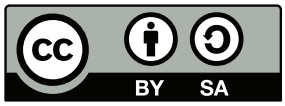
# 机器学习理论学习笔记(3):
# 集中不等式：用经验误差估计泛化误差的依据下篇

王龙腾
longtenwang@gmail.com

2017 年 3 月 11 日






## 1 回顾

在上一篇笔记中，我们已经从马尔科夫不等式出发，利用Chernoff bound这个工具证明了Hoeffding不等式：

对于n个独立随机变量序列$X_1, X_2, \ldots, X_n$，若$X_i \in [a_i, b_i], i = 1, 2, \ldots, n$,记$S_m = \sum_{i=1}^{m} X_i$,则

$$P(S_m - \mathbb{E}[S_m] \geq \epsilon) \leq \exp\left\{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2\right\} \tag{1}$$

在上篇的最后，我们讲过，根据（1），我们进行随机变量组成的函数偏离其期望的估计时不用考虑任何随机变量分布的信息，然而作为代价，我们要知道随机变量有确定的上界和下界，如果我们预先知道了方差和期望这样的关于随机变量分布的信息，并且我们只知道随机变量的上界，那么，我们能得到什么样的估计呢。此外，如果（1）中的$S_m$不是多个随机变量的和，而是多个随机变量组成的函数，那么，我们可以得到什么样子的估计呢？本文就从这两个思路，介绍Bennett 不等式，Bernstein不等式，以及Azuma不等式和McDiarmid不等式。

## 2 Bennett不等式与Bernstein不等式

说明：推理思路来自 [2]的附录D的练习题，证明引理1.1的构造不等式的技巧，受文献[1]的启发。

**引理1.1：** 如果随机变量$X$满足$X \le c, \mathbb{E}[X] = 0, \mathbb{E}[X^2] = \sigma^2, c > 0$，那么$\forall t > 0$：

$$\mathbb{E}[e^{tX}] \le e^{f(\sigma^2/c^2)} \tag{2}$$

其中$f(x) = \log\left(\frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}\right)$

证明. 1.构造函数

$$g(x) = \begin{cases} \frac{e^x - 1 - x}{x^2}, x \neq 0 \\ \frac{1}{2}, \qquad x = 0 \end{cases} \tag{3}$$

经过计算，我们可以得到

$g'(0) = 0, x^3 g'(x) = h(x), h(x) = xe^x - 2e^x + x + 2, h(0) = 0, h'(0) = 0, h''(x) = xe^x > 0$

因此，我们会有在$(0, \infty)$上$h''(x) > 0$,在$(-\infty, 0)$上$h''(x) < 0$,故$h'(x) \ge 0$那么我们最终得到$g'(x) = h(x)/x^3 > 0, x \neq 0$,也就是说$g(x)$是个单调增的函数。

2.由题设，$X \le c, c > 0$,故$X \le c$,故$g(t(X + \sigma^2/c)) \le g(t(c + \sigma^2/c))$，代入（3），我们得到

$$\frac{e^{t(X+\sigma^2/c)} - 1 - t(X + \sigma^2/c)}{t^2(X + \sigma^2/c)^2} \le \frac{e^{t(c+\sigma^2/c)} - 1 - t(c + \sigma^2/c)}{t^2(c + \sigma^2/c)^2}$$

两边同乘以$(X + \sigma^2/c)^2$ 后根据期望的性质，给不等号两边取期望，化简可得:

$$\mathbb{E}[e^{tX}]e^{t\sigma^2/c} - 1 - t\sigma^2/c \le \frac{\sigma^2(1 + \sigma^2/c^2)}{c^2(1 + \sigma^2/c^2)^2}(e^{t(c+\sigma^2/c)} - 1 - t(c + \sigma^2/c))$$

将上式整理即得所求结论。 □

**引理1.2：** 对于$f(x) = \log\left(\frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}\right)$,有$\forall x > 0, t > 0, f''(x) < 0$

证明. 令$u(x) = \frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}$,计算可得$u''u - (u')^2 < 0, \forall x \ge 0$，

故对于$f(x) = \log(u(x)), f'(x) = \frac{u'}{u}, f''(x) = \frac{u''u - (u')^2}{u^2} \le 0$ □

**引理1.3：** 对于$f(x) = \log\left(\frac{1}{1+x}e^{-ctx} + \frac{x}{1+x}e^{ct}\right)$，我们有

$$f(x) \le (e^{ct} - 1 - ct)x, \forall x \ge 0 \tag{4}$$

证明. 计算可得$f'(0) = e^{ct} - 1 - ct$ 在$x = 0$处做泰勒展开

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2}f''(\theta), 0 \le \theta \le x$$

