AI2613 随机过程 2020–2021 春季学期

Lecture 1 – Review of Probability

2021 年 2 月 22 日

Lecturer: 张驰豪 *Scribe*: 杨宽

课程信息

课程大纲：

- 马尔可夫链 Markov Chains: discrete / continuous
- 泊松过程 Poisson Process
- 鞅 Martingale
- 布朗运动 Brownian Motion
- AI / 大数据 / 机器学习中的算法应用

参考资料：

- Richard Durret, *Essentials of Stochastic Processes*
- Sheldon M. Ross, *Introduction to Probability Models*
- http://www.stat.yale.edu/~pollard/Courses/251.spring2013/Handouts/Chang-notes.pdf

1 Probability and Random Variable

Definition 1 (Probability space). A *probability space* consists of a 3-ary tuple $(\Omega, \mathcal{F}, \mathbf{Pr}[\cdot])$:

- Ω is a set of "outcomes" (countable or uncountable);
- $\mathcal{F} \subseteq 2^{\Omega}$ is a σ-algebra (a set of all possible "events") on which we can define probability, and here we say $\mathcal{F}$ is a σ-algebra if $\mathcal{F}$ satisfies
  - $\emptyset \in \mathcal{F}$,
  - $A \in \mathcal{F} \Longrightarrow A^{\mathsf{c}} \in \mathcal{F}$,
  - a countable sequence of sets $A_1, \ldots, A_n, \ldots \in \mathcal{F} \Longrightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$;
- Probability $\mathbf{Pr}[\cdot] : \mathcal{F} \to [0,1]$ is a function s.t.

1. $\mathbf{Pr}[\emptyset] = 0$,
2. $\mathbf{Pr}[\Omega] = 1$,
3. if $A_1, \ldots, A_n, \ldots \in \mathcal{F}$ are disjoint, then $\mathbf{Pr}\left[\cup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbf{Pr}[A_i]$.

Example 2 (6-face dice). $\Omega = [6] = \{1,2,3,4,5,6\}$, $\mathcal{F} = 2^{[6]}$, $\mathbf{Pr}[i] = 1/6$.
Generally, in a *discrete space*, Ω is countable. Define $\mathcal{F} = 2^{\Omega}$ and $\hat{p}\colon \Omega \to [0,1]$ s.t. $\sum_{\omega\in\Omega} \hat{p}(\omega) = 1$, then

$$\forall\, A \in \mathcal{F}, \quad \mathbf{Pr}[A] \triangleq \sum_{\omega \in A} \hat{p}(\omega)\,.$$

Question. How to define a probability on $\mathbb{R}$?
Or, what do we mean by drawing a uniform real in $(0,1)$?

Example 3 (Uniform real in $(0,1)$). Define the probability of uniformly drawing real numbers as follows:

- $\Omega = (0,1)$;
- $\mathcal{F}$ is the σ-algebra consisting of all "Borel sets" on $(0,1)$, namely the collection of subsets of $(0,1)$ obtained from all open intervals by repeatedly taking *countable* unions and complements;
- $\forall$ interval $I = (a,b)$, $\mathbf{Pr}[I] = (b-a)$. (*Lebesgue measure*)

Remark. $\mathcal{F}$ is called the Borel algebra, which is the smallest σ-algebra containing all open intervals. All Borel sets are measurable. The existence of non-Borel set is independent of ZF (Zermelo-Fraenkel set theory).

