Lecture 2: 集中不等式、Chaining

2024.2.29

Lecturer: 丁虎　　　　*Scribe:* 王运韬

1　集中不等式

Chernoff 界可以扩展到任何次高斯（subgaussian）随机变量。

Definition 1.1. 随机变量被称为次高斯随机变量, 如果存在 $c > 0, \sigma$, 使得对充分大的$t, \mathbb{P}(x \geq t) \leq ce^{-\frac{t^2}{2\sigma^2}}$.

Example 1.2. *支撑集为闭区间 (或其他 $\mathbb{R}^d$ 上紧集) 的连续分布. 原因是在一个区间内有界而他处为 0.*

Definition 1.3. 对次高斯随机变量, 可以定义范数 [2]:

$$\|x\|_{sg} = \inf_{s>0} \mathbb{E}[e^{\frac{x^2}{s^2}} \leq 2].$$

在机器学习等任务中，时常需要计算数据的统计量, 如均值、最大值等。受限于巨量数据，难以精确计算。故而可以使用集中不等式来估计。

Theorem 1.4 (次高斯分布的 Hoeffding 界)**.** *$x_1, x_2, \ldots, x_n$ 是独立的次高斯随机变量. 均值皆为 0. 则有*

$$\mathbb{P}\left(|\sum_1^n x_i| \geq t\right) \leq 2e^{-\frac{2t^2}{\sum \|x_i\|_{sg}^2}}$$

Example 1.5. *若 $X_i \in [a_i, b_i]$, 则有*

$$\mathbb{P}\left(|\sum_1^n (x_i - \mathbb{E}X_i)| \geq t\right) \leq 2e^{-\frac{2t^2}{\sum (b_i - a_i)^2}}$$

Remark 1.6. 事实上, 上式对任意随机变量都成立, 并不局限于次高斯变量.

接下来我们引入一个有用的概念.

Definition 1.7 (鞅)**.** 对于随机变量序列 $z_i, z_2, \ldots,$ 有

$$\mathbb{E}[z_j|z_1, z_2. \ldots, z_{j_1}] = (/ \leq / \geq) z_{j-1}, \quad \forall j \in \mathbb{N}.$$

则称之为一个鞅 (martingale)(/下鞅 (submartingale)/上鞅 (supermartingale)) 序列.

Example 1.8. *一个人参加一系列赌博游戏, 每次游戏带来的收益均值都为 0. 则此人的本金 (构成的序列) 是一个鞅.*

Theorem 1.9 (Azuma 不等式)**.** *对于上述的鞅/下鞅, 如满足 $\forall i, z_i - z_{i-1} \in [a_i, b_i], |a_i - b_i| \leq c_i$, 则*

$$\mathbb{P}\left(z_j - z_0 \leq -t\right) \leq e^{-\frac{t^2}{\sum c_i^2}}$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(z_j - z_0 \leq -t\right) &= \mathbb{P}\left(\sum_{i=1}^{j}(z_{i-1} - z_i) \geq t\right) \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda\sum_{i=1}^{j}(z_{i-1}-z_i)}] && \text{(Chernoff 界)} \\
&= e^{-\lambda t}\mathbb{E}\left[\mathbb{E}[e^{\lambda\sum_{i=1}^{j}(z_{i-1}-z_i)}|z_{j-1}, \ldots, z_1]\right] && \text{(全期望公式)} \\
&= e^{-\lambda t}\mathbb{E}\left[e^{\lambda\sum_{i=1}^{j-1}(z_{i-1}-z_i)}\mathbb{E}[e^{\lambda(z_{j-1}-z_j)}|z_{j-1}, \ldots, z_1]\right] \\
&= e^{-\lambda t}\mathbb{E}\left[e^{\lambda\sum_{i=1}^{j-1}(z_{i-1}-z_i)}\mathbb{E}[e^{\lambda(\mathbb{E}[z_j]-z_j)}|z_{j-1}, \ldots, z_1]\right] && \text{(鞅的定义)} \\
&\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda\sum_{i=1}^{j-1}(z_{i-1}-z_i)}e^{-\lambda^2\frac{(b_j-a_j)^2}{8}}|z_{j-1}, \ldots, z_1\right] && \text{(上节课讲义中的 Hoeffding 引理)} \\
&\leq e^{-\lambda t}e^{-\frac{\lambda^2\sum_i(b_i-a_i)^2}{8}} && \text{(递推)} \\
&\leq e^{-\frac{t^2}{\sum c_i^2}} && \text{(优化}\lambda\text{)}
\end{aligned}$$

□

Remark 1.10. 同理可证 $\mathbb{P}\left(z_j - z_0 \geq t\right) \leq e^{-\frac{t^2}{\sum c_i^2}}$

2 Chaining

上述集中不等式提供了随机变量均值的尾概率, 要研究一族随机变量的上界 (同样是随机变量) 的尾概率, 我们引入 Chaining 方法. 以下我们研究一个具体的案例, 从中管窥 chaining 方法的一角. 更多应用参见 [1].

假设我们被给定了一族有界向量 $T \subset \mathbb{R}^n$,其在 X 范数下的直径为$\rho_X(T)$. 又有随机向量 $g \in \mathbb{R}^n$ 的每一项都是独立的标准高斯分布 (均值为 0, 方差为 1). 我们研究 $(X_t)_{t\in T}, X_t := \langle g, t\rangle$. 利用正态分布线性组合均值、方差的性质, 可以算出

$$\forall s,t \in T, \mathbb{P}(|X_s - X_t| > \lambda) \lesssim e^{-\lambda^2/(2\|s-t\|_2^2)}. \tag{1}$$

上式中,$\lesssim$ 表示右端乘以某常数 c 后 $\leq$ 成立. 接下来, 我们展示三种计算 $g(T) := \mathbb{E}_g \sup\limits_{t\in T} X_t$ 的尾分布的方法.

