6 集中不等式(Concentration Inequalities)

6.1 为什么研究集中不等式?

通常给定一个训练数据集

$$S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\},$$

其中, $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$表示第$i$个训练数据的特征, $y_i \in \mathcal{Y} = \{0, 1\}$表示第$i$个训练数据的标记, 为了简单起见, 这里仅仅考虑二分类问题. 假设$\mathcal{D}$是空间$\mathcal{X} \times \mathcal{Y}$的一个联合分布, 其实际应用中未知不可见. 机器学习的经典假设是训练数据集S_n中每个数据$(\boldsymbol{x}_i, y_i)$ 是根据分布$\mathcal{D}$独立同分布采样所得.

给定一个函数或分类器$f \colon \mathcal{X} \to \{0, 1\}$, 可定义函数$f$在训练数据集$S_n$ 上的分类错误率为

$$\hat{R}(f, S_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i),$$

这里$\mathbb{I}(\cdot)$表示指示函数, 当论断为真时其返回值为1, 否则为0.

实际中更为关心函数f在分布$\mathcal{D}$上的分类错误率, 即定义

$$R(f, \mathcal{D}) = E_{(\boldsymbol{x}, y) \sim \mathcal{D}}(\mathbb{I}(f(\boldsymbol{x}) \neq y)) = \Pr_{(\boldsymbol{x}, y) \sim \mathcal{D}}[f(\boldsymbol{x}) \neq y].$$

由于分布$\mathcal{D}$未知不可见, 不能直接计算$R(f, \mathcal{D})$. 我们仅有一个训练数据集S_n以及训练错误率$\hat{R}(f, S_n)$, 如何有效估计$R(f, \mathcal{D})$? 在实际中, 我们非常关心

$$\Pr_{S_n \sim \mathcal{D}^n} \left[|\hat{R}(f, D_m) - R(f)| \geq t \right] \text{ 是否足够小?}$$

即我们能否以很大的概率保证

$$|\hat{R}(f, D_m) - R(f)| < t,$$

从而以很大的概率保证$\hat{R}(f, D_m)$是$R(f)$的一个有效估计. 上述性质在机器学习中被称为泛化性, 是机器学习模型理论研究的根本性质, 研究模型能否从可见的训练数据推导出对未见数据的处理能力.

例6.1. *对二分类问题, 假设训练数据集$S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$根据分布$\mathcal{D}$独立采样所得, 一个分类器$f$在训练数据集$S_n$的错误率为$\hat{p}$,*

- *若$\hat{p} \neq 0$, 求函数f在分布$\mathcal{D}$上的错误率介于$\hat{p}/2$和$3\hat{p}/2$之间的概率;*
- *若$\hat{p} = 0$, 求函数f在分布$\mathcal{D}$上的错误率介于0和$\epsilon > 0$之间的概率.*

解. 为简单起见, 我们引入随机变量

$$X_i = \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i].$$

那么训练错误率

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

不妨假设函数f在分布$\mathcal{D}$上的错误率为

$$p=\Pr_{(\boldsymbol{x},y)\sim\mathcal{D}}[f(\boldsymbol{x})\neq y]=E[X_i]=E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right].$$

由此可得随机变量$X_i\sim\mathrm{Ber}(p)$, 其方差为$p(1-p)$. 根据Chebyshcve不等式有

$$\Pr[|p-\hat{p}|>\epsilon]\leq\frac{1}{\epsilon^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{p(1-p)}{n\epsilon^2}\leq\frac{1}{4n\epsilon^2}.$$

取$\epsilon=\hat{p}/2$, 有$\Pr[|p-\hat{p}|>\hat{p}/2]\leq 1/n\hat{p}^2$, 因此$p\in(\hat{p}/2,3\hat{p}/2)$至少以$1-1/n\hat{p}^2$的概率成立.

当$\hat{p}=0$时, 根据独立性条件有

$$\begin{aligned}\Pr[p\geq\epsilon,\hat{p}=0]&\leq\Pr[X_i=\mathbb{I}[f(\boldsymbol{x}_i)\neq y_i]=0(i=1,\ldots,n)|p\geq\epsilon]\\&=\prod_{i=1}^{n}\Pr[X_i=\mathbb{I}[f(\boldsymbol{x}_i)\neq y_i]=0|p\geq\epsilon]\leq(1-\epsilon)^n\leq\exp(-n\epsilon).\end{aligned}$$

因此$p\in(0,\epsilon)$至少以$1-\exp(-n\epsilon)$的概率成立. □

从上例的求解可知, 假设随机变量

$$X_i=\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i),$$

则机器学习问题可通过概率统计抽象描述为: 假设有m个独立同分布的随机变量X_1,X_2,X_m, 如何从m个独立同分布的随机变量中以很大概率地获得期望$E[X]$ 的一个估计, 即

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m}X_i-E(X_i)\right|>\epsilon\right]\quad\text{非常小.}$$

后续研究将不再给出机器学习的实际应用, 仅仅讨论概率论中的随机变量, 但大家要了解随机变量背后的实际应用. 其次, 上例通过Chebyshcve不等式得到概率的上界, 有没有更紧的上界, 这就是本章讨论的问题：集中不等式.

