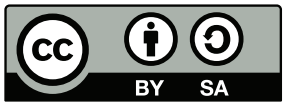
机器学习理论学习笔记(2):
集中不等式：用经验误差估计泛化误差的依据上篇

王龙腾
longtenwang@gmail.com

2017 年 3 月 11 日



在这一篇笔记中，我们将回顾与介绍几个重要的不等式，他们被称为集中不等式*(concentration inequalities)*。这些不等式给出了随机变量偏离某些值特别是期望值的界的估计，这些不等式，为我们在机器学习中用经验误差估计泛化误差提供了理论依据。考虑篇幅，笔者将这个主题分为两篇。熟悉1，2两部分内容的读者可以跳过直接从第3部分开始看。相关内容请参看课程CS229,以及 [3]和 [5]

1 概率论中的知识回顾

这里，我们假定读者已经在概率论课本中接触了随机变量以及分布函数的知识，并且知道如下我们会在后文中用到的几个关于数学期望的知识：

$$\mathbb{E}[X] \geq 0, \textit{if } X \geq 0 \tag{1}$$

$$\mathbb{E}[aX] = a\mathbb{E}[X], a \textit{ is a constant} \tag{2}$$

$$\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] \tag{3}$$

$$\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i], \textit{if } X_i, i = 1, 2, \ldots, n \textit{ are independent} \tag{4}$$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] \tag{5}$$

此外，我们还要引入一个有用的叫做矩母函数这样的东西：

$$M(t) = \mathbb{E}[e^{tX}] \tag{6}$$

对于矩母函数，我们有

$$M^{(n)}(t) = \mathbb{E}[X^n] \tag{7}$$

根据矩母函数的定义，我们可以计算服从标准正态分布$\mathcal{N}(0,1)$的随机变量Z的矩母函数：

$$\begin{aligned} M_Z(t) &= \mathbb{E}[e^{tZ}] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{x^2 - 2tx}{2} \right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{(x-t)^2}{2} + \frac{t^2}{2} \right\} dx \\ &= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

最后一个积分读者们应该都会算吧。。。如若读者想复习相关概率论的基本知识，请参看 [4]。

2 单个随机变量偏移其期望的概率估计

第一个集中不等式为马尔科夫不等式:

定理2.1(Markov's inequality)：对于非负随机变量X,若$\mathbb{E}[X] < +\infty$则$\forall t \geq 0$我们有

$$P(X \geq t\mathbb{E}[X]) \leq \frac{1}{t} \tag{8}$$

证明. 注意到$P(X \geq t\mathbb{E}[X]) = \mathbb{E}[\mathbb{I}(X \geq t\mathbb{E}[X])]$,其中$\mathbb{I}$为指示函数,如果$X \geq t\mathbb{E}[X]$,那么有$X/t\mathbb{E}[X] \geq 1 \geq \mathbb{I}(X \geq t\mathbb{E}[X])$，当$X < t\mathbb{E}[X]$，我们依然有$X/t\mathbb{E}[X] \geq 0 = \mathbb{I}(X \geq t\mathbb{E}[X])$，于是，我们恒有：

$$X/t\mathbb{E}[X] \geq \mathbb{I}(X \geq t\mathbb{E}[X])$$

根据(1)和(3),我们有:

$$E[X] \geq E[Y], \textit{if } X \geq Y \tag{9}$$

故我们有

$$\begin{aligned} P(X \geq t\mathbb{E}[X]) &= \mathbb{E}[\mathbb{I}(X \geq t\mathbb{E}[X])] \\ &\leq \mathbb{E}[X/t\mathbb{E}[X]] \\ &= \mathbb{E}[X]/t\mathbb{E}[X] \\ &= \frac{1}{t} \end{aligned}$$

□

令(8)中的$tE[X] = \epsilon$,则我们有马尔科夫不等式另一种等价形式：

$$P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon} \tag{10}$$

由(10)我们可以得到：

$$P(|X - \mathbb{E}[x]| \geq \epsilon) = P((X - \mathbb{E}[X])^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[((X - \mathbb{E}[X])^2]}{\epsilon^2} = \frac{\mathrm{Var}(X)}{\epsilon^2}$$

