[AI2613 随机过程][第九讲] σ-代数与条件期望

张驰豪

最后更新：*2025* 年 *5* 月 *1* 日

目录

1 σ-代数与信息

我们今天从另外一视角来看 σ-代数，即看成信息的集合。为了说明这一点，我们回顾一下随机变量的定义。给定一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$，我们说函数 $X\colon \Omega \to \mathbb{R}$ 是一个随机变量，当且仅当 X 是一个可测函数，也就是说对于任何 $B \in \mathscr{B}(\mathbb{R})$，我们有 $X^{-1}(B) \in \mathscr{F}$。这个时候，我们也称 X 是 $\mathscr{F}$-可测的。同理，对于定义在 Ω 上的任意一个 σ-代数 $\mathscr{G}$ 和一个函数 $Y\colon \Omega \to \mathbb{R}$，我们说 Y 是 $\mathscr{G}$-可测的，当且仅当对于任何 $B \in \mathscr{B}(\mathbb{R})$，$Y^{-1}(B) \in \mathscr{G}$。

$\mathscr{B}(\mathbb{R})$ 是 $\mathbb{R}$ 上所有 Borel 集的集合

反过来，给定一个函数 $X\colon \Omega \to \mathbb{R}$，我们用 $\sigma(X)$ 表示使得 X 可测的最小的 σ-代数，容易验证，$\sigma(X)$ 总是存在的。直观上，对于**离散**的 X 我们可以把 $\sigma(X)$ 理解成 $\left\{X^{-1}(x) : x \in \mathsf{Im}(X)\right\}$ 所构成的 Ω 的分划所生成的 σ-代数。这个直观帮助我们理解这个概念很重要。实际上，对于一般的 X 我们有下面命题。

命题 1. $\sigma(X) = \left\{X^{-1}(B) : B \in \mathscr{B}(\mathbb{R})\right\}$.

我们可以自然地把定义推广到多个随机变量 $X_1, \ldots, X_n$ 上。我们用 $\sigma(X_1, \ldots, X_n)$ 表示使得 $(X_1, \ldots, X_n)$ 的联合分布可测的最小的 σ-代数。容易验证

$$\sigma(X_1, \ldots, X_n) = \sigma\left(\bigcup_{i\in[n]} \sigma(X_i)\right).$$

命题的验证很简单，首先根据定义 $\sigma(X)$ 必须包含 $\left\{X^{-1}(B) : B \in \mathscr{B}(\mathbb{R})\right\}$。其次，我们可以通过定义直接验证 $\left\{X^{-1}(B) : B \in \mathscr{B}(\mathbb{R})\right\}$ 本身是一个 σ-代数。

同样，如果是无穷多个随机变量 $\{X_\alpha : \alpha \in I\}$，那么 $\sigma(\{X_\alpha : \alpha \in I\}) \coloneqq \sigma\left(\bigcup_{\alpha\in I} \sigma(X_i)\right)$。

我们今天会有很多比较抽象的概念，因此，脑子里一直有下面这个运行例子是比较重要的。我们考虑投掷一个公平的六面骰子的概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$，其中 $\Omega = [6]$，$\mathscr{F} = 2^\Omega$，$\forall i \in \Omega, \mathbb{P}[\{i\}] = \frac{1}{6}$。我们定义四个随机变量：

- $X_1\colon i \in \Omega \mapsto i$，即 X_1 表示掷出来的点数；
- $X_2\colon i \in \Omega \mapsto \mathbb{I}_{[i\geq 4]}$，即 X_2 表示掷出来的点数是“大”还是“小”；
- $X_3\colon i \in \Omega \mapsto i \mod 2$，即 X_3 表示掷出来的点数除 2 之后的余数；
- $X_4\colon i \in \Omega \mapsto i \mod 4$，即 X_4 表示掷出来的点数除 4 之后的余数。