6.2 基础不等式

首先给出一些基础的概率或期望不等式. 首先研究Markov不等式:

定理6.1 (Markov不等式)**.** *设随机变量$X\geq 0$, 对任意$\epsilon>0$, 有*

$$P(X\geq\epsilon)\leq\frac{E(X)}{\epsilon}.$$

Proof. 利用全期望公式考虑随机事件$X\geq\epsilon$, 有

$$E[X]=E[X|X\geq\epsilon]P(X\geq\epsilon)+E[X|X\leq\epsilon]P(X\leq\epsilon)\geq Pr(X\geq\epsilon)\epsilon$$

从而完成证明. □

利用Markov不等式可以推导Chebyshev不等式:

定理6.2 (Chebyshev不等式). 设随机变量X的均值为μ, 则

$$P(|X-\mu|>\epsilon)\le\frac{Var(X)}{\epsilon^2}.$$

Proof. 根据Markov不等式有

$$P(|X-\mu|>\epsilon)=P((X-\mu)^2\ge\epsilon^2)\le\frac{E(X-\mu)^2}{\epsilon^2}=\frac{Var(X)}{\epsilon^2}.$$

□

比Chebyshev不等式更紧地Cantelli不等式, 又被成为单边Chebyshev不等式.

引理6.1 (Cantelli不等式). 假设X是一个均值为$\mu>0$, 方差为σ^2的随机变量. 对任意$\epsilon>0$, 有

$$P(X-\mu\ge\epsilon)\le\frac{\sigma^2}{\sigma^2+\epsilon^2}\quad 和\quad P(X-\mu\le-\epsilon)\le\frac{\sigma^2}{\sigma^2+\epsilon^2}.$$

Proof. 设随机变量$Y=X-\mu$, 则有$E(Y)=0$以及$Var(Y)=\sigma^2$. 对任意$u>0$, 有

$$\begin{aligned}P(X-\mu\ge\epsilon) &= P(Y\ge\epsilon)=P(Y+u\ge\epsilon+u)\le P((Y+u)^2\ge(\epsilon+u)^2)\\ &\le \frac{E((Y+u)^2)}{(\epsilon+u)^2}=\frac{\sigma^2+u^2}{(\epsilon+u)^2}\end{aligned}$$

设$u=\sigma^2/\epsilon$, 由此得到

$$P(X-\mu\ge\epsilon)\le\min_{u>0}\frac{\sigma^2+u^2}{(\epsilon+u)^2}=\frac{\sigma^2}{\epsilon^2+\sigma^2}.$$

另一方面, 对任意$u>0$, 有

$$\begin{aligned}P(X-\mu\le-\epsilon) &= P(Y\le-\epsilon)=P(Y-u\le-\epsilon-u)\le P((Y+u)^2\ge(\epsilon+u)^2)\\ &\le \frac{E((Y+u)^2)}{(\epsilon+u)^2}=\frac{\sigma^2+u^2}{(\epsilon+u)^2}\end{aligned}$$

类似完成证明. □

下面介绍Chebyshev不等式的推论.

推论6.1. 对n个独立同分布的随机变量$X_1,X_2,\dots,X_n$, 如果满足$E(X_i)=\mu$和$Var(X_i)\le\sigma^2$, 则对任意$\epsilon>0$, 有

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X_i-\mu\right|\ge\epsilon\right)\le\frac{\sigma^2}{n\epsilon^2}$$

Proof. 根据Chebyshev不等式有

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^n X_i-\mu\right|\ge\epsilon\right)\le\frac{1}{\epsilon^2}\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^n X_i\right).$$

而根据方差的性质有

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}Var\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{n}Var(X_i)\leq\frac{\sigma^2}{n}.$$

由此得到

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

从而完成证明. □

引理6.2 (Young不等式)**.** 给定非负实数a,b, 对任意满足$1/p+1/q=1$的非负实数p,q, 有

$$ab\leq\frac{1}{p}a^p+\frac{1}{q}b^q.$$

Proof. 根据凸函数性质有

$$\begin{aligned}ab &= \exp(\ln(ab))=\exp(\ln a+\ln b)=\exp\left(\frac{1}{p}\ln a^p+\frac{1}{q}\ln b^q\right)\\ &\leq \frac{1}{p}\exp\left(\ln a^p\right)+\frac{1}{q}\exp\left(\ln b^q\right)=\frac{1}{p}a^p+\frac{1}{q}b^q.\end{aligned}$$

引理得证. □

根据Young不等式可证明著名的Hölder不等式.

引理6.3 (Hölder不等式)**.** 对任意随机变量X和Y以及实数$p>0$和$q>0$, 满足$1/p+1/q=1$, 有

$$E(|XY|)\leq(E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}.$$

特别地, 当$p=q=2$时*Hölder*不等式变成为*Cauchy-Schwartz*不等式.