于是，我们有了第二个集中不等式，切贝谢夫不等式：

推论2.1.1(Chebyshev's inequality)：如果对于任意随机变量X只有有限的方差，则$\forall \epsilon \geq 0$

$$P(|X - \mathbb{E}[x]| \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} \tag{11}$$

马尔科夫不等式和切贝谢夫不等式的重要性在于，在我们只知道分布的期望，或者分布的期望和方差的情况下，计算概率上界。

结合我们在第一部分提到的矩母函数的概念，在随机变量X的矩母函数$M(t)$已知时,利用马尔科夫不等式，我们可以得到更加有效的估计概率$P(X \geq a)$的上界：

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \mathbb{E}[e^{tX}]e^{-ta}, \forall t > 0$$

$$P(X \leq a) = P(e^{tX} \geq e^{ta}) \leq \mathbb{E}[e^{tX}]e^{-ta}, \forall t < 0$$

于是我们有：

推论2.1.2(Chernoff bound):

$$\begin{aligned} P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq e^{-ta}\mathbb{E}[e^{tX}], \forall t > 0 \\ P(X \leq a) = P(e^{tX} \geq e^{ta}) \leq e^{-ta}\mathbb{E}[e^{tX}], \forall t < 0 \end{aligned} \tag{12}$$

显然，如果我们想知道概率$P(X \geq a)$的最小上界(上确界)，则我们可以通过求解最值问题：$\max\limits_{t>0} e^{-ta}\mathbb{E}[e^{tX}]$来实现。

3 多个随机变量的组成的函数偏移其期望的概率估计：初步结果

在第1篇笔记中，我们讲过，在机器学习的实际应用中，我们是用样本空间的采样来训练学习算法的。在机器学习的一个分支————统计学习里，有一个假设是，我们的样本空间中的所有样本服从一个未知的分布$\mathcal{D}$,这个分布我们称其为*数据生成分布* [1]。我们假设样例集中的所有样本都是独立地从数据生成分布中采样而得，用术语来说就是，我们假设所有用来训练，校正，测试学习算法的样本均为独立同分布(independent and identical distributed,*i.i.d*)样本。

对于n个*i.i.d*样本$X_1, X_2, ..., X_n$，由于他们的分布相同，所以他们的期望与方差也都也相同，假设他们的期望也都有限，不妨设为μ, σ令$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则根据(2),(3),(4)可得$\mathbb{E}[\bar{X}] = \mu, \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$代入(11)得：

$$P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \tag{13}$$

根据(13),我们可以导出*弱大数定律*：

定理2.2 **弱大数定理**(Weak law of large numbers)：对于独立同分布随机变量序列，其期望与方差存在且有限，则$\forall \epsilon > 0$:

$$\lim_{n|\infty} P(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon) = 0 \tag{14}$$

4 多个随机变量的组成的函数偏移其期望的概率估计：Hoeffding不等式

让我们把限制放宽一点，如果我们只要求随机变量序列$X_1, X_2, \ldots, X_n$彼此独立就好了，应用切尔诺夫界与(4)，我们可以得到：

$$\begin{aligned} P(\sum_{i=1}^{n} X_i \geq a) &\leq e^{-ta}\mathbb{E}[e^{t\sum_{i=1}^{n} X_i}], \forall t > 0 \\ &= e^{-ta}\mathbb{E}[\prod_{i=1}^{n} e^{tX_i}] \\ &= e^{-ta}\prod_{i=1}^{n}\mathbb{E}[e^{tX_i}] \end{aligned} \tag{15}$$

在实际中，我们的样本通常是有界的，这启示我们，如果假设随机变量$X_i \in [a, b]$，应用（15）我们可以得到什么？Hoeffding给出了一个回答 [2]

定理2.3(Hoeffding inquality):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$S_n = \sum_{i=1}^{n} X_i$,则

$$P(S_m - \mathbb{E}[S_m] \geq \epsilon) \leq \exp\left\{ -2\epsilon^2 / \sum_{i=1}^{n} (b_i - a_i)^2 \right\} \tag{16}$$