Definition 4 (Random variable). A *random variable* is a function or mapping from the probability space to a field. Given $\left(\Omega, \mathcal{F}, \mathbf{Pr}[\cdot]\right)$, a real-valued random variable is a function of Ω:

$$X\colon \Omega \to \mathbb{R}.$$

Definition 5 (Distribution (discrete)). For a countable Ω and a random variable X, the distribution of X is given by

$$\forall a \in \mathsf{Range}(X), \qquad \mu(a) = \mathbf{Pr}[X = a] \triangleq \mathbf{Pr}\left[\{\omega : X(\omega) = a\}\right] = \mathbf{Pr}\left[X^{-1}(a)\right].$$

Example 6 (Binomial distribution). Toss a *biased* coin (Head with probability p and Tail with probability $1-p$) n times. Let X be the number of Heads, then the distribution of X is called the *binomial distribution*:

$$\Omega = \{0,1\}^n, \qquad X \sim \mathsf{Binom}(n,p) \iff \mathbf{Pr}[X = k] = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}.$$

Example 7 (Geometric distribution). Toss a biased coin. Let X be the number of trials until the first Head, then the distribution of X is called the *geometric distribution*:

$$\Omega = \{0,1\}^*, \qquad X \sim \text{Geometric}(n,p) \iff \mathbf{Pr}[X = k] = (1-p)^{k-1}p\,.$$

Definition 8 (Distribution (continuous)). For an uncountable Ω and a random $\mathrm{d}x$ variable X, if there exists a nonnegative function $f(x)$ s.t.

$$\mathbf{Pr}[a \le X \le b] = \int_a^b f(x)\,\mathrm{d}x\,,$$

then $f(x)$ is called the *probability density function* of X.
The function

$$F(x) = \mathbf{Pr}[X \le x] = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

is called the *cumulative distribution function* of X.

Example 9 (Uniform distribution on (a,b)). The probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b; \\ 0, & \text{otherwise}\,. \end{cases}$$

Example 10 (Exponential distribution). The probability density function of the exponential distribution with $\lambda > 0$ is defined as

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \ge 0; \\ 0, & \text{otherwise}\,. \end{cases}$$

Example 11 (Gaussian / Standard normal distribution). The probability density function is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,.$$

2 Expectation

Definition 12 (Expectation). Given a probability space $\left(\Omega, \mathcal{F}, \mathbf{Pr}[\cdot]\right)$ and a random variable X, the *expectation* of X is defined as:

$$\mathbb{E}[X] = \sum_a a \cdot \mathbf{Pr}[X = a] \qquad \text{(discrete)}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f(t)\,\mathrm{d}t \qquad \text{(continuous)}$$

Example 13 (Uniform distribution on (a,b)).

$$\mathbb{E}[X] = \int_a^b t \cdot \frac{1}{b-a}\,\mathrm{d}t = \frac{1}{b-a}\cdot\frac{t^2}{2}\Big|_a^b = \frac{b+a}{2}$$

Example 14 (Exponential distribution).

$$\mathbb{E}[X] = \int_0^\infty t\cdot\lambda \mathrm{e}^{-\lambda t}\,\mathrm{d}t = -\int_0^1 t\,\mathrm{d}\mathrm{e}^{-\lambda t} = -t\mathrm{e}^{-\lambda t}\Big|_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda t}\,\mathrm{d}t = -\frac{1}{\lambda}\cdot\mathrm{e}^{-\lambda t}\Big|_0^\infty = \frac{1}{\lambda}$$

Definition 15 (Variance). Given a probability space $\big(\Omega,\mathcal{F},\mathbf{Pr}[\cdot]\big)$ and a random variable X, the *variance* of X is defined as:

$$\mathsf{Var}[X] = \mathbb{E}\big[(X-\mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}[X]\big)^2.$$

Definition 16 (Independence). Given a probability space $\big(\Omega,\mathcal{F},\mathbf{Pr}[\cdot]\big)$ and two random variables X and Y, X and Y are *independent* ($X \perp Y$) if

$$\forall\, A,B\subseteq\mathbb{R}, \qquad \mathbf{Pr}[X\in A\wedge Y\in B] = \mathbf{Pr}[X\in A]\cdot\mathbf{Pr}[Y\in B]\,.$$

Proposition 1 (Linearity of expectation). *If $X_1,X_2,\ldots,X_n$ are n random variables (not necessarily independent), then*

$$\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i]\,.$$

Proposition 2. *If $X_1,X_2,\ldots,X_n$ are n "mutually independent" random variables, then*

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i]\,,$$
$$\mathsf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathsf{Var}[X_i]\,.$$

3 Conditional Probability and Conditional Expectation

In this section, assume the set of outcomes Ω is finite. Then case for infinite Ω is more subtle.

