机器学习讲义(L12)：概率论复习

授课教师：王贝伦 / 助教：张嘉琦，黄旭，谈笑，徐浩卿

1 概率论与机器学习

事件的概率是衡量该事件发生的可能性的量度。虽然在一次随机试验中某个事件的发生是带有偶然性的，但那些可在相同条件下大量重复的随机试验却往往呈现出明显的数量规律。
不确定性和随机性可能来自多个方面，使用概率论来量化不确定性。 概率论在机器学习中扮演着一个核心角色，因为机器学习算法的设计通常依赖于对数据的概率假设。

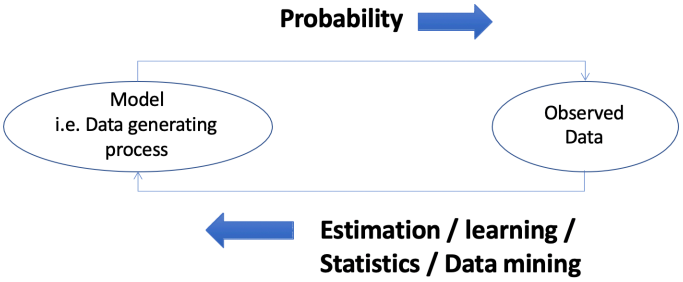

Figure 1: 概率论与机器学习

如图所示，可以说机器学习的核心就是探讨如何从数据中提取人们需要的信息或规律，我们通过构建的模型对数据进行预测。

2 事件与事件空间（Events and Event space）

2.1 事件与基本事件

定义：基本事件

在试验中可直接观察到的、最基本的不能再分解的结果称为基本事件。

例 2.1. 抛掷一个硬币，我们以“H”表示硬币正面朝上，“T”表示硬币反面朝上，两个硬币最后朝上的面会产生两种结果，构成两项基本事件：$\{H\},\{T\}$。

定义：事件

试验中部分结果的集合/样本空间的子集被称为事件。

例 2.2. 抛掷一个硬币，我们以“H”表示硬币正面朝上，“T”表示硬币反面朝上，一个硬币最后朝上的面会产生四项事件：$\emptyset,\{H\},\{T\},\{H,T\}$。

两者联系：基本事件（也称为原子事件或简单事件）是一个仅在样本空间中单个结果的事件，不会出现事件中的空集，更不会出现多元集，而事件则包含对所有结果的组合，甚至包括空集和样本空间。

2.2 样本空间与事件空间

定义：样本空间

试验中所有可能结果组成的集合被称为样本空间。

我们这里用字母 O 表示样本空间。

例 2.3. 抛掷一个硬币，我们以“H”表示硬币正面朝上，“T”表示硬币反面朝上，一个硬币最后朝上的面会产生两种结果，这个试验的样本空间，可以表示为：$O=\{H,T\}$。

定义：事件空间

实验中所有可能事件的集合被称为事件空间。

我们这里用字母 S 表示事件空间。

例 2.4. 抛掷一个硬币，我们以“*H*”表示硬币正面朝上，“*T*”表示硬币反面朝上，一个硬币最后朝上的面会产生四种事件，这个试验的是事件空间，可以表示为：$S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$。

两者联系：样本空间包含着所有的基本事件，所有的基本事件的发生概率相加为 1，而事件空间则包含着所有的事件，事件之间不一定相互独立，所有事件的发生概率相加为 2^{n-1}（n 表示基本事件个数）。

