

第三讲

1 Gaussian 尾界

集中不等式关心随机变量 X 偏离其期望 $\mathbb{E}(X)$ 的概率，各位熟悉 3σ 准则则是其中一种。

引理 1.1: 3σ 准则

设 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则

$$P\{|X - \mu| \leq 3\sigma\} \approx 0.9973.$$

下面我们将 3σ 准则推广至一般的情形。

引理 1.2: Gaussian 尾界

设 $X \sim \mathcal{N}(0, 1)$, 则对任意 $t > 0$

$$P\{X \geq t\} \leq \frac{1}{t\sqrt{2\pi}} \exp(-t^2/2).$$

证明. 已知

$$P\{X \geq t\} = \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx.$$

令 $x = t + y$, 其中 $y \geq 0$. 上式可以改写为

$$\begin{aligned} P\{X \geq t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\{-(t+y)^2/2\} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{t^2}{2}\right) \cdot \exp\left(-\frac{ty}{2}\right) \cdot \exp\left(-\frac{y^2}{2}\right) dy \\ &\leq \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{t^2}{2}\right) \cdot \exp\left(-\frac{ty}{2}\right) dy \\ &= \frac{1}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right). \end{aligned}$$

□

2 概率估计间隙的一个例子

在给出例子之前, 先回忆一些概率论的基本概念和结论.

定理 2.1: Lindeberg-Lévy 中心极限定理

设 X_1, X_2, \dots 为一组独立同分布的随机变量序列, 对应的分布期望为 μ , 方差为 σ^2 . 记它们的和为

$$S_N := X_1 + X_2 + \dots + X_N,$$

并将其标准化为均值为 0, 方差为 1 的随机变量

$$Z_n = \frac{S_N - \mathbb{E}S_N}{\sqrt{\text{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^N (X_i - \mu).$$

当 $N \rightarrow \infty$, Z_N 依概率收敛于标准正态分布 $\mathcal{N}(0, 1)$.

评论. 设 $X \sim \mathcal{N}(0, 1)$, 上述定理的依概率分布收敛意味着,

$$\lim_{N \rightarrow \infty} P\{S_n \geq t\} = P\{X \geq t\} \quad \text{对任意 } t \in \mathbb{R}.$$

若 Lindeberg-Lévy 中心极限定理中随机变量序列的元素服从伯努利分布 $Ber(p)$, 即

$$P\{X_i = 1\} = p, P\{X_i = 0\} = 1 - p,$$

则有 $\mathbb{E}X_i = p$, $\text{Var}(X_i) = p(1 - p)$.

其 N 项和 S_N 服从二项分布 $\text{Binom}(N, p)$, 有 $\mathbb{E}S_N = Np$, $\text{Var}(S_N) = Np(1 - p)$.

由 Lindeberg-Lévy 中心极限定理知, 当 $N \rightarrow \infty$

$$\frac{S_N - Np}{\sqrt{Np(1 - p)}} \text{ 依概率收敛于 } \mathcal{N}(0, 1).$$

中心极限定理只提供了一个收敛结局, 但没有关于其收敛快慢的分析. Berry-Esseen 中心极限定理作为其补充.

定理 2.2: Berry-Esseen 中心极限定理

在 Lindeberg-Lévy 中心极限定理的条件下, 对任意 $N \in \mathbb{N}$ 及任意 $t \in \mathbb{R}$,

$$|P\{Z_N \geq t\} - P\{X \geq t\}| \leq \frac{\rho}{\sqrt{N}},$$

其中 $\rho = \mathbb{E}|X_1 - \mu|^3 / \sigma^3$, $X \sim \mathcal{N}(0, 1)$.

概率估计. 投掷一枚均匀硬币 N , 至少有 $\frac{3}{4}N$ 次头像向上的概率是多少?

设投掷该硬币 N 次, 头像朝上的总次数为随机变量 $S_N \sim \text{Binom}(N, \frac{1}{2})$,

$$\mathbb{E}S_N = \frac{N}{2}, \quad \text{Var}(S_N) = \frac{N}{4}.$$

由 Chebyshev 不等式,

$$P\{S_N \geq \frac{3}{4}N\} \leq P\{|S_N - \frac{N}{2}| \geq \frac{N}{4}\} \leq \frac{(N/4)}{(N/4)^2} = \frac{4}{N}.$$

此概率上界线性依赖于样本容量 N : N 增加一个尺度, 概率上界下降一个尺度.

另一方面, 我们从中心极限定理出发估计此概率上界. 当 $N \rightarrow \infty$,

$$\frac{S_N - N/2}{\sqrt{N/4}} \text{ 依概率收敛于 } \mathcal{N}(0, 1).$$

由 Gaussian 尾界

$$\begin{aligned} P\left\{S_N \geq \frac{3}{4}N\right\} &= P\left\{\frac{S_N - N/2}{\sqrt{N/4}} \geq \frac{\sqrt{N}}{2}\right\} \\ &\leq \frac{1}{\sqrt{2\pi}\sqrt{N/4}} \exp\left(-\frac{N}{8}\right) \\ &\leq \exp\left(-\frac{N}{8}\right) \quad \left(\frac{2}{\sqrt{2\pi N}} \leq 1\right) \end{aligned}$$

此概率上界随着样本容量 N 增加指数下降: N 增加一个尺度, 概率上界下降数倍尺度.

在上述例子中, 取 $N = 80$,

$$\text{(Chebyshev 不等式估计)} \quad P\{S_N \geq 60\} \leq \frac{4}{80} = 0.05$$

$$\text{(Gaussian 尾界估计)} \quad P\{S_N \geq 60\} \leq \exp(-10) \approx 0.000045.$$

这是否意味着 Gaussian 尾界估计更准确? 事实并非如此, 在 Gaussian 尾界估计中, 我们假设了 $\frac{S_N - N/2}{\sqrt{N/4}}$ 近似服从标准正态分布, 此近似的误差依赖于 Lindeberg-Lévy 中心极限定理的收敛速度, 即 Berry-Esseen 中心极限定理的结论. 因此 Gaussian 尾界估计还依赖于 $\mathcal{O}(1/\sqrt{N})$, 其上界为

$$\mathcal{O}\left(1/\sqrt{N} + \exp(-N/8)\right) = \mathcal{O}(1/\sqrt{N}).$$

从中心极限定理出发能否得到一个比 $\mathcal{O}(1/\sqrt{N})$ 更好的上界估计? 答案是否定的. 考虑如下概率

$$P\{S_N = \frac{N}{2}\} = 2^{-N} \binom{N}{N/2} \asymp \frac{1}{\sqrt{N}} \quad (\text{Stirling 公式}),$$

由中心极限定理

$$P\{Z_N = 0\} \asymp \frac{1}{\sqrt{N}}$$

另一方面 $P\{X = 0\} = 0$, 若 $X \sim \mathcal{N}(0, 1)$. 故

$$|P\{Z_N = 0\} - P\{X = 0\}| \asymp \frac{1}{\sqrt{N}}.$$

这说明 **Berry-Esseen** 中心极限定理估计误差 $\mathcal{O}(1/\sqrt{N})$ 不可避免.

为了获得更优的概率上界估计, 在下一讲中我们将绕开中心极限定理, 采用一种更直接的概率上界估计方法.