Definition 17 (Conditional probability). Given a probability space $\big(\Omega,\mathcal{F},\mathbf{Pr}[\cdot]\big)$, let A and B are two events, then the probability of A conditioned on B is

$$\mathbf{Pr}[A\mid B] \triangleq \frac{\mathbf{Pr}[A\cap B]}{\mathbf{Pr}[B]}.$$

Similarly, we would like to define conditional expectation. However, to formally deal with conditional expectation, we should introduce measurable function first.

Definition 18 (Measurable functions). Let $\mathcal{F}$ be a σ-algebra and X be a function. Then X is $\mathcal{F}$-*measurable* if

$$\forall\, a \in \mathsf{Range}(X), \qquad X^{-1}(a) \in \mathcal{F}\,.$$

Denote by $\sigma(X)$ the minimum σ-algebra $\hat{\mathcal{F}}$ such that X is $\hat{\mathcal{F}}$-measurable.

Now we can define conditional expectation. Note that we only define it in discrete cases here.

Definition 19 (Conditional expectation (discrete)). Suppose $X, Y : \Omega \to \mathbb{R}$ are two random variables, and $A \subseteq \Omega$ is a event. Then the expectation of X conditioned on A is

$$\mathbb{E}[X \mid A] = \sum_x x \cdot \mathbf{Pr}[X = x \mid A]\,.$$

Specifically, let $A = Y^{-1}(a)$ be the event that $Y = a$,

$$\mathbb{E}[X \mid Y = a] = \sum_b b \cdot \mathbf{Pr}[X = b \mid Y = a]\,.$$

Given Y, $f_Y = \mathbb{E}[X \mid Y]$ is a function (a random variable) on Ω which satisfies

$$\forall\, \omega \in \Omega, \qquad f_Y(\omega) = \mathbb{E}[X \mid Y = Y(\omega)]\,.$$

Proposition 3. *Conditional expectation has the following two important propositions:*

1. $\mathbb{E}[X \mid Y]$ *is* $\sigma(Y)$*-measurable;*
2. $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[f_Y] = \mathbb{E}[X]$.

Proof. Item 1 is trivial. We only prove item 2 here.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_y \mathbb{E}\big[X \mid Y = y\big] \cdot \mathbf{Pr}\big[Y = y\big] \\
&= \sum_x \sum_y x \cdot \mathbf{Pr}\big[X = x \mid Y = y\big] \cdot \mathbf{Pr}\big[Y = y\big] \\
&= \sum_x x \cdot \sum_y \mathbf{Pr}\big[X = x \mid Y = y\big] \cdot \mathbf{Pr}\big[Y = y\big] \\
&= \sum_x x \cdot \sum_y \mathbf{Pr}\big[X = x \wedge Y = y\big] \\
&= \sum_x x \cdot \mathbf{Pr}[X = x] \\
&= \mathbb{E}[x]\,.
\end{aligned}$$

□

Example 20. Consider the probability space $\left(\Omega, \mathcal{F}, \mathbf{Pr}[\cdot]\right)$ where Ω is the set of all Chinese people and $\mathbf{Pr}[\cdot]$ is the uniform probability.
Let X and Y be two random variables that X is the height of a person and Y is the gender of a person. Then $\mathbb{E}[X]$ is the average height of Chinese people.
Let $f_Y = \mathbb{E}[X \mid Y] : \Omega \to \mathbb{R}$ be the random variable that $f_Y(\omega)$ is the average height of people with the same gender as ω. Then $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$.