Proof. 设$c=(E(|X|^p))^{\frac{1}{p}}$和$d=(E(|Y|^q))^{\frac{1}{q}}$, 根据Young不等式有

$$\frac{|X|}{c}\frac{|Y|}{d}\leq\frac{1}{p}\frac{|X|^p}{c^p}+\frac{1}{q}\frac{|Y|^q}{d^q}.$$

对上式两边同时取期望有

$$\frac{E(|XY|)}{cd}\leq\frac{1}{p}\frac{E(|X|^p)}{c^p}+\frac{1}{q}\frac{E(|Y|^q)}{d^q}=\frac{1}{p}+\frac{1}{q}=1,$$

从而完成证明. □

6.3 Chernoff不等式

6.3.1 矩生成函数(Moment Generating Function)

首先给出随机变量的矩生成函数定义为:

关于t的函数.

定义6.1. 定义随机变量X的矩生成函数为$M_X(t)=E[e^{tX}]$.

下面给出关于矩生成函数的一些性质:

定理6.3. 设随机变量X的矩生成函数为$M_X(t)$, 对$\forall n\geq 1$, 有

$$E[X^n]=M_X^{(n)}(0),$$

→ 同时被定义为随机变量X的一阶矩

这里$M_X^{(n)}(t)$表示矩生成函数在$t=0$的n阶导数, 而$E[X^n]$被称为随机变量X的n阶矩(*moment*).

Proof. 首先由Tayler公式有

$$e^{tX}=\sum_{i=1}^{\infty}\frac{(tX)^i}{i!}.$$

两边取期望有

$$E[e^{tX}]=\sum_{i=1}^{\infty}\frac{t^i}{i!}E[X^i].$$

对上式两边分别对t求n阶导数、并取$t=0$, 有$M_X^{(n)}(t)=E[X^n]$. □

→ 类似分布函数、密度函数性质.

定理6.4. 对随机变量X,Y, 如果存在常数$\delta>0$, 当$t\in(-\delta,\delta)$时有$M_X(t)=M_Y(t)$成立, 那么X与Y有相同的分布.

上述定理表明随机变量的矩生成函数可唯一确定随机变量的分布, 其证明超出了本书的范围. 若随机变量X与Y独立, 则有

$$M_{X+Y}(t)=E[e^{(X+Y)t}]=E[e^{tX}e^{tY}]=E[e^{tX}]\cdot E[e^{tY}]=M_X(t)M_Y(t).$$

于是得到

推论6.2. 对独立随机变量X和Y, 有$M_{X+Y}(t)=M_X(t)M_Y(t)$.

6.3.2 Chernoff方法

作用：当只知道变量X的一些性质，(比如它是哪些随机变量的和)，而X的具体分布却很难求解时，可用Chernoff方法进行估计

给定任意随机变量X, 以及任意$t>0$和$\epsilon>0$, 利用Markov不等式有

$$\Pr[X\geq\epsilon]=\Pr[e^{tX}\geq e^{t\epsilon}]\leq e^{-t\epsilon}E[e^{tX}].$$

特别地, 有

$$\boxed{\Pr[X\geq\epsilon]\leq\min_{t>0}[e^{-t\epsilon}E[e^{-tX}]].}$$

类似地, 对$\forall\epsilon>0$, $t<0$, 有

$\Pr[e^{tX}\leq e^{t\epsilon}]$

$$\Pr[X\leq\epsilon]=\Pr[tX\geq t\epsilon]\leq e^{-t\epsilon}E[e^{tX}].$$

同样有

$$\boxed{\Pr[X\leq\epsilon]\leq\min_{t<0}[e^{-t\epsilon}E[e^{tX}]].}$$

上述方法称为'**Chernoff方法**', 是证明集中不等式的重要方法. 下面针对特定的分布或特定的条件, 可解$E[e^{tX}]$, 进而求解最小时t的取值.

6.3.3 二值随机变量和Chernoff不等式

定理6.5. 设$X_1, X_2, \cdots, X_n$是n个独立的Bernoulli随机变量，且满足$X_i \sim Ber(p_i)$. 令$X = \sum_{i=1}^n X_i$, $\mu = \sum_{i=1}^n p_i$, 则有 $\mu = \sum_{i=1}^n p_i = \sum_{i=1}^n E[X_i]$

- 对$\forall \epsilon > 0$, 有
$$\Pr[X-\mu \geq \epsilon\mu] = \Pr[X \geq (1+\epsilon)\mu] < \left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}}\right)^\mu.$$
- 对$\forall \epsilon \in (0,1)$, 有
$$\Pr[X-\mu \geq \epsilon\mu] = \Pr[X \geq (1+\epsilon)\mu] \leq e^{-\mu\epsilon^2/3}.$$

X: 能被实验观测到
μ: 不能被实验观测到
当$n \to \infty$时,
$\Pr[X \geq (1+\epsilon)\mu] \to 0$
即随机试验次数足够多时,
二者越来越接近.

上述第一个不等式给出了最紧的不等式上界, 第二个不等式是第一个不等式的适当放松.

