_2, \ldots, C_n|X)} \tag{11}$$

解释：这个推广形式适用于复杂场景，如机器学习中的贝叶斯网络或因果推理，允许在多个条件和额外条件 X 的约束下估计某个事件 D 发生的概率。

全概率公式（Law of Total Probability）

“全概率公式”（Law of Total Probability）帮助我们在已知一些互不相交的事件覆盖了所有可能情况时，计算某个事件发生的总概率。

1. **全概率公式的定义**：
    - 假设有几个事件 $A_1, A_2, \ldots, A_n$，它们互不相交（也就是说，任意两个事件 A_i 和 A_j 如果 $i \neq j$，那么 $A_i \cap A_j = \emptyset$），并且它们的并集等于整个事件空间 E（也就是 $A_1 + A_2 + \cdots + A_n = E$）。

- 对于任意一个事件 X（属于可测集 $\mathfrak{F}$），它的概率 $P(X)$ 可以用以下公式计算：

$$P(X) = \sum_{i=1}^{n} P(X|A_i) \cdot P(A_i), \tag{12}$$

其中 $P(X|A_i)$ 是“在 A_i 发生的情况下，X 发生的条件概率”，$P(A_i)$ 是事件 A_i 的概率。

2. **证明思路**：
   - 因为 X 属于整个事件空间 E，并且 E 可以被分成 $A_1, A_2, \ldots, A_n$ 的并集，所以 X 也可以表示为这些事件与 X 的交集的并集：

$$X = E \cap X = \left(\bigcup_{i=1}^{n} A_i \right) \cap X = \bigcup_{i=1}^{n} (A_i \cap X)。 \tag{13}$$

   - 根据概率的可加性（因为 A_i 互不相交），X 的概率等于每个 $A_i \cap X$ 概率的总和：

$$P(X) = \sum_{i=1}^{n} P(A_i \cap X)。 \tag{14}$$

   - 利用条件概率定义 $P(A_i \cap X) = P(X|A_i) \cdot P(A_i)$，我们得到：

$$P(X) = \sum_{i=1}^{n} P(X|A_i) \cdot P(A_i)。 \tag{15}$$

   - 这就是全概率公式的证明。

简单来说，全概率公式让我们通过已知一些互不相交、覆盖所有可能性的事件（比如天气状况：晴天、雨天、雪天），结合这些事件发生的概率和条件概率，计算某个事件（比如下雨）的总概率。这在决策分析和预测中非常有用。

贝叶斯定理（Bayes' Theorem）

“贝叶斯定理”（Bayes' Theorem）帮助我们在已知某个事件发生的情况下，推导出另一个事件发生的反向概率（即从结果推导原因）。

1. **贝叶斯定理的定义**：
   - 假设有几个互不相交的事件 $A_1, A_2, \ldots, A_n$，它们覆盖了整个事件空间 E（也就是说，$A_1 + A_2 + \cdots + A_n = E$，且如果 $i \neq j$，则 $A_i \cap A_j = \emptyset$）。
   - 对于任意一个事件 X（属于可测集 $\mathfrak{F}$），某个特定事件 A_i 在 X 发生情况下的条件概率（即“A_i 是 X 的原因”的概率）可以用以下公式计算：

$$P(A_i|X) = \frac{P(A_i) \cdot P(X|A_i)}{\sum_{j=1}^{n} P(A_j) \cdot P(X|A_j)}。 \tag{16}$$

   - 这里的分子 $P(A_i) \cdot P(X|A_i)$ 是 A_i 发生的概率乘以在 A_i 发生时 X 发生的条件概率，分母是所有可能事件 A_j 的类似概率之和，用来“标准化”结果。

2. **证明思路**：
   - 贝叶斯定理的证明基于条件概率的定义和总和法则（Sum Rule）。

- 根据条件概率定义，$P(A_i|X)$ 是“在 X 发生时 A_i 发生的概率”，可以用联合概率 $P(A_i, X)$ 除以 $P(X)$：

$$P(A_i|X) = \frac{P(A_i, X)}{P(X)}。 \tag{17}$$

- 联合概率 $P(A_i, X) = P(A_i) \cdot P(X|A_i)$（因为 $P(A_i, X)$ 等于 A_i 的概率乘以在 A_i 发生时 X 的条件概率）。
- 事件 X 的总概率 $P(X)$ 可以通过全概率公式计算：

$$P(X) = \sum_{j=1}^{n} P(A_j) \cdot P(X|A_j)。 \tag{18}$$

- 代入上述公式，得到贝叶斯定理的表达式。

简单来说，贝叶斯定理让我们从观察到的结果（比如症状）推导出可能的原因（比如疾病），通过结合原因的先验概率（$P(A_i)$）和结果的条件概率（$P(X|A_i)$）。这在医学诊断、垃圾邮件过滤和机器学习中非常重要。

贝叶斯定理的广义含义

关于概率论中贝叶斯定理（Bayes' Theorem）的应用，具体介绍当我们观察到新数据时如何用贝叶斯定理更新对假设的信念。有一个更通用的贝叶斯公式形式，特别适用于机器学习和数据分析。

1. **贝叶斯定理的通用形式**：
   - 假设 X 是一个假设或模型（比如一个分类器或疾病的可能原因），D 是一些观察到的数据（比如测试结果或症状）。
   - 我们想计算“在观察到数据 D 后，假设 X 为真的后验概率” $P(X|D)$。这个后验概率可以用以下公式计算：

$$P(X|D) = \frac{P(X) \cdot P(D|X)}{P(D)}, \tag{19}$$

     其中：
     - $P(X)$ 是 X 的先验概率（在看到数据之前对 X 的初始信念）。
     - $P(D|X)$ 是似然（likelihood），即在 X 为真时数据 D 发生的概率。它并不是 X 的概率，而是关于 X 的函数。
     - $P(D)$ 是证据（evidence for the model），即数据 D 在所有可能假设下的总概率，用来“标准化”结果：

$$P(D) = \sum_{X \in \mathcal{X}} P(D|X) \cdot P(X), \tag{20}$$

       其中 $\mathcal{X}$ 是所有可能假设 X 的集合。

2. **贝叶斯定理的含义**：
   - 这个公式将后验概率 $P(X|D)$ 分解为三个部分：先验概率 $P(X)$、似然 $P(D|X)$ 和证据 $P(D)$。
   - 先验概率 $P(X)$ 反映我们对假设 X 的初始信念（比如某种疾病在人群中的普遍性）。
   - 似然 $P(D|X)$ 反映数据 D 在假设 X 下的可能性（比如在某种疾病下观察到特定症状的概率）。

- 证据 $P(D)$ 确保后验概率总和为 1，是所有可能假设下数据 D 发生的加权平均概率。

3. **应用场景**：
   - 这个公式用于“在观察数据 D 时更新对假设 X 的信念”。比如，在医学中，我们可以用患者症状（数据 D）和疾病的先验概率（$P(X)$）来更新对特定疾病（假设 X）的诊断概率。
   - 似然 $P(D|X)$ 不一定是我们在看到数据之前就知道的概率，而是基于所有可能数据在假设 X 下的分布：

$$P(X, d) = \sum_{d \in \mathcal{D}} P(X, d) \text{ under all possible data.} \tag{21}$$

   - 这在机器学习中非常重要，比如在贝叶斯分类或贝叶斯网络中更新模型。

简单来说，贝叶斯定理让我们从新观察的数据中更新对假设的信念，通过结合先验知识和数据证据。这在科学推理、医学诊断和机器学习中广泛应用。