3 常用公理

定理 3.1. 对于事件空间中的任意事件 α，其概率大小必定遵守不等式：

$$0 \leq P(\alpha) \leq 1$$

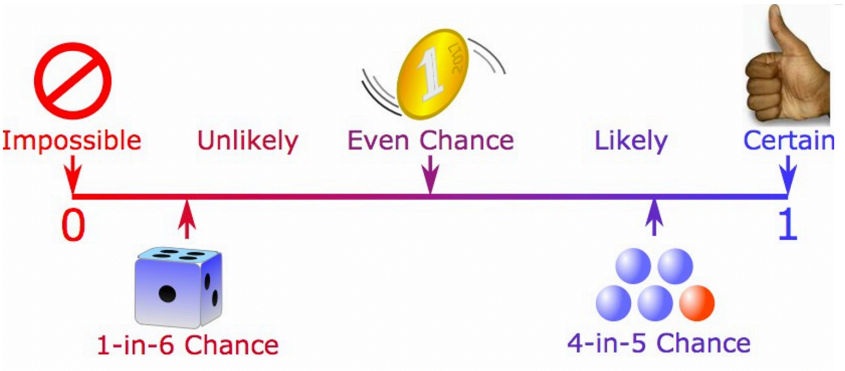

Figure 2: 概率的范围

定理 3.2. 如果事件 *A*，*B* 是互斥，$A \cup B$ 表示 A 或 B，即事件 *A* 与事件 *B* 的并集，那么

$$P(A \cup B) = P(A) + P(B)$$

推论 3.2.1. 对于一个样本空间来说，我们用 B_i 表示基本事件，因为基本事件之间是互斥的，故样本空间的概率可以表示为所有基本事件的和：

$$P(O) = \sum_i P(B_i) = 1$$

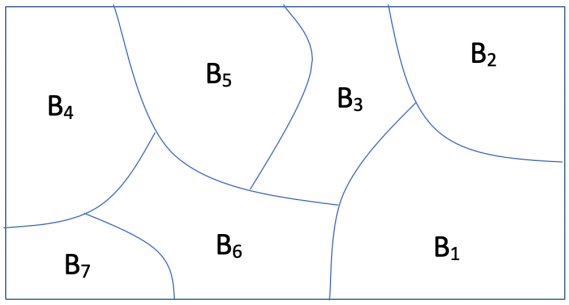

Figure 3: 样本空间的概率

定理 3.3. 概率中的容斥原理：有两个不同事件，分别用 A 和 B 表示，$A \cap B$ 表示 A 且 B，即事件 *A* 与事件 *B* 的交集，有等式：

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

当然我们也可以用 *or* 和 *and* 的逻辑符表达该式：

$$P(A\,or\,B) = P(A) + P(B) - P(A\,and\,B)$$

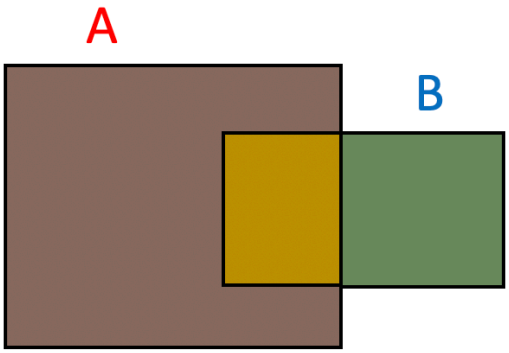

Figure 4: 概率中的容斥原理

推论 3.3.1. 对于定理 *3.3*，对等式两边的式子稍作变换，$\sim B$ 表示事件 B 的补集，已知

$$P(A\cap \sim B) = P(A \cup B) - P(B)$$

故有推论：

$$P(A) = P(A \cap B) + P(A\cap \sim B)$$

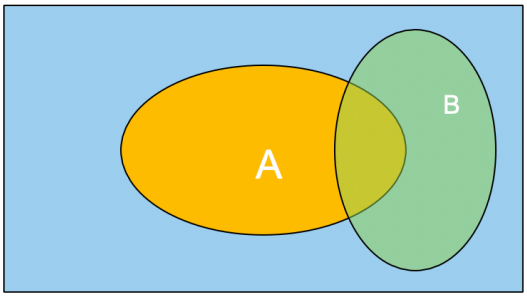

Figure 5: 概率中的容斥原理

推论 3.3.2. 对于定理 *3.3*，若 $B =\sim A$, 则 $P(A\cup \sim A) = 1$，则等式可以改写为：

$$P(not\ A) = P(\sim A) = 1 - P(A)$$

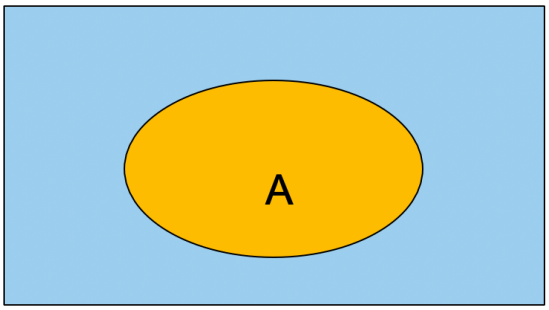

Figure 6: 概率中的容斥原理

4 随机变量

一个随机试验可产生多个结果，我们有时候想要了解满足某个属性的事件的概率是多少。例如样本空间 O 为全体学生，我们想要知道这些学生的学习情况，可以定义如下的事件：

- Grade_A= 所有等级为 A 的学生
- Grade_B= 所有等级为 B 的学生
- Hardworking_Yes= 所有勤奋的学生

- ……

这样子对于每一个属性的值都要定义一个事件，显得十分麻烦。我们需要一个函数，从样本空间 O 映射到属性空间 T。例如，定义 H 表示学生是否勤奋，那么 $H = \text{Yes}$ 表示勤奋的学生的集合。下面给出随机变量的定义