Proof. 对任意$t > 0$ 根据Chernoff方法有
$$\Pr[X \geq (1+\epsilon)\mu] = \Pr[e^{tX} \geq e^{t(1+\epsilon)\mu}] \leq e^{-t(1+\epsilon)\mu} E[e^{tX}].$$
利用随机变量的独立性以及$1 + x \leq e^x$, 有

独立性.
$$E[e^{tX}] = E[e^{\sum_{i=1}^n tX_i}] = \prod_{i=1}^n E[e^{tX_i}]$$
$$= \prod_{i=1}^n [(1-p_i) + p_i e^t] = \prod_{i=1}^n [1 + p_i(e^t - 1)] \leq \exp\left(\sum_{i=1}^n p_i(e^t-1)\right) = \exp(\mu(e^t-1)).$$
直接利用0-1分布

由此可得
$$\Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(-t(1+\epsilon)\mu + \mu(e^t-1)\right). = e^{-\mu(1+\epsilon)t+\mu(e^t-1)}$$
对上式求最小值解得$t_{\min} = \ln(1+\epsilon)$, 代入后得到

chernoff方法
$$\Pr[X \geq (1+\epsilon)\mu] \leq \left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}}\right)^\mu.$$

对定理中第二个不等式, 当$\epsilon \in (0,1)$, 我们只需要证明
$$f(\epsilon) = \ln\left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}}\right) + \frac{\epsilon^2}{3} = \epsilon - (1+\epsilon)\ln(1+\epsilon) + \frac{\epsilon^2}{3} \leq 0.$$
易知$f(0) = 0$和$f(1) < 0$. 当$\epsilon \in (0,1)$,
$$f'(\epsilon) = -\ln(1+\epsilon) + 2\epsilon/3, \quad f''(\epsilon) = -\frac{1}{1+\epsilon} + \frac{2}{3}.$$
于是得到$f'(0) = 0$, $f'(1) = -0.0265 < 0$和$f'(1/2) = -0.0721 < 0$, 由连续函数性质有$f'(\epsilon) \leq 0$, 即函数$f(\epsilon)$在$[0,1]$上单调递减. 当$\epsilon \geq 0$ 时有$f(\epsilon) \leq f(0) = 0$, 所以$\exp(f(\epsilon)) \leq 1$. □

下面的定理给出了$\Pr[X \leq (1-\epsilon)\mu]$的估计, 证明与前面的定理证明类似, 作为练习题留给学生自己完成.

同理: $\Pr[X \leq (1-\epsilon)\mu] = \Pr[tX \geq (1-\epsilon)\mu t] = \Pr[e^{tX} \geq e^{(1-\epsilon)\mu t}]$
$\leq e^{-\mu t(1-\epsilon)} E[e^{tX}] = e^{-\mu t(1-\epsilon)} \cdot e^{\mu(e^t-1)} = e^{\mu(e^t-1)-\mu t(1-\epsilon)}$
$\leq e^{-\epsilon\mu - \mu(1-\epsilon)\ln(1-\epsilon)} = \left(\frac{1}{e^\epsilon(1-\epsilon)^{1-\epsilon}}\right)^\mu, \ (\epsilon \in (0,1)).$

定理6.6. 设$X_1, X_2, \cdots, X_n$是n个独立的Bernoulli随机变量，且满足$X_i \sim Ber(p_i)$. 设$X = \sum_{i=1}^n X_i$, $u = \sum_{i=1}^n p_i$, 对$\forall \epsilon \in (0,1)$, 有

$$\Pr[X \leq (1-\epsilon)\mu] < \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{(1-\epsilon)}}\right)^{\mu} \leq \exp(-\mu\epsilon^2/2).$$

对于更为特殊的随机变量$X \in \{+1, -1\}$, 且满足

$$\Pr(X = +1) = \Pr(X = -1) = 1/2,$$

我们有如下定理：

定理6.7. 设$X_1, X_2, \cdots, X_n$是n个独立同分布随机变量，满足$\Pr(X_i = 1) = \Pr(X_i = -1) = 1/2$, 则有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp(-n\epsilon^2/2), \qquad \Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \leq -\epsilon\right) \leq \exp(-n\epsilon^2/2).$$

Proof. 根据$\exp(t)$和$\exp(-t)$的Taylor展开式有

很牛

$$\frac{1}{2}\exp(t) + \frac{1}{2}\exp(-t) = \sum_{i\geq 0}\frac{t^{2i}}{(2i)!} \leq \sum_{i \geq 0}\frac{(t^2/2)^i}{i!} = \exp(t^2/2).$$

$\frac{1}{2}e^t + \frac{1}{2}e^{-t} \leq e^{\frac{t^2}{2}}$

对随机变量$X \in \{+1, -1\}$且满足$\Pr(X = 1) = \Pr(X = -1) = 1/2$, 有

$$E[e^{tX}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} \leq \exp(t^2/2).$$

对任意$t > 0$, 根据Chernoff方法有

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp(-nt\epsilon)E\left(\exp(\sum_{i=1}^n tX_i)\right) = \exp(-nt\epsilon)\prod_{i=1}^n E\left(\exp(tX_i)\right) \leq \exp(-nt\epsilon + nt^2/2).$$

通过对上式右边求最小值解得$t = \epsilon$, 带入上式得到

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \epsilon\right) \leq \exp(-n\epsilon^2/2).$$

同理证明另外一个不等式. □

6.3.4 有界随机变量和Chernoff不等式

→ 本质上是在变量有界且未知变量分布的情况下对$E[e^{tX}]$的估计.

