

高维概率

High-Dimensional Probability

一、高维概率简介

@滕佳焯

- 参考教材:

High-Dimensional Probability: An Introduction with Applications in Data Science (Roman Vershynin 2018)

High-Dimensional Probability

An Introduction with Applications in Data Science

Roman Vershynin

University of California, Irvine

August 1, 2018

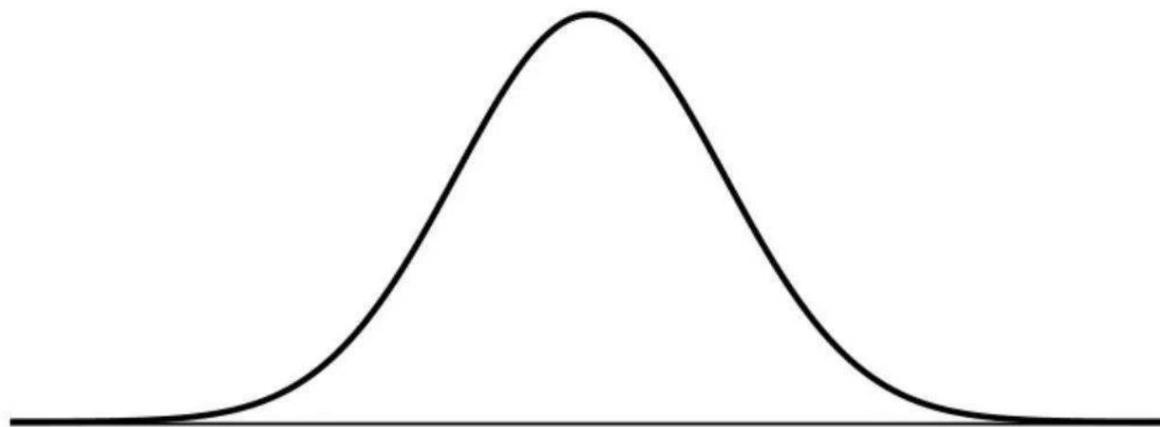
<https://www.math.uci.edu/~rvershyn/>

- 高维空间与低维空间

- 高维空间与低维空间

先考虑我们在低维空间中最常用的正态分布 $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



- 高维空间与低维空间

那高维正态分布呢？ $X \sim N(\mu, \Sigma)$

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

- 高维空间与低维空间

那高维正态分布什么样呢？ $X \sim N(0, I_n)$

$$f(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}x^T x\right)$$

$$\|X\| \approx \sqrt{n}$$

所有的点都聚集在一个球壳周围！

- 高维空间与低维空间



Figure 3.6 A Gaussian point cloud in two dimensions (left) and its intuitive visualization in high dimensions (right). In high dimensions, the standard normal distribution is very close to the uniform distribution on the sphere of radius \sqrt{n} .

- Concentration: 讨论的核心

什么是Concentration? ——随机变量的“聚集”情况

例如: 对于正态分布, 大部分的mass都留在均值 μ 的周围(3σ)
—— 3σ 原则!

我们来通过一个例子, 说说为什么Concentration这么重要

- Concentration: 一个小例子

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

- Concentration: 一个初步的尝试

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

切比雪夫不等式： $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

由于 $\mathbb{E}S_N = 0, \mathbb{D}S_N = N$ ，代入，则有

$$P\left(S_N \geq \frac{1}{2}N\right) \leq P\left(|S_N - 0| \geq \frac{1}{2}N\right) \leq \frac{4}{N}$$

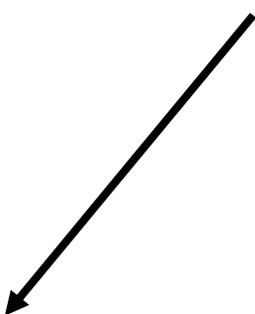
- Concentration: 一个初步的尝试

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

切比雪夫不等式: $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

线性收敛

由于 $\mathbb{E}S_N = 0, \mathbb{D}S_N = N$, 代入, 则有

$$P\left(S_N \geq \frac{1}{2}N\right) \leq P\left(|S_N - 0| \geq \frac{1}{2}N\right) \leq \frac{4}{N}$$