定义：随机变量（Random Variable）

样本空间 O 上的单值实值函数 $X = X(\omega)$，$\omega \in O$，其中 ω 为样本点，称为随机变量。

随机变量在不同的条件下由于偶然因素影响，可能取各种不同的值，故其具有不确定性和随机性，但这些取值落在某个范围内的概率是一定的。随机变量通常用大写英文字母或小写希腊字母来表示，从上面的定义注意到，随机变量的实质上是函数。

例 4.1. 随机掷两个骰子，整个样本空间由 *36* 个元素组成：

$$O = \{(i, j)|i = 1, \ldots, 6; j = 1, \ldots, 6\}$$

这里可以构成多个随机变量，比如随机变量 X 为两个骰子的点数和，随机变量 Y 为两个骰子的点数差。X 可以取 *11* 个整数值，而 Y 只能取 *6* 个。

$$X(i, j) = i + j, x = 2, 3, \ldots, 12$$

$$Y(i, j) = |i - j|, y = 0, 1, 2, 3, 4, 5$$

随机变量又可分为离散型随机变量和连续型随机变量。

定义：离散型随机变量

如果随机变量 X 的取值是有限的或者是可数无穷尽的值，则称 X 为离散型随机变量。

定义：连续性随机变量

如果随机变量 X 由全部实数或者一部分区间组成，则称 X 为连续型随机变量。

参数为 k 的离散型随机变量 X 可以取集合 $\{x_1, \ldots, x_k\}$ 中的一个值。离散型随机变量的概率质量函数（Probability Mass Function）为：$P(X = x_i)$。PMF 具有以下的性质：

- $\sum_i P(X = x_i) = 1$
- $P(X = x_i \cap X = x_j) = 0 (i \neq j)$
- $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j) (i \neq j)$
- $P(X = x_1 \cup X = x_2 \cup \cdots \cup X = x_k) = 1$

例 4.2. 投一枚硬币，只会产生两个结果，正面朝上或反面朝上，其中正面朝上的概率为 p。像这种只产生两种可能结果的试验称之为伯努利试验。如果投一枚硬币 k 次，随机变量 X 为正面朝上的次数，那么 X 为离散型随机变量，取值范围为 $\{1, 2, \ldots, k\}$，X 满足参数为 k 和 p 的二项分布。

离散型随机变量常见的分布有离散均匀分布和二项分布。离散均匀分布即随机变量 X 可以取 N 个值，取到每个值的概率是相同的，即 $P(X = i) = \frac{1}{N}$。例如投骰子，每个面朝上的概率都为 $\frac{1}{6}$。二项分布为重复 k 次独立的伯努利试验成功次数的分布，每次成功概率为 p 的话，$P(X = i) = \binom{k}{i} p^i (1 - p)^{k-i}$，记作 $X \sim B(k, p)$。例如投 k 次硬币，其中正面朝上的次数 $X \sim B(k, \frac{1}{2})$。当 $k = 1$ 时，又称为伯努利分布或 0-1 分布。

5 联合概率，边缘概率，条件概率

定义：联合概率

联合概率是指在多元概率分布中，多个随机变量分别满足各自条件的概率。

$P(A, B)$ 表示事件 A 和事件 B 同时发生的概率。

定义：边缘概率

边缘概率指在概率论和统计学的多维随机变量中，只包含其中部分变量的概率。

例 5.1. 图7表示某随机变量 X 和 Y 的联合概率分布和边缘分布。边缘概率即单独考虑 X 或 Y 的概率。

Y \ X	x_1	x_2	x_3	x_4	$p_Y(y)\downarrow$
y_1	4/32	2/32	1/32	1/32	**8/32**
y_2	3/32	6/32	3/32	3/32	**15/32**
y_3	9/32	0	0	0	**9/32**
$p_X(x)\rightarrow$	**16/32**	**8/32**	**4/32**	**4/32**	32/32