本小节研究随机变量$X_i \in [a, b]$, 其对应的Chernoff不等式. 首先介绍著名的Chernoff引理

引理6.4. 设随机变量$X \in (0, 1)$的期望$\mu = E[x]$. 对任意$t > 0$, 有

$$E[e^{tX}] \leq \exp(t\mu + t^2/8).$$

Proof. 由凸函数的性质可知 $X \in [0,1]$ 对左式两边同时取期望.

$$e^{tX} \le Xe^t + (1-X)e^0 \Rightarrow E(e^{tX}) \le 1-\mu+\mu e^t = \exp(\ln(1-\mu+\mu e^t)) \tag{2}$$

令$f(t) = \ln(1-\mu+\mu e^t)$, 有$f(0) = 0$, 以及

$$f'(t) = \frac{\mu e^t}{1-\mu+\mu e^t} \Rightarrow f'(0) = \mu.$$

进一步有

$$f''(t) = \frac{\mu e^t}{1-\mu+\mu e^t} - \frac{\mu^2 e^{2t}}{(1-\mu+\mu e^t)^2} \le 1/4.$$

a a^2

根据泰勒中值定理有

$$f(t) = f(0) + tf'(0) + f''(\xi)t^2/2 \le t\mu + t^2/8.$$

引理得证. □

由上面的Chernoff引理进一步推导出

作变量代换 $Y = \frac{x-a}{b-a} \in [0,1]$即可

推论6.3. 设随机变量$X \in (a,b)$的期望$\mu = E[x]$. 对任意$t > 0$, 有

$$E(e^{tX}) \le \exp(\mu t + t^2(b-a)^2/8).$$

根据上述推论, 我们得到有界随机变量的Chernoff不等式:

定理6.8. 假设$X_1, \ldots, X_n$是n独立的随机变量且满足$X_i \in (a,b)$. 对任意$\epsilon > 0$, 我们有

$$\begin{aligned}
\Pr\left[\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n E[X_i] \ge \epsilon\right] &\le \exp(-2n\epsilon^2/(b-a)^2),\\
\Pr\left[\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n E[X_i] \le -\epsilon\right] &\le \exp(-2n\epsilon^2/(b-a)^2).
\end{aligned}$$

Proof. 这里给出第一个不等式的证明, 第二个不等式证明将作为习题. 对任意$t > 0$, 根据Chernoff方法我们有

$$\begin{aligned}
&\Pr\left[\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n E[X_i] \ge \epsilon\right] = \Pr\left[\sum_{i=1}^n t(X_i - E[X_i]) \ge nt\epsilon\right]\\
\le\ & \exp(-nt\epsilon)E\left[\exp\left(\sum_{i=1}^n t(X_i - E[X_i])\right)\right] = \exp(-nt\epsilon)\prod_{i=1}^n E\left[\exp(t(X_i - E[X_i]))\right].
\end{aligned}$$

根据Chernoff引理以及简单整理可得, 对任意$X_i \in [a,b]$有

$$E\left[\exp(t(X_i - E[X_i]))\right] \le \exp((b-a)^2t^2/8).$$

由此得到

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n E[X_i] \ge \epsilon\right] \le \exp(-nt\epsilon + nt^2(b-a)^2/8).$$

对上式右边取最小值求解$t = 4\epsilon/(b-a)^2$, 然后带入上式可得:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geq \epsilon\right] \leq \exp(-2n\epsilon^2/(b-a)^2).$$

从而完成证明. □

6.4 Gaussian和Sub-Gaussian随机变量不等式

讨论无界变量：高斯变量之和的估计.

首先考虑独立同分布的Gaussian随机变量:

→不再借助chernoff方法求解$E[e^t]$，而直接利用Gauss分布的性质

定理6.9. 设$X_1, \ldots, X_n$是n独立同分布的随机变量, 且服从$X_i \sim \mathcal{N}(\mu, \sigma)$. 对任意$\epsilon > 0$, 有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geq \epsilon\right] = \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \leq -\epsilon\right] \leq \frac{1}{2}\exp(-n\epsilon^2/2\sigma^2).$$

Proof. 如果$X_i \sim \mathcal{N}(\mu, \sigma)$, 根据正太分布的性质有

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \sim \mathcal{N}(0, \sigma^2/n) \quad \Rightarrow \quad \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i - \mu) \sim \mathcal{N}(0, 1).$$

若$X' \sim \mathcal{N}(0,1)$, 对任意$\epsilon > 0$, 根据以前的定理有

$$P(X' \geq \epsilon) \leq \frac{1}{2}e^{-\epsilon^2/2}.$$

$$\begin{aligned} Pr(X' \geq \epsilon) &= \int_{\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(t+\epsilon)^2/2} dt' \\ &\leq \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2+\epsilon^2}{2}} dt \\ &= \frac{1}{2} \cdot e^{-\frac{\epsilon^2}{2}} \end{aligned}$$