- Concentration: 进一步尝试

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

中心极限定理 $\sum X_i \sim \mathcal{N}(\mu, \sigma^2)$

由于 $S_N = \sum X_i$, $\mathbb{E}X_i = 0$, $\mathbb{D}X_i = 1$, 代入，则有

$$S_N \sim \mathcal{N}(0, N)$$

$$P(S_N \geq \frac{1}{2}N) \approx P(Z \geq \sqrt{N/4}) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8) \quad (\text{对足够大的}N\text{成立})$$

- Concentration: 进一步尝试

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

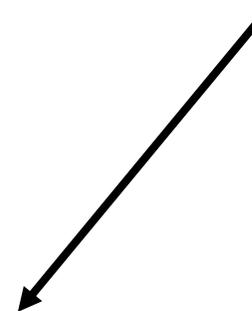
中心极限定理 $\sum X_i \sim \mathcal{N}(\mu, \sigma^2)$

由于 $S_N = \sum X_i$, $\mathbb{E}X_i = 0$, $\mathbb{D}X_i = 1$, 代入，则有

$$S_N \sim \mathcal{N}(0, N)$$

$$P(S_N \geq \frac{1}{2}N) \approx P(Z \geq \sqrt{N/4}) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8) \quad (\text{对足够大的}N\text{成立})$$

指数收敛



- Concentration: 对比

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

切比雪夫不等式：线性收敛

$$P\left(S_N \geq \frac{1}{2}N\right) \leq \frac{4}{N}$$

中心极限定理：指数收敛

$$P(S_N \geq \frac{1}{2}N) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8)$$

- Concentration: 对比

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

切比雪夫不等式：线性收敛

$$P\left(S_N \geq \frac{1}{2}N\right) \leq \frac{4}{N}$$

中心极限定理：指数收敛

估计值，而非准确值

$$P(S_N \geq \frac{1}{2}N) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8)$$


- Concentration: 发现问题!

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

那么中心极限定理估计的误差是多少呢？（Berry-Esseen central limit theorem）

$$\frac{\rho}{\sqrt{N}}$$

$$P(S_N \geq \frac{1}{2}N) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8)$$

$$P(S_N \geq \frac{1}{2}N) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8) + \frac{\rho}{\sqrt{N}}$$

- Concentration: 更好的解决方案

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

有没有更好的解决方案呢？

有！霍弗丁不等式(Hoeffding's inequality)

$$P(S_N \geq \frac{1}{2}N) \leq \exp(-N/8)$$

- Concentration: 更好的解决方案

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷 N 次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

注意霍弗丁不等式和中心极限定理的差距
二者都是**指数型**的界。

但是中心极限定理只对 $N \rightarrow \infty$ （渐进）成立
而霍弗丁不等式对任意的 N （非渐进）成立

下一个视频我们将会专门考虑这个不等式，暂时我们默认它成立。

- Concentration: 小总结

对于一个硬币而言，正面得分+1，反面得分-1，那么投掷N次后，最终得分 $S_N \geq \frac{1}{2}N$ 的概率有多高？

最后一个对比吧！

切比雪夫：线性收敛；非渐进；

$$P\left(S_N \geq \frac{1}{2}N\right) \leq \frac{4}{N}$$

中心极限定理：指数收敛；渐进；

$$P(S_N \geq \frac{1}{2}N) \leq \frac{1}{\sqrt{2\pi}} \exp(-N/8)$$

霍弗丁不等式：指数收敛；非渐进；

$$P(S_N \geq \frac{1}{2}N) \leq \exp(-N/8)$$

这些式子都代表了某个随机变量尾部取值的界，都属于Concentration

谢谢!

@滕佳焯