X，Y的联合概率 $P(X=x_i,Y=y_j)$

Y的边缘概率$P(Y=y_j)$

X的边缘概率$P(X=x_i)$

Figure 7: 联合概率与边缘概率

定义：条件概率

条件概率指事件 A 在事件 B 发生的条件下发生的概率，记作 $P(A|B)$。

条件概率有时也称为后验概率。条件概率的定义式为：

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

下面介绍这几个概率常用计算方法：

- 联合概率

  联合概率可以用链式法则来计算，链式法则与条件概率的定义式是一致的

$$\begin{aligned} P(A,B)&=P(B|A)P(A)\\ &=P(A|B)P(B)\end{aligned} \tag{2}$$

- 边缘概率

  边缘概率则可以用全概率公式来计算：若 $\sum_i P(B_i)=1$ 且 B_i 之间两两互斥，有

$$P(A)=\sum_i P(B_i)P(A|B_i) \tag{3}$$

图8很好的表示了全概率公式的意义。事件 A 与不同的 B 事件有重合，$P(B_i)P(A|B_i)$ 就表示 A 与 B_i 同时发生的概率，相当于图中 A 与 B_i 相交的部分。把 A 与每个 B_i 相交的部分都相加，便可以得到 A 的概率。

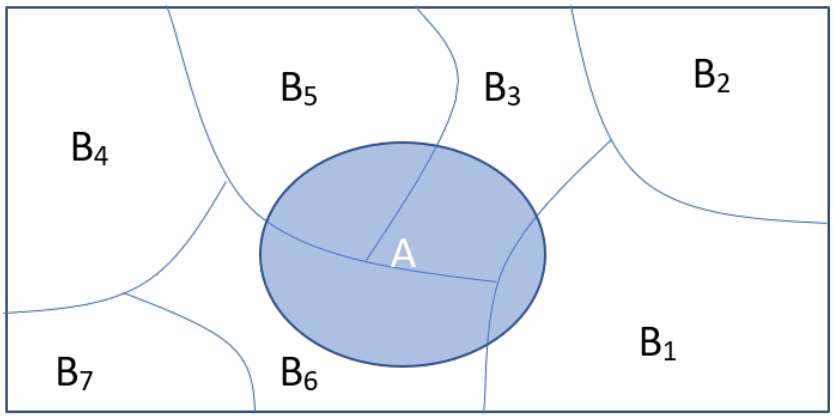

Figure 8: 全概率公式示意图。

又如图7所示的例子中，$P(X=x_i)$ 的概率可以表示为

$$\begin{aligned} P(X=x_i)&=\sum_j P(X=x_i|Y=y_j)P(Y=y_j)\\ &=\sum_j P(X=x_i,Y=y_j)\end{aligned} \tag{4}$$

- 条件概率

  条件概率除了使用定义式以外，还可以使用贝叶斯定理

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{5}$$

贝叶斯定理不难使用条件概率定义和链式法则推出，且定理中的分母常可以用全概率公式来表示。

6 独立性

6.1 随机事件的独立性

在有些情况下，两个随机事件 A, B 各自发生的概率与彼此是否发生无关，举例来说，第一次掷硬币出现正面，与第二次掷硬币是否出现正面在常识中是无关的。对于一般的随机事件，可以引入下列独立的概念：

定义：随机事件的独立性

设 A, B 是两个随机事件，若

$$P(AB) = P(A)P(B) \tag{6}$$

则称 A, B 是相互独立的。

由条件概率公式容易得到以下定理：

定理 6.1. 设 A, B 是两个概率非 0 的随机事件，则 A, B 相互独立的充要条件是下式之一成立：

$$P(B|A) = P(B),\ P(A|B) = P(A) \tag{7}$$

容易将两个事件的独立性推广到 n 个随机事件：

定义：多随机事件的独立性

对 n 个随机事件 $A_1, A_2, \cdots, A_n$，称它们是相互独立的当且仅当其中任意 $k(2 \le k \le n)$ 个事件满足

$$P(A_{i_1}A_{i_2}\cdots A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k}) \tag{8}$$

例 6.1. 一个袋子中装有红、黄、蓝等不同颜色小球若干个，如果我们有放回地从中每次摸出一个小球，记事件 A_i：第 i 次摸出的小球是红色，那么任意多个 A_i 都是相互独立的。

6.2 随机变量的独立性

利用随机事件的独立性，可以定义随机变量的独立性：

定义：随机变量的独立性

设随机变量 X 和 Y 的联合分布函数为 $F(x,y)$，$F_X(x)$,$F_Y(y)$ 分别为各自的分布函数。若对任意 $x, y \in \mathbb{R}$ 有

$$F(x,y) = F_X(x)F_Y(y) \tag{9}$$

则称 X,Y 相互独立，记为 $X \perp Y$。

对于离散和连续的两种情况，有以下定理：

定理 6.2. 设离散随机变量 X 和 Y 的联合分布律为

$$P(X = x_i, Y = y_j) = p_{ij},\ i, j = 1, 2, \cdots \tag{10}$$

则 X,Y 相互独立的充要条件是对一切的 x_i, y_i，

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \tag{11}$$

定理 6.3. 设连续随机变量 X 和 Y 的联合概率密度函数为 $f(x,y)$，$f_X(x)$,$f_Y(y)$ 分别为各自的边缘密度函数，则 X,Y 相互独立的充要条件是对一切的 x, y，