因此得到

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geq \epsilon\right] = \Pr\left[\frac{1}{\sqrt{n}\sigma}\sum_{i=1}^{n}(X_i - \mu) \geq \epsilon\sqrt{n}/\sigma\right] \leq \frac{1}{2}\exp(-n\epsilon^2/2\sigma^2).$$

定理得证. □

下面研究Sub-Gaussian随机变量, 可以将有界随机变量和Gaussian随机变量统一起来:

定义6.2. 对任意$t \in \mathcal{R}$, 如果随机变量X满足

$$E[e^{(X-\mu)t}] \leq \exp(bt^2/2),$$

则称随机变量X是服从参数为b的亚高斯(*Sub-Gaussian*)随机变量.

直观而言：亚高斯分布随机变量表示随机变量的尾部分布不会比一个高斯分布更严重.

例6.2. 对任意有界的随机变量 $X \in [a,b]$, 根据 *Chernoff* 引理有

$$E[e^{(X-\mu)t}] \leq \exp(t^2(b-a)^2/8),$$

因此有界随机变量是服从参数为$(b-a)^2/4$的亚高斯型随机变量.

☆ 高斯函数积分的统一思路：都能把变量在概率密度函数的指数上集中，然后通过换元，使概率密度函数的积分上下限变为 $-\infty/+\infty$ 或 $0/+\infty$.

例6.3. 如果随机变量X服从高斯分布$\mathcal{N}(\mu,\sigma^2)$, 我们有

配方

$$E[e^{(X-\mu)t}] = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{xt} e^{-x^2/2\sigma^2} dx = e^{\sigma^2 t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(t\sigma - x/\sigma)^2/2} d(x/\sigma) = e^{\sigma^2 t^2/2}.$$

因此*Gaussian*型随机变量是服从参数为σ^2的亚高斯型随机变量.

由前面的两个例子可知高斯随机变量和有界随机变量都是亚高斯随机变量. 对于亚高斯型随机变量, 我们有如下不等式:

定理6.10. 假设$X_1,\dots,X_n$是独立地、服从参数为b的亚高斯随机变量, 对任意$\epsilon>0$, 有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \ge \epsilon\right] \le \exp(-n\epsilon^2/2b), \qquad \Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \le -\epsilon\right] \le \exp(-n\epsilon^2/2b).$$

$\delta \to \epsilon = \sqrt{\frac{2b\ln\frac{1}{\delta}}{n}}$

Proof. 对任意$t>0$, 根据Chernoff方法有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \ge \epsilon\right] \le e^{-tn\epsilon} \prod_{i=1}^n E[e^{(X_i-\mu)t}] \le e^{-tn\epsilon + nbt^2/2}$$

通过求解上式最小值可得$t_{\min}=\epsilon/b$, 从而完成证明. □

也就是以至少 $1-\delta$ 的概率 $\frac{1}{n}\sum_{i=1}^n X_i - E[X_i] < \sqrt{\frac{2b\ln\frac{1}{\delta}}{n}}$

对于亚高斯型随机变量, 我们还可以估计最大值期望的不等式:

定理6.11. 假设$X_1,\dots,X_n$是相互独立地、服从参数为b的亚高斯随机变量, 且满足$E[X_i]=0$. 我们有

$$E\left[\max_{i\in[n]} X_i\right] \le \sqrt{2b\ln n}.$$

Proof. 对任意$t>0$, 根据Jensen不等式有

$$\exp\left(tE\left[\max_{i\in[n]} X_i\right]\right) \le E\left[\exp\left(t\max_{i\in[n]} X_i\right)\right] = E\left[\max_{i\in[n]} \exp(tX_i)\right] \le \sum_{i=1}^n E[\exp(tX_i)] \le n\exp(t^2 b/2).$$

对上式两边同时取对数和除以t有

$$E\left[\max_{i\in[n]} X_i\right] \le \frac{\ln n}{t} + \frac{bt}{2}.$$

通过求解上式最小值可得$t_{\min}=\sqrt{2\ln n/b}$, 从而完成证明. □

前面所讲的概率不等式, 可以以另外一种表达形式给出, 这里以定理 6.10为例, 假设$X_1,\dots,X_n$是独立地、服从参数为b的亚高斯随机变量, 对任意$\epsilon>0$, 有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \ge \epsilon\right] \le \exp(-n\epsilon^2/2b).$$

令$\delta=\exp(-n\epsilon^2/2b)$, 求解出

$$\epsilon = \sqrt{2b\ln(1/\delta)/n},$$

带入得到下面的不等式至少以$1-\delta$的概率成立：

$$\frac{1}{n}\sum_{i=1}^n (X_i-\mu) \le \sqrt{\frac{2b}{n}\ln\frac{1}{\delta}}.$$

所有不等式都可以采用$1-\delta$的形式描述.