结合引理1.2，得$f(x) \le f(0) + xf'(0) = (e^{ct} - 1 - ct)x$ □

**定理**1(Bennett inequality): 对于独立的，均值分别为零的随机变量$X_1, X_2, \ldots, X_m$，满足$X_i \leq c, i = 1, 2, \ldots, m$，令$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\sigma_{X_i}^2$,记$\theta(x) = (1+x)\log(1+x) - x$,则我们有：

$$P[\frac{1}{m}\sum_{i=1}^{m}X_i \geq \epsilon] \leq \exp\left(-\frac{m\sigma^2}{c^2}\theta(\frac{\epsilon c}{\sigma^2})\right) \tag{5}$$

证明. 根据Chernoff bound，我们有：

$$P[\frac{1}{m}\sum_{i=1}^{m}X_i \geq \epsilon] \leq e^{-tm\epsilon+\sum_{i=1}^{m}f(\sigma_{X_i}^2/c^2)}$$

把（4）代入：

$$P[\frac{1}{m}\sum_{i=1}^{m}X_i \geq \epsilon] \leq e^{-tm\epsilon+(\sigma_{X_i}^2/c^2)\sum_{i=1}^{m}(e^{ct}-1-ct)}$$

对$F(t) = -tm\epsilon+(\sigma_{X_i}^2/c^2)\sum_{i=1}^{m}(e^{ct}-1-ct)$，分析可得，其最大值点为$t = \frac{1}{c}\log\left(\frac{m\epsilon}{c\sum_{i=1}^{m}\sigma_{X_i}^2/c^2}\right)$代入上式，化简既可证得所求。 □

一个更松的估计是Bernstein inequalities

**定理**2(Bernstein inequality): 对于独立的，均值分别为零的随机变量$X_1, X_2, \ldots, X_m$，满足$X_i \leq c, i = 1, 2, \ldots, m$，令$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\sigma_{X_i}^2$,记$\theta(x) = (1+x)\log(1+x) - x$,则我们有：

$$P[\frac{1}{m}\sum_{i=1}^{m}X_i \geq \epsilon] \leq \exp\left(-\frac{m\epsilon^2}{2\sigma^2+2c\epsilon/3}\right) \tag{6}$$

证明. 构造函数$\varphi(x) = \theta(x) - \frac{3}{2}\frac{x^2}{x+3} = (1+x)\log(1+x) - x - \frac{3}{2}\frac{x^2}{x+3}$，计算可得

$$\varphi'(x) = \frac{27}{2(x+3)^2} + \log(x+1) - 3/2$$

而$\varphi''(x) = (x^2(9+x))/((1+x)(3+x)^3) \geq 0$故$\forall x \geq 0, \varphi'(x) \geq \varphi'(0), \varphi(x) \geq \varphi(0)$，故

$$\theta(x) \geq \frac{3}{2}\frac{x^2}{x+3}$$

最终得到：

$$-\frac{m\sigma^2}{c^2}\theta(\frac{\epsilon c}{\sigma^2}) \leq -\frac{m\epsilon^2}{2\sigma^2+2c\epsilon/3}$$

结合(5)，我们即得所求。 □

# 3 Azuma不等式和McDiarmid不等式

在这之前，让我们先定义，所谓Martignale difference sequnce(Martignale(鞅)，没找到合适的翻译,所以直接用英文了),

**定义**: 随机变量序列$V_1, V_2, \ldots$称为随机变量序列$X_1, X_2, \ldots$的 Martignale difference sequnce, 如果满足$\forall i > 0, V_i = f(X_i, \ldots, X_i), \mathbb{E}[V_{i+1}|X_1, \ldots, X_i] = 0$ **引理**3.1: 如果随机变量$V, Z$满足$\mathbb{E}[V|Z] = 0, \exists f, c \geq 0$，

$$f(Z) \leq Z \leq f(Z) + c$$

则$\forall t > 0$,

$$\mathbb{E}[e^{tV}|Z] \leq e^{t^2c^2/8} \tag{7}$$

证明的思路同上一篇的中证明(17)式的一样，不过是把$X$换成了$V|X$ 而$a = f(Z), b = f(Z) + c$就不再赘述了。

**引理**3.2(Azuma's inquality): 如果随机变量序列$V_1, V_2, \ldots$为随机变量序列$X_1, X_2, \ldots$的Martignale difference sequnce，满足$\forall i \geq 1, i = 1, 2, \ldots, \exists c_i > 0, Z_i = g(V_1, \ldots, V_{i-1})$,且有

$$Z_i \leq V_i \leq Z_i + c$$

那么，我们有$\forall \epsilon > 0$

$$P[\sum_{i=1}^{m} V_i \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right) \tag{8}$$