$$f(x,y) = f_X(x)f_Y(y) \tag{12}$$

对于6.1中随机事件的条件概率充要条件和多个随机事件的独立性等定义、定理，也可以类似地推广到随机变量中。

例 6.2. 同时掷一枚硬币和一颗骰子，以 $X = 0, 1$ 分别代表硬币正、反，$Y = 1, \cdots, 6$ 代表骰子的点数，则 (X, Y) 的联合分布律为

$$P(X = i, Y = j) = \frac{1}{12},\ i = 0, 1, j = 1, \cdots, 6 \tag{13}$$

显然，由常识可知

$$\begin{aligned} P(X = i) &= \frac{1}{2}, i = 0, 1 \\ P(Y = j) &= \frac{1}{6}, j = 1, \cdots, 6 \end{aligned} \tag{14}$$

容易验证，

$$P(X=i,Y=j)=\frac{1}{12}=\frac{1}{2}\times\frac{1}{6}=P(X=i)P(Y=j) \tag{15}$$

于是 X,Y 相互独立。而我们通过常识也可以判断，掷硬币和掷骰子是不相干的两次试验，它们不会影响彼此的结果，故这两个事件一定是独立的。

6.3 条件独立性

若 A,B 两个事件的独立性由第三个事件 C 决定，那么此时 A,B 就被称为条件独立。

定义：条件独立

设随机变量 X 和 Y 的联合分布函数为 $F(x,y)$，$F_X(x)$,$F_Y(y)$ 分别为各自的分布函数。若对任意 $x,y\in\mathbb{R}$ 并且对任意满足 $P(Z\leq z)>0$ 的 z 有

$$F(x,y|Z=z)=F_X(x|Z=z)F_Y(y|Z=z) \tag{16}$$

则称 X,Y 在 Z 的条件下相互独立，记为 $X\perp Y|Z$。

对6.1和6.2中的定理，也可以推广到条件独立性中来。需要注意的是，条件独立与独立无关，并非包含与被包含关系。也就是说，两变量条件独立并不意味着它们独立，反之亦然。

例 6.3. 有两袋小球，甲袋中装有 *1* 个红球和 *1* 个蓝球，乙袋中装有 *2* 个黄球，现任选一个袋子，从中先后拿出 *2* 个小球，记 A 为第一次拿出的小球颜色，B 为第二次拿出的小球颜色，C 为选取的袋子。若以 $X,Y=0,1,2$ 分别表示第一次、第二次取出红、蓝、黄球，以 $Z=1,2$ 表示甲、乙袋，可以写出在 $Z=2$ 条件下的联合分布律：

$$P(X=i,Y=j|Z=2)=\begin{cases}1, & i=j=2\\ 0, & \text{otherwise}\end{cases} \tag{17}$$

此时两随机变量 X,Y 各自的边缘分布律为：

$$P(X=i|Z=2)=P(Y=j|Z=2)=\begin{cases}1, & i \text{ or } j=2\\ 0, & \text{otherwise}\end{cases} \tag{18}$$

不难验证，

$$P(X=i,Y=j|Z=2)=P(X=i|Z=2)P(Y=j|Z=2) \tag{19}$$

这表明两次拿出的小球颜色在 $Z=2$ 的条件下是相互独立的。显然，乙袋中只有黄球，因此不论哪一次，取出的球必定是黄色，从而自然相互独立。

但若选取的是甲袋，情况就不一样了。容易验证：

$$P(X=0,Y=0|Z=1)=0 \tag{20}$$

但我们知道

$$\begin{aligned}P(X=0|Z=1)=0.5\\ P(Y=0|Z=1)=0.5\end{aligned} \tag{21}$$

因而

$$P(X=0,Y=0|Z=1)\neq P(X=0|Z=1)P(Y=0|Z=1) \tag{22}$$

即 X,Y 在 $Z=1$ 的条件下并不相互独立。从常识出发也容易理解，甲袋中有 *2* 个颜色不同的球，此时第一次取出的球之颜色决定了第二次取出的颜色，意味着 X 与 Y 存在相关性。

引用

[1]https://en.wikipedia.org/wiki/Joint_probability_distribution

[2]https://en.wikipedia.org/wiki/Marginal_distribution

[3]https://en.wikipedia.org/wiki/Conditional_probability

[4] https://github.com/scutan90/DeepLearning-500-questions