6.5 Bennet和Bernstein不等式

6.5以下的证明完整式均为单边.

通过考虑随机变量的方差, 可能推导出更紧地集中不等式, 下面介绍两种基于方差的集中不等式.

定理6.12 (Bennet不等式). 假设$X_1, \ldots, X_n$是n独立同分布的随机变量且满足$X_i - E[X_i] \le 1$. 设随机变量的均值为μ和方差为σ^2, 则有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\ge\epsilon\right]\le\exp(-n\epsilon^2/(2\sigma^2+2\epsilon/3)).$$

Proof. 对任意$t>0$, 根据Chernoff方法有

独立同分布性质.

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\ge\epsilon\right]\le e^{-nt\epsilon}E\left[\exp\left(\sum_{i=1}^{n}t(X_i-\mu)\right)\right]=e^{-nt\epsilon}\left(E[e^{t(X_1-\mu)}]\right)^n$$

设$Y=X_1-\mu$, 利用公式$\ln z\le z-1$得到

$\ln x \le x-1$ $E[Y]=0$

$$\begin{aligned}\ln E[e^{t(X_1-\mu)}] &= \ln E[e^{tY}]\le E[e^{tY}]-1=t^2E\left[\frac{e^{tY}-tY-1}{t^2Y^2}Y^2\right]\\ &\le t^2E\left[\frac{e^t-t-1}{t^2}Y^2\right]=(e^t-t-1)\sigma^2\end{aligned}$$

$\boxed{\because Y=X_1-\mu\in[0,1] \quad \therefore Yt\le t, \text{代换即可}}$

这里利用$tY\le t$以及$(e^z-z-1)/z^2$是一个非单调递减的函数. 进一步有

泰勒展开+放缩.

$$e^t-t-1\le\frac{t^2}{2}\sum_{k=0}^{\infty}(t/3)^k=\frac{t^2}{2(1-t/3)}.$$

因此可得

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\ge\epsilon\right]\le\exp\left(-nt\epsilon+\frac{nt^2\sigma^2}{2(1-t/3)}\right).$$

忽视项目的估计.

猜出$t=\epsilon/(\sigma^2+\epsilon/3)$, 带入完成证明. □

对于Bennet指数不等式, 设

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\ge\epsilon\right]\le\exp(-n\epsilon^2/(2\sigma^2+2\epsilon/3))=\delta$$

其另外一种表述为: 至少以$1-\delta$的概率有 消去ϵ,得到δ,用概率的形式讨论.

$$\frac{1}{n}\sum_{i=1}^{n}X_n\le\mu+\frac{2\ln 1/\delta}{3n}+\sqrt{\frac{2\sigma^2}{n}\ln\frac{1}{\delta}}.$$

当σ^2非常小, 或趋于0时, 得到更紧的收敛率$\bar{X}_n\le\mu+O(1/n)$.

下面考虑另一种基于方差的不等式, 与Bennet不等式不同之处在于约束随机变量的势:

定理6.13 (Bernstein不等式). 假设$X_1, \ldots, X_n$是n独立同分布的随机变量. 设随机变量的均值为μ, 以及方差为σ^2. 如果存在常数$b>0$, 使得对任意$m\ge 2$, 有$E[X_i^m]\le m!b^{m-2}\sigma^2/2$成立, 那么有

(对于$\epsilon\in(0,\frac{1}{2})$)

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\ge\epsilon\right]\le\exp\left(-\frac{n\epsilon^2}{2\sigma^2+2b\epsilon}\right).$$

TODO

Proof. 对任意$t>0$, 根据Chernoff方法有

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\geq\epsilon\right]\leq e^{-nt\epsilon}E\left[\exp\left(\sum_{i=1}^{n}(X_i-\mu)\right)\right]=e^{-nt\epsilon-n\mu t}\left(E[e^{tX_1}]\right)^n$$

利用公式$\ln z\leq z-1$有

代入矩函数的条件,

$$\ln E[e^{tX_1}]\leq E[e^{tX}]-1=\sum_{m=1}^{\infty}E[X^m]\frac{t^m}{m!}\leq t\mu+\frac{t^2\sigma^2}{2}\sum_{m=2}^{\infty}(bt)^{m-2}=t\mu+\frac{t^2\sigma^2}{2(1-bt)}.$$

由此得到

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\geq\epsilon\right]\leq=\exp\left(-nt\epsilon+\frac{t^2n\sigma^2}{2(1-bt)}\right)$$

取$t=\epsilon/(V+b\epsilon)$完成证明. □

Bennet 和 Berstein 区别在于:
Bennet 基于 $X_i-E[X_i]$ 的约束
而 Berstein 基于矩函数性质来对 $E[e^t]$ 进行放缩

习题. 给出Bernstein不等式的$1-\delta$表述.