证明. $\forall k \in [1, m]$,令$S_k = \sum_{i=1}^{k} V_k$,根据Chernoff bound，结合引理3.1，我们有$\forall t > 0$：

$$\begin{aligned} P[S_m \geq \epsilon] &\leq e^{-t\epsilon}\mathbb{E}[e^{tS_m}] \\ &= e^{-t\epsilon}\mathbb{E}[e^{tS_{m-1}}\mathbb{E}[e^{tV_m}|X_1, \ldots, X_{m-1}]] \\ &\leq e^{-t\epsilon}e^{t^2\sum_{i=1}^{m} c_i^2/8} \end{aligned}$$

右边正数第二个式子的依据是$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$,第三个式子是反复用引理3.1得到的。找使得$e^{-t\epsilon}e^{t^2\sum_{i=1}^{m} c_i^2/8}$最小的t，我们得到$t = 4\epsilon/\sum_{i=1}^{m} c_i^2$,将其代入$e^{-t\epsilon}e^{t^2\sum_{i=1}^{m} c_i^2/8}$，我们即得到所求。 □

之前我们说过，注意到(1)的左边，我们如果把$S_m = \sum_{i=1}^{m} X_i$做进一步的抽象，将其看作为m个随机变量的函数$f(X_1, X_2, \ldots, X_m)$，那么我们会得到什么样的估计呢？

在实际应用中，我们常常考虑一种特殊的函数，有有界变差的函数：

$\forall 1 \leq i \leq m$,函数$f$满足

$$\sup_{x_1, \ldots, x_m, x_i'} |f(x_1, \ldots, x_i', \ldots, x_m) - f(x_1, \ldots, x_i, \ldots, x_m)| \leq c_i$$

对于这类函数，我们有McDiarmid's inequality:

**定理**3(McDiarmid's inequality): 若函数$f(X_1, X_2, \ldots, X_m)$有有界变差，也就是说$\forall 1 \leq i \leq m$,函数$f$满足

$$\sup_{x_1,\ldots,x_m,x_i'} |f(x_1, \ldots, x_i', \ldots, x_m) - f(x_1, \ldots, x_i, \ldots, x_m)| \leq c_i$$

那么我们就有：

$$P[f - \mathbb{E}[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) \tag{9}$$

注意到(8),(9)的右边是相似的，因此我们的思路是从(8)推出(9),因此，我们需要先根据定理3题设构造出Martignale difference sequnce。我们照着这个思路进行证明。

证明. 设$X^i = (X_1, \ldots, X_i), X^0 = 1$ 构造随机变量$V_k, k \in [1, m], V = f - \mathbb{E}[f], V_k = \mathbb{E}[f|X^k] - \mathbb{E}[f|X^{k-1}]$,则

$$\sum_{i=1}^n V_i = \sum_{i=1}^n \mathbb{E}[f|X^i] - \mathbb{E}[f|X^{i-1}] = \mathbb{E}[f|X^n] - \mathbb{E}[f] = f - \mathbb{E}[f] = V$$

进一步，我们有：

$$\mathbb{E}[\mathbb{E}[V|X^k]|X^{k-1}] = \mathbb{E}[V|X^{k-1}]$$

故

$$\mathbb{E}[V_k|X^{k-1}] = \mathbb{E}[(\mathbb{E}[V|X^k] - \mathbb{E}[V|X^{k-1}])|X^{k-1}] = 0$$

因此$V_1, \ldots, V_m$是Martignale difference sequnce。此外，定义$V_k$的上界$W_k$和下界$U_k$:

$$W_k = \sup_x \mathbb{E}[f|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f|X_1, \ldots, X_{k-1}]$$
$$U_k = \inf_x \mathbb{E}[f|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f|X_1, \ldots, X_{k-1}]$$

根据题设，我们有：

$$W_k - U_k = \sup_{x,x'} \mathbb{E}[f|X_1, \ldots, X_{k-1}, x] - \mathbb{E}[f|X_1, \ldots, X_{k-1}, x'] \leq c_k$$

因此，我们有$U_k \leq V_k \leq U_k + c_k$，于是我们应用Azuma's inquality，得到$P[\sum_{i=1}^m V_i \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right)$，而$\sum_{i=1}^n V_i = f - \mathbb{E}[f]$，故定理3到证明。 □

我们到现在为止，已经谈了很多的概率论了，下篇笔记开始，我们就要开始用这些升级的概率论知识，来讲述PAC理论。

# 参考文献

[1] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.