我们可以分别计算 $\sigma(X_i)$。由于 X_i 是离散的随机变量，我们只需要给出分划 $\left\{X^{-1}(x) : x \in \mathsf{Im}(X)\right\}$ 就可以了。回忆到我们之前介绍过，对于一个集族 $\mathscr{A} \subseteq 2^\Omega$，$\sigma(\mathscr{A})$ 为包含 $\mathscr{A}$ 的最小 σ-代数。于是，稍作思索可以得到

- $\mathscr{F}_1 = \sigma(X_1) = \sigma(\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\})$；
- $\mathscr{F}_2 = \sigma(X_2) = \sigma(\{\{1, 2, 3\}, \{4, 5, 6\}\})$；

- $\mathscr{F}_3 = \sigma(X_3) = \sigma(\{\{1,3,5\},\{2,4,6\}\})$；
- $\mathscr{F}_4 = \sigma(X_4) = \sigma(\{\{4\},\{1,5\},\{2,6\},\{3\}\})$。

回忆我们说一个函数 $f\colon \mathbb{R} \to \mathbb{R}$ 是 Borel 的，当且仅当 f 是 $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ 可测的。下面命题可以说明，为什么我们把 σ-代数称为信息的集合。

命题 2. *随机变量 Y 是 $\sigma(X)$-可测的当且仅当存在一个 Borel f 使得 $Y = f(X)$。*

这个命题想说明这样一件事情：一个随机变量 Y 是另一个随机变量 X 生成的 σ-代数可测，意味着如果知道了 X 的取值，那么 Y 的取值也就知道。换句话说，X 包含了 Y 的所有信息，这等价于 $\sigma(Y) \subseteq \sigma(X)$。也就是说，如果我想知道随机变量 Y 的取值，我并不需要知道随机试验得到了哪个样本点 $\omega \in \Omega$，而只需知道随机试验得到的样本点在 X 上的取值即可。

命题的证明见我的概率论讲义。

我们用前面的例子来检查一下这个结论，希望大家能够仔细弄清楚。

- 首先，由于 $\mathscr{F}_1 = \mathscr{F}$，因此 X_1, X_2, X_3, X_4 均是 $\mathscr{F}_1$-可测的。这是很显然的，因为 $X_1(i) = i$ 就返回随机实验得到的样本点本身，$\mathscr{F}_1$ 包含了“投一个公平六面骰子”的全部信息。
- X_3 是 $\mathscr{F}_4$ 可测的。这个从 $\mathscr{F}_3$ 和 $\mathscr{F}_4$ 的定义上可以看出来，但直观上它想说的事情是，“如果我们知道一个数除 4 的余数，那自然也就知道其除 2 的余数”。因此，X_3 可以写成 X_4 的函数（$X_3 = X_4 \mod 2$）。但是反过来就不对，因为我们知道一个数除 2 的余数，并不能够得到其除 4 的余数，$\sigma(X_3)$ 包含的信息严格少于 $\sigma(X_4)$。
- $\mathscr{F}_2$ 和 $\mathscr{F}_3$ 是不能够比较的，因此，X_2 和 X_3 互相不能写成对方的函数。因为，知道一个数是否大于等于 4 不能确定其除 2 的余数，反之亦然。

2 条件期望的定义

我们接着来引入概率论里面的一个核心概念：条件期望，这是我们在未来继续学习随机过程的时候必不可少的语言。在今天的讨论里，我们还是固定一个概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$。给定事件 $A, B \in \mathscr{F}$，在 $\mathbb{P}[B] > 0$ 的时候，我们定义过条件概率 $\mathbb{P}[A \mid B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$。我们也定义过给定事件 B 之后随机变量 X 的条件概率：$\mathbf{E}[X \mid B] := \frac{\mathbf{E}[X \cdot \mathbb{1}[B]]}{\mathbb{P}[B]}$。对于两个随机变量 X 和 Y，我们今天的目标是定义记号 $\mathbf{E}[Y \mid X]$。在我们今天所有的讨论中，均假设 X 和 Y 是可积的。

