Information Theory: Probability Background

zqy1018

2020. 4. 6

Contents

1 Famous Inequalities In Probability Theory

In this section, we review some famous and also useful inequalities.

Theorem 1. (Markov's Inequality) For any nonnegative random variable X and any $t > 0$,

$$p(X \geq t) \leq \frac{EX}{t}$$

Proof.

$$tp(X \geq t) = \int_t^{+\infty} tp(X = x)\mathrm{dx} \leq \int_t^{+\infty} xp(X = x)\mathrm{dx} \leq \int_0^{+\infty} xp(X = x)\mathrm{dx} = EX$$

So $p(X \geq t) \leq \frac{EX}{t}$.

We can easily construct a random variable that satisfies the equality. For example, let $X = 1$ (i.e. $p(X = 1) = 1$). Then $p(X \geq 1) = \frac{EX}{1} = 1$. □

Theorem 2. (Chebyshev's Inequality) For any random variable Y with mean value μ and variance σ^2,

$$p(|Y - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof. Let $X = (Y - \mu)^2$. Then by Markov's inequality,

$$p(|Y - \mu| > \epsilon) = p(X > \epsilon^2) \leq \frac{EX}{\epsilon^2} = \frac{D(Y - \mu) + [E(Y - \mu)]^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

□

Note that we will use the Chebyshev's inequality to prove the weak law of large number.

2 Convergence of Random Variables

We define convergence on a sequence of random variables $X_1, X_2, \cdots, X_n, \cdots$. We usually use three different definitions of convergence.

2.1 Convergence In Probability

Definition. $X_1, X_2, \cdots$ converges to X **in probability** if

$$\forall \epsilon > 0, \lim_{n \to \infty} p(|X_n - X| > \epsilon) = 0$$

Or to write it in epsilon-N language

$$\forall \epsilon > 0, \forall \delta > 0, \exists N \in \mathbb{N}, \forall n > N, p(|X_n - X| > \epsilon) < \delta$$

Usually denoted as $X_n \xrightarrow{p} X$.

2.2 Convergence In Mean

Definition. $X_1, X_2, \cdots$ converges to X **in the p-th mean** (or **in the L^p-norm**) if

$$\lim_{n\to\infty} E\left(|X_n - X|^p\right) = 0, 1 \le p < +\infty$$

Or to write it in epsilon-N language

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \forall n > N, E\left(|X_n - X|^p\right) = 0 < \epsilon$$

Usually denoted as $X_n \xrightarrow{L^p} X$.

Note.

(1) We usually use $p = 2$, and $X_n \xrightarrow{L^2} X$ is also called X_n converges **in mean square**.

(2) $1 \le q < p < +\infty, X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{L^q} X$.

2.3 Convergence With Probability 1

Definition. $X_1, X_2, \cdots$ converges to X **with probability 1** (or **almost surely**) if

$$p\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Or more explicitly,

$$p\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1$$

Usually denoted as $X_n \xrightarrow{a.s.} X$.

2.4 Relationship

We can prove that the $X_1, X_2, \cdots$ converges to X either in mean or with probability 1 implies that it also converges in probability.

Theorem 3. $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$.

Theorem 4. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X$.

3 Law of Large Number

Definition. $X_1, X_2, \cdots$ are **i.i.d.** if they are independent of each other and obey the same distribution.

Note. $X_1, X_2, \cdots$ can be treated as a sequence or many random variables. It depends on the context.

Theorem 5. (Strong Law of Large Number, Strong LLN) For i.i.d. $X_1, X_2, \cdots$, let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$p\left(\lim_{n\to\infty} \overline{X}_n = E(X_1)\right) = 1$$

Or $\overline{X}_n \overset{a.s.}{\to} E(X_1)$.

Theorem 6. (Weak Law of Large Number, Weak LLN) For i.i.d. $X_1, X_2, \cdots$, let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\overline{X}_n \overset{p}{\to} EX_1$.