2.1 方法一: 合并界 (Union bound)

回顾概率论的知识, 我们有:

$$\mathbb{E}\,|Z| = \int_0^\infty \mathbb{P}(Z > u)du.$$

从而

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t = &\int_0^\infty \mathbb{P}(\sup_{t\in T} X_t > u)du \\
&\leq \int_0^{2\rho_{\ell_2}(T)\sqrt{2\log|T|}} \overbrace{\mathbb{P}(\sup_{t\in T} X_t > u)}^{\leq 1} du + \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}} \mathbb{P}(\sup_{t\in T} X_t > u)du \\
&\leq \rho_{\ell_2}(T)\sqrt{2\log|T|} + \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}}^\infty \sum_{t\in T} \mathbb{P}(X_t > u)du \text{ (合并界)} \\
&\leq \rho_{\ell_2}(T)\sqrt{2\log|T|} + |T| \cdot \int_{\rho_{\ell_2}(T)\sqrt{2\log|T|}}^\infty e^{-u^2/(2\rho_{\ell_2}(T)^2)} du \\
&= \rho_{\ell_2}(T)\sqrt{2\log|T|} + \rho_{\ell_2}(T) \cdot |T| \cdot \int_{\sqrt{2\log|T|}}^\infty e^{-\nu^2/2} d\nu \text{(变量代换)} \\
&\lesssim \rho_{\ell_2}(T) \cdot \sqrt{\log|T|}
\end{aligned} \tag{2}$$

2.2 方法二: $\epsilon-$ 网

设 $T' \subseteq T$ 为 T 的 ε-网, 定义为满足对任意 $t \in T$ 都存在 $t' \in T'$ 使得 $\|t - t'\|_2 \leq \varepsilon$ 的集合. 由 $\langle g, t\rangle = \langle g, t' + (t - t')\rangle$, 有

$$X_t = X_{t'} + X_{t-t'}.$$

从而

$$g(T) \leq g(T') + \mathbb{E}\sup_{t\in T} \langle g, t - t'\rangle\,.$$

由(2), $g(T') \lesssim \rho_{\ell_2}(T') \cdot \sqrt{\log|T'|} \le \rho_{\ell_2}(T) \cdot \sqrt{\log|T'|}$. 又有 $\langle g, t-t' \rangle \le \|g\|_2 \cdot \|t-t'\| \le \varepsilon \|g\|_2$, 以及

$$\mathbb{E}\|g\|_2 \le (\mathbb{E}\|g\|_2^2)^{1/2} \le \sqrt{n}.$$

上式第一个不等号将 $\|g\|_2 * 1$ 视作两项的乘积，作积分的 Cauchy-Schwarz 不等式，第二个不等号源于直接计算。得出：

$$\begin{aligned} g(T) &\leqslant \rho_{\ell_2}(T) \cdot \sqrt{\log|T'|} + \varepsilon\sqrt{n} \\ &= \rho_{\ell_2}(T) \cdot \log^{1/2} \mathcal{N}(T, \ell_2, \varepsilon) + \varepsilon\sqrt{n} \end{aligned} \tag{3}$$

此处 $\mathcal{N}(T, d, u)$ 表示度量熵（metric entropy）或覆盖数（covering number），定义为 $d-$ 度量空间中使用半径为 u 的球覆盖 T 所需的最小个数。易见这即为 u-网大小的下确界。可以通过优化(3)中的参数 ϵ 来得出更精细的界。不论如何，比起(2), 至少这一上界对于 T 为无穷集的情形不再平凡了.

2.3 方法三：Dudley 不等式（Chaining）

Chaining 的核心思想是，使用越来越细密，乃至可数多个网，来减小上界。设 $T_r \subset T$ 为 T 的 $2^{-r}\rho_{\ell_2}(T)$-网，其覆盖半径记为 ϵ_r，t_r 是 T_r 中距离 t 最近的点。则有

$$\langle g, t \rangle = \langle g, t_0 \rangle + \sum_{r=1}^{\infty} \langle g, t_r - t_{r-1} \rangle$$

这是因为 t_r 距离 t 趋于零。

$$\begin{aligned} g(T) &\le \sum_{r=1}^{\infty} \mathbb{E} \sup_{t \in T} \langle g, t_r - t_{r-1} \rangle \\ &\lesssim \sum_{r=1}^{\infty} \frac{\rho_{\ell_2}(T)}{2^r} \cdot \log^{1/2}(\mathcal{N}(T, \ell_2, \frac{\rho_{\ell_2}(T)}{2^r})^2) \qquad \text{(由 (2) 及三角不等式)} \\ &\lesssim \sum_{r=1}^{\infty} \frac{\rho_{\ell_2}(T)}{2^r} \cdot \log^{1/2} \mathcal{N}(T, \ell_2, \frac{\rho_{\ell_2}(T)}{2^r}). \end{aligned} \tag{4}$$

在此对 (4) 做一解释：由三角不等式，可知 $\|t_r - t_{r-1}\|$ 不超过 $\epsilon_r + \epsilon_{r-1}$，注意到 $|T_r - T_{r-1}| \le |T_r| \cdot |T_{r-1}| \le |T_r|^2$，$T_r$ 选为最小的 ϵ_r-网，直接利用合并界 (2) 即可。

同学们可以尝试将 T 取成欧氏空间内的单位球或方块，具体计算一下。将会发现通过 Chaining 得到的界更优。

References

[1] J. Nelson. *Chaining introduction with some computer science applications*. Bulletin of EATCS 3, no. 120, 2016.

[2] R. Vershynin. *Concentration of Sums of Independent Random Variables*, page 11–37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.