随机变量 X 可积（integrable）指的是 $\mathbf{E}[|X|] < \infty$。

2.1 X 是离散随机变量的场合

我们首先假设 X 是离散的随机变量，即 X 的取值 $\mathrm{Im}(X) = \{x_1, x_2, \dots\}$。对于每一个 x_i，我们知道 $[X = x_i]$ 是一个概率非零的事件，因此按照我们上面的定义

$$\mathbf{E}[Y \mid X = x_i] = \frac{\mathbf{E}[Y \cdot \mathbb{1}[[X = x_i]]]}{\mathbb{P}[X = x_i]}.$$

显然这是一个关于 x_i 的函数，换句话说，我们可以找到一个 Borel 函数 $f\colon \mathbb{R} \to \mathbb{R}$ 满足

$$f\colon x_i \mapsto \mathbf{E}[Y \mid X = x_i].$$

于是，我们定义 $\mathbf{E}[Y \mid X] := f(X)$。换句话说，$\mathbf{E}[Y \mid X]$ 是一个随机变量，满足

$$\mathbf{E}[Y \mid X] : \omega \in \Omega \mapsto f(X(\omega)).$$

我们应该这样看待这个定义：X 定义了样本空间的一个分划 $\Omega = \bigsqcup_{n\geq 1} \Lambda_n$，其中 $\Lambda_n = X^{-1}(x_n)$。对于每一个 $\omega \in \Omega$，如果其属于 Λ_k，则 $\mathbf{E}[Y \mid X](\omega)$ 的值为 Y 在 Λ_k 上的条件期望，即 $\mathbf{E}[Y \mid \Lambda_k]$。这是理解条件期望以及它的相关性质的最重要的直观。

这里 $A \sqcup B$ 是非交并的意思，它只对 $A \cap B = \varnothing$ 的集合定义。

继续考虑投掷一个公平的六面骰子的概率空间 $(\Omega, \mathscr{F}, \mathbb{P})$，其中 $\Omega = [6]$，$\mathscr{F} = 2^{\Omega}$, $\forall i \in \Omega, \mathbb{P}[\{i\}] = \frac{1}{6}$。我们定义四个随机变量

- $X_1\colon i \in \Omega \mapsto i$，即 X_1 表示掷出来的点数；
- $X_2\colon i \in \Omega \mapsto \mathbb{1}[[i \geq 4]]$，即 X_2 表示掷出来的点数是“大”还是“小”；
- $X_3\colon i \in \Omega \mapsto i \mod 2$，即 X_3 表示掷出来的点数除 2 之后的余数；
- $X_4\colon i \in \Omega \mapsto i \mod 4$，即 X_4 表示掷出来的点数除 4 之后的余数。

那么我们有

- $\mathbf{E}[X_1 \mid X_1](i) = X_1(i)$;
- $\mathbf{E}[X_1 \mid X_2](i) = \begin{cases} 5 & \text{if } i \geq 4; \\ 2 & \text{if } i < 4. \end{cases}$
- $\mathbf{E}[X_3 \mid X_4](i) = X_3(i)$;
- $\mathbf{E}[X_4 \mid X_2](i) = \begin{cases} \frac{2+0+2}{3} & \text{if } i \text{ is even}; \\ \frac{1+3+1}{3} & \text{if } i \text{ is odd}. \end{cases}$

此外，我们还可以注意到一个事实，就是 $\mathbf{E}[Y \mid X]$ 的定义实际上只与 X 所定义出来的分划 $\Lambda_1, \Lambda_2, \dots$ 有关，而与 $x_1, x_2, \dots$ 的具体取值无关。换句话说，$\mathbf{E}[Y \mid X]$ 实际上只与 X 生成的 σ-代数 $\sigma(X)$ 有关。