Proof. We assume that $DX_1 = \sigma^2$.

By Chebyshev's inequality, we have

$$p(|\overline{X}_n - EX_1| > \epsilon) \leq \frac{D\overline{X}_n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

where $D\overline{X}_n = \frac{\sum_{i=1}^{n} DX_i}{n^2} = \frac{\sigma^2}{n}$.

Then we take $n \to \infty$ and we have $\overline{X}_n \overset{p}{\to} EX_1$. □

Note. Sometimes if the random variables are not well-defined, then the strong one may not hold. For example, when EX_1 does not exist. However, they share the same premises. So if the strong one holds, then the weak one will hold.

4 Stochastic Process

Definition. A **(discrete) stochastic process** is an indexed sequence of random variables.

The random variables can be related to each other. For example, $X_{i+1} = X_i + 1$.

There are many different types of stochastic processes. Here we focus on the stationary process.

Definition. A stochastic process is **stationary** if the joint distribution of *any subset* of the sequence of random variables is *time-shift-invariant*. That is,

$$\forall n, t, p(X_{i_1} = x_1, X_{i_2} = x_2, \cdots, X_{i_n} = x_n) = p(X_{i_1+t} = x_1, X_{i_2+t} = x_2, \cdots, X_{i_n+t} = x_n)$$

Note. The random variables in a stationary stochastic process obeys the same distribution since $\forall x, p(X_1 = x) = p(X_2 = x) = \cdots$.

5 Markov Chain

5.1 Basic Definition

Definition. For random variables $X_1, X_2, \cdots, X_n$, where $n \geq 3$, $X_1 \to X_2 \to \cdots \to X_n$ form a **Markov chain** if

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1})$$

Note. It is easy to check that $X_1 \to X_2 \to \cdots \to X_n$ iff $X_n \to X_{n-1} \to \cdots \to X_1$. So sometimes we use another notation $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_n$ to represent this symmetry.

Actually, there is another equivalent definition, often seen in books about stochastic processes.

Definition. A discrete stochastic process $X_1, X_2, \cdots, X_n$ is said to be a **Markov chain** if

$$p(x_{n+1}|x_n, x_{n-1}, \cdots, x_1) = p(x_{n+1}|x_n)$$

We can see their equivalence by chain rule.

5.2 Basic Properties

Theorem 7. $X_1 \to X_2 \to \cdots \to X_n$ iff

$$\begin{aligned} &X_1 \to X_2 \to X_3 \\ &(X_1, X_2) \to X_3 \to X_4 \\ &\vdots \\ &(X_1, X_2, \cdots, X_{n-2}) \to X_{n-1} \to X_n \end{aligned}$$

Proof. By induction. □

Theorem 8. $X \to Y \to Z \iff X \perp Z|Y$, i.e. X and Z and X are conditionally independent given Y.

Proof. Notice that $X \to Y \to Z \iff p(x,y,z) = p(x)p(y|x)p(z|y) \iff p(x,z|y) = p(x|y)p(z|y)$. □

Corollary 1. $X \to Y \to Z \iff I(X;Z|Y) = 0$.

Corollary 2. If $Z = f(Y)$, then $X \to Y \to Z$.

5.3 Time-invariance And Transition Matrix

Definition. A Markov chain is **time-invariant** if $p(x_{n+1}|x_n)$ is independent of n. That is, $\forall n, p(X_{n+1} = a|X_n = b) = p(X_2 = a|X_1 = b)$.

Note. We assume that all X_i's are defined in the same alphabet.

For convenience, we usually represent $p(x_2|x_1)$ with a **transition matrix** P, where $P_{ij} = p(X_2 = x_j|X_1 = x_i)$. And sometimes $p(y|x)$ just denotes the transition matrix from X to Y.

Acknowledgment

The contents are mainly based on the course materials of CS258, Shanghai Jiao Tong University and Wikipedia.