证明. 首先，我们先考虑一个随机变量的情况。对于随机变量$X \in [a, b]$,注意到$\mathbb{E}[X - \mathbb{E}] = 0$,因此，我们不妨考虑只考虑当$\mathbb{E}[X] = 0$的情况,当$\mathbb{E}[X] = 0$时,考察(12)中的最后一行中含有矩母函数那一项：

$$\begin{aligned}
\mathbb{E}[e^{tX}] &\leq \mathbb{E}[\frac{b - X}{b - a} e^{ta} + \frac{X - a}{b - a} e^{tb}], \textit{use convexity of } f(x) = e^x \\
&= \mathbb{E}[\frac{b}{b - a} e^{ta} + \frac{-a}{b - a} e^{tb}], \textit{use } \mathbb{E}[X] = 0
\end{aligned}$$

这提示我们通过对函数$f(t) = \frac{b}{b-a} e^{ta} + \frac{-a}{b-a} e^{tb}$的上界进行估计，从而我们可以估计$\mathbb{E}[e^{tX}]$的上界。

为了估计函数$f(t) = \frac{b}{b-a} e^{ta} + \frac{-a}{b-a} e^{tb}$的上界，令$h = t(b - a), p = -\frac{a}{b-a}$,解得$a = -\frac{h}{t}, b = (1 - p)\frac{h}{t}$，$f(t)$变为$f(h) = (1 - p + pe^h)e^{-hp}$ 对于函数$g(h) = -hp + ln(1 - p + pe^h)$，我们可以得到：

$$g''(h) = \frac{pe^h(1 - p)}{(1 - p + pe^h)^2}$$

考察分母$(1 - p + pe^h)^2$,因为$(1 - p - pe^h)^2 \geq 0$,故$(1 - p + pe^h)^2 \geq 4pe^h(1 - p)$,代入上面这个二阶导数的表达式，我们可以得到：

$$g''(h) \leq \frac{1}{4}$$

我们发现，$g(0) = 0, g'(0) = 0$,对$g(h)$在$h = 0$处的领域$(-\delta, \delta)$做二阶泰勒展开，我们可以得到：

$$\begin{aligned}
g(h) &= g(0) + g'(0)h + \frac{g''(\theta)h^2}{2}, \theta \in (-\delta, \delta) \\
&\leq \frac{h^2}{2} \times \frac{1}{4} \\
&= \frac{h^2}{8}
\end{aligned}$$

回想起$h = t(b - a)$，故最终我们得到：

$$\mathbb{E}[e^{tX}] \leq \exp\left\{ \frac{t^2(b - a)^2}{8} \right\} \tag{17}$$

(17)被称为Hoeffding 引理，是我们证明Hoeffding 不等式的基础。根据(17),结合切尔诺夫界,我们得到：

$$\begin{aligned} P(S_m - \mathbb{E}[S_m] \geq \epsilon) &\leq e^{-t\epsilon} \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] \\ &\leq e^{-t\epsilon} \exp\left\{ \frac{t^2(b_n - a_n)^2}{8} \right\} \prod_{i=1}^{n-1} \mathbb{E}[e^{tX_i}] \\ &\leq e^{-t\epsilon} \exp\left\{ \sum_{i=1}^{n} \frac{t^2(b_i - a_i)^2}{8} \right\} \\ &\leq \exp\left\{ \sum_{i=1}^{n} \frac{t^2(b_i - a_i)^2}{8} - t\epsilon \right\} \end{aligned}$$

求解函数$f(t) = \exp\left\{ \sum_{i=1}^{n} \frac{t^2(b_i - a_i)^2}{8} - t\epsilon \right\}$的最值，其为$\exp\{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2\}$，于是我们最终得到了

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

□

证明(16)之后，我们先比较一下(11)和(16)之间的区别。注意到两式的左边的形式是相同的，但是(16)的右边并不需要方差的信息，不过作为代价，我们需要知道随机变量的取值范围。

参考文献

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[2] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

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

[4] Sheldon Ross. *A first course in probability*. Pearson, 2015.

[5] 周志华. 机器学习. 清华大学出版社, 2016.