2.2 X 与 Y 有联合密度函数的场合

设 $f_{XY}(x,y)$ 是 X 与 Y 的联合密度函数。在边缘密度函数 $f_X(x) \neq 0$ 的时候，我们之前定义过条件期望

$$\mathbf{E}[Y \mid X = x] = \frac{\int_{\mathbb{R}} y f_{XY}(x,y)\,\mathrm{d}y}{f_X(x)}.$$

可以看出，这也是一个关于 x 的函数。我们可以找到一个 Borel 函数 $f\colon \mathbb{R} \to \mathbb{R}$ 满足

$$f\colon x \mapsto \mathbf{E}[Y \mid X = x].$$

于是我们可以类似离散场合定义 $\mathbf{E}[Y \mid X] \coloneqq f(X)$。

2.3 一般随机变量的场合

对于一般的随机变量，合理的定义出 $\mathbf{E}[Y \mid X]$ 不是一件简单的事情。事实上，我们先抽象出前面两种特殊场合定义的条件期望满足的两个重要性质：

1. $\mathbf{E}[Y \mid X]$ 是 $\sigma(X)$-可测的；
2. 对于任何 $A \in \sigma(X)$，$\int_A Y\,\mathrm{d}\mathbb{P} = \int_A \mathbf{E}[Y \mid X]\,\mathrm{d}\mathbb{P}$。

因为我们知道存在一个 Borel f 满足 $\mathbf{E}[Y \mid X] = f(X)$，所以 $\mathbf{E}[Y \mid X]$ 是 $\sigma(X)$-可测是显然的。我们现在分别对于离散和具有联合分布函数的两种场合验证第二点。

离散随机变量 我们知道 $\Omega = \bigsqcup_{n\geq 1} \Lambda_n$。我们知道对于每一个 $A \in \sigma(X)$，都可以写成若干 Λ_k 的并，因此根据积分的可加性，我们只需要对 $A = \Lambda_k$ 证明即可。这个时候，我们知道根据定义，对于每一个 $\omega \in \Lambda_k$，$\mathbf{E}[Y \mid X](\omega)$ 的取值均为 $\mathbf{E}[Y \mid X = x_k]$，所以

$$\int_{\Lambda_k} \mathbf{E}[Y \mid X]\,\mathrm{d}\mathbb{P} = \mathbf{E}[Y \mid X = x_k] \cdot \mathbb{P}(\Lambda_k) = \mathbf{E}[Y \cdot \mathbb{1}[[X = x_k]]] = \int_{\Lambda_k} Y\,\mathrm{d}\mathbb{P}.$$

具有联合分布的随机变量 对于 $A \in \sigma(X)$，我们知道，存在 $B \in \mathcal{B}$ 使得 $A = X^{-1}(B)$。我们首先有

$$\int_A Y\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}}\int_{\mathbb{R}} y \cdot f_{XY}(x,y) \cdot \mathbb{1}[[x \in B]]\,\mathrm{d}x \otimes \mathrm{d}y.$$

另一方面

$$\begin{aligned}\int_A \mathbf{E}[Y \mid X]\,\mathrm{d}\mathbb{P} &= \int_B f_X(x)\cdot \mathbf{E}[Y \mid X=x]\,\mathrm{d}x \\ &= \int_B f_X(x)\cdot\left(\int_{\mathbb{R}} y\cdot f_{Y|X}(y|x)\,\mathrm{d}y\right)\mathrm{d}x \\ \text{(Fubini-Tonelli)}\qquad &= \int_{\mathbb{R}} y\cdot\left(\int_B f_{XY}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} y\cdot f_{XY}(x,y)\cdot \mathbb{1}[[x\in B]]\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

我们把上面两个性质当做条件期望的定义。我们更一般的给出一个随机变量 Y 在一个 σ-代数的条件下的条件期望定义。

定义 3 (条件期望). *设 $(\Omega,\mathcal{F},\mathbb{P})$ 是个概率空间，X 是一个定义在其上的可积的随机变量，$\mathcal{G}\subseteq\mathcal{F}$ 是一个子 σ-代数。我们说一个随机变量 $Z\colon\Omega\to\mathbb{R}$ 是给定 $\mathcal{G}$ 后 X 的条件期望，并记作 $Z=\mathbf{E}[X\mid\mathcal{G}]$，当且仅当其满足：*

1. *Z 是 $\mathcal{G}$-可测的；*
2. *对于每一个 $A\in\mathcal{G}$，$\int_A Z\,\mathrm{d}\mathbb{P}=\int_A X\,\mathrm{d}\mathbb{P}$。*

在这个定义的基础上，对于随机变量 X，我们定义 $\mathbf{E}[Y\mid X]:=\mathbf{E}[Y\mid\sigma(X)]$。

我们有的时候也使用“条件概率”$\mathbb{P}[A\mid\mathcal{G}]$ 的记号，它被定义为 $\mathbf{E}[\mathbb{1}[A]\mid\mathcal{G}]$。我们对于条件期望的定义比较抽象，它和我们之前遇到过的大多数数学对象都不一样，是通过“描述性质”的方法来定义的。所以我们必须说明其合理性。首先是“唯一性”：

命题 4. *如果 Z 和 Z' 都是满足上面两个条件的随机变量，那么 $Z=Z'$ a.e.*

证明. 根据定义的第二条，我们知道 Z 和 Z' 都是可积的。设 $A=\{\omega\in\Omega : Z(\omega)>Z'(\omega)\}$。于是

$$\int_A Z-Z'\,\mathrm{d}\mathbb{P}=\int_A Z\,\mathrm{d}\mathbb{P}-\int_A Z'\,\mathrm{d}\mathbb{P}=\int_A X\,\mathrm{d}\mathbb{P}-\int_A X\,\mathrm{d}\mathbb{P}=0.$$

即 $\mathbb{P}[Z>Z']=0$。同理 $\mathbb{P}[Z<Z']=0$。因此 $\mathbb{P}[Z=Z']=1$。 □

最后，我们要说明这样一个 Z 总是存在的。它是测度论里面的 *Radon-Nikodym 定理*的推论，它的证明超出了这门课的范围。感兴趣的读者可以参考概率论的教材。

3 条件期望的性质

我们现在讨论条件期望的性质。可以很容易验证，我们定义的期望，本身也是条件期望的一个特殊情况，即 $\mathcal{G}=\{\varnothing,\Omega\}$ 是最简单 σ-代数。

1. $\mathbf{E}[X] = \mathbf{E}[X \mid \{\varnothing, \Omega\}]$。

2. 如果 X 是 $\mathcal{G}$-可测的，那么 $\mathbf{E}[X \mid \mathcal{G}] = X$ a.e.

   证明. 根据条件期望的定义，这是显然的（X 是 $\mathcal{G}$-可测且对任意 $A \in \mathcal{G}$，

   $$\int_A X \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P}.$$

   □

3. 条件期望的一个核心的性质是所谓的“tower rule”。假设 $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{F}$，并且满足 $\mathcal{G}_1 \subseteq \mathcal{G}_2$。换句话说，$\mathcal{G}_1$ 是比 $\mathcal{G}_2$ 更粗的 σ-代数。那么

   $\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbf{E}[X \mid \mathcal{G}_1]$。

   也就是说，当条件期望复合出现的时候，最终剩下的总是更“粗”的 σ-代数。

   证明. $\mathbf{E}[X \mid \mathcal{G}_1]$ 是 $\mathcal{G}_1$-可测的，因此也是 $\mathcal{G}_2$-可测的。于是根据性质 (2)，

   $$\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}_1] \mathcal{G}_2] = \mathbf{E}[X \mid \mathcal{G}_1].$$

   另一方面，对于任意一个 $A \in \mathcal{G}_1$，我们知道其也 $\in \mathcal{G}_2$，

   $$\int_A \mathbf{E}[X \mid \mathcal{G}_2] \, \mathrm{d}\mathbb{P} = \int_A X \, \mathrm{d}\mathbb{P} = \int_A \mathbf{E}[X \mid \mathcal{G}_1] \, \mathrm{d}\mathbb{P}.$$

   又 $\mathbf{E}[X \mid \mathcal{G}_1]$ 是 $\mathcal{G}_1$-可测的，所以

   $$\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbf{E}[X \mid \mathcal{G}_1].$$

   □

4. $\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}]] = \mathbf{E}[X]$

   这个性质是 (1) 和 (3) 的简单推论，但是在很多概率的计算中非常有用。我们通常使用的方式是 $\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid Y]]$，它可以解读成“为了计算 X 的平均值，我们先按照 Y 分类，对 Y 的每种情况计算对应 X 的平均值，再对 Y 的取值求平均”。比如说，我们让 $(\Omega, \mathcal{F}, \mathbb{P})$ 为班上所有同学的均匀分布。X 为同学的身高，Y 为同学的性别，那么 $\mathbf{E}[X \mid Y]$ 就表示随机抽一个同学，和该同学同性别的同学的平均身高。而 $\mathbf{E}[\mathbf{E}[X \mid Y]] = \mathbf{E}[X]$ 的直观含义就是，先统计男生平均身高和女生平均身高，然后再按照男生女生人数的比例对这两个数平均，就得到了全班同学的身高。

5. 如果 X 是 $\mathcal{G}$-可测的，并且 XY 是可积的，那么 $\mathbf{E}[XY \mid \mathcal{G}] = X\mathbf{E}[Y \mid \mathcal{G}]$ a.e.

   这个性质和（2）一样告诉我们，在做计算的时候，如果 X 是 $\mathcal{G}$-可测的，说明它在“已知 $\mathcal{G}$”的信息下，它没有什么随机性，因此可以当成一个常数一样从期望里拿出来。

   对于离散的随机变量 X，可以使用定义简单的证明。对于一般的 X，我们可以通过对 $\underline{X}_k$ 的情况取极限得到。这个证明留做练习。

大量关于期望的性质都可以推广到条件期望，我们罗列如下。他们均可以通过定义简单证明。

(6) $\mathbf{E}[aX + bY \mid \mathcal{G}] = a\mathbf{E}[X \mid \mathcal{G}] + b\mathbf{E}[Y \mid \mathcal{G}]$ a.e.

(7) 如果 $X \geq 0$ a.e., 那么 $\mathbf{E}[X \mid \mathcal{G}] \geq 0$ a.e.

(8) $|\mathbf{E}[X \mid \mathcal{G}]| \leq \mathbf{E}[|X| \mid \mathcal{G}]$ a.e.

(9) 如果 X 和 $\mathcal{G}$ 独立，那么 $\mathbf{E}[X \mid \mathcal{G}] = \mathbf{E}[X]$。

(10) 如果 X_n 和 X 均可积，并且 $X_n \uparrow X$，那么 $\mathbf{E}[X_n \mid \mathcal{G}] \to \mathbf{E}[X \mid \mathcal{G}]$ a.e.

(11) 【琴生不等式】如果函数 ϕ 在定义域内是 convex 的，并且 $\phi(X)$ 是可积的。那么 $\phi(\mathbf{E}[X \mid \mathcal{G}]) \leq \mathbf{E}[\phi(X) \mid \mathcal{G}]$。

我们将在之后几次课大量使用这些性质进行计算。但在计算的时候，需要非常小心。试看下面一例：

示例 5. 假设我独立的投掷两个 6 面骰子，X 表示第一个的点数，Y 表示第二个的点数。那么 $\mathbf{E}[\mathbf{E}[X + Y \mid X] \mid Y]$ 是多少？

根据定义，我们知道 $\mathbf{E}[X + Y \mid X] = X + \mathbf{E}[Y \mid X] = X + 3.5$ 是一个 $\sigma(X)$ 可测的随机变量。于是

$$\mathbf{E}[X + 3.5 \mid Y] = 3.5 + \mathbf{E}[X] = 7.$$