6.6 应用: 随机投影(Random Projection)

本节介绍不等式的一种应用: 随机投影. 假设$\mathbb{R}^d$维空间有n个点: $\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n$ (d非常大, 例如大约100万). 我们能否找到一个保距变换: $f:\mathbb{R}^d\to\mathbb{R}^k$ ($k\ll d$), 使得以较大概率有

降维.

$$(1-\epsilon)\|\bar{x}_i-\bar{x}_j\|_2^2\leq\|f(\bar{x}_i)-f(\bar{x}_j)\|_2^2\leq(1+\epsilon)\|\bar{x}_i-\bar{x}_j\|_2^2.$$

映射后的距离不能偏离太多

随机投影广泛应用于高维数据的机器学习问题, 例如最近邻, k-近邻, 降维, 聚类等问题.

行向量形式

随机投影函数可以简单的表示为$f(\bar{x})=\bar{x}P/c$, 其中P为$d\times k$的一个随机矩阵, c为一常数(根据随机矩阵P唯一确定). 下面介绍三种常见的随机矩阵

① • $P=(p_{ij})_{d\times k}\in\mathbb{R}^{d\times k}$, $p_{ij}\sim\mathcal{N}(0,1)$, 此时$c=\sqrt{k}$;

② • $P=(p_{ij})_{d\times k}\in\{-1,1\}^{d\times k}$, $\Pr(p_{ij}=1)=\Pr(p_{ij}=-1)=1/2$, 此时$c=\sqrt{k}$; 【Rademacher随机变量】

③ • $P=(p_{ij})_{d\times k}\in\{-1,0,1\}^{d\times k}$, $\Pr(p_{ij}=0)=2/3$, $\Pr(p_{ij}=1)=\Pr(p_{ij}=-1)=1/6$, 此时$c=\sqrt{k/3}$. 【主要用于sparse投影, 减少计算量】

生成更加稀疏的降维结果

下面我们重点理论分析 Gaussian 随机变量, 其它随机变量可参考相关资料, 对 Gaussian 随机变量, 这里介绍著名的 Johnson–Lindenstrauss 引理, 简称 JL 引理.

引理6.5. 设$\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n\in\mathbb{R}^d$, 随机矩阵$P=(p_{ij})_{d\times k}\in\mathbb{R}^{d\times k}$, 其元素$p_{ij}\sim\mathcal{N}(0,1)$, 设

$$\bar{y}_i=f(\bar{x}_i)=\bar{x}_iP/\sqrt{k},\qquad i\in[n]$$

将d维空间中n个点$\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n$通过随机矩阵$P$投影到$k$维空间($k\ll d$). 对任意$\epsilon\in(0,1/2)$, 如果$k\geq 8\log 2n/(\epsilon^2-\epsilon^3)$, 则至少以1/2的概率有

$$(1-\epsilon)\|\bar{x}_i-\bar{x}_j\|_2^2\leq\|\bar{y}_i-\bar{y}_j\|_2^2\leq(1+\epsilon)\|\bar{x}_i-\bar{x}_j\|_2^2\qquad(i,j\in[n]).$$

(概率的大小固定, 对 k 和 ε 的关系作出约束.
k 越大, 降维后的距离差距就越小.

由于 $k\geq 8\log 2n/\varepsilon^2-\varepsilon^3$, 因此在 $n\in d$ (事实上通常情况下有 $n\ll d$) 时, $k\sim\log d$.
也就是降维可降维至 $\log d$ 的规模.

P是随机变量构成的矩阵.

Proof. 下面分三步证明 J-L 引理.

第一步: 对任意非零 $\bar{x}=(x_1,x_2,\cdots,x_d)\in\mathbb{R}^d$, 我们需证明: $E[\|\bar{x}P\|_2^2]=\|\bar{x}\|_2^2$.

设$\bar{y}=\bar{x}P$, 根据$P=(p_{ij})_{d\times k}$ $(p_{ij}\sim\mathcal{N}(0,1))$, 我们有

其实是很显然的结论.

$$E\left[\|\bar{y}\|_2^2\right]=E\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{d}\frac{x_ip_{ij}}{\sqrt{k}}\right)^2\right]=\sum_{j=1}^{k}\frac{1}{k}E\left[\left(\sum_{i=1}^{d}x_ip_{ij}\right)^2\right]=\sum_{j=1}^{k}\frac{1}{k}\sum_{i=1}^{d}x_i^2=\frac{1}{k}\sum_{j=1}^{k}\|\bar{x}\|_2^2=\|\bar{x}\|_2^2.$$

$= \sum_{j=1}^{k}\frac{1}{k}\sum_{i=1}^{d}x_i^2E[p_{ij}^2]$

即在期望的情况下, $\bar{x}$与$\bar{y}$具有相等的到原点的距离.

$= \sum_{j=1}^{k}\frac{1}{k}\cdot\sum_{i=1}^{d}x_i^2 = \sum_{j=1}^{k}\frac{1}{k}\cdot\|x\|_2^2 = \|x\|_2^2$

第二步: 证明: $\Pr[\|\bar{y}\|_2^2\ge(1+\epsilon)\|\bar{x}\|_2^2]\le\exp(-(\epsilon^2-\epsilon^3)k/4)$.

为了简洁起见, 我们将矩阵表示为$P=(P^1,P^2,\ldots,P^k)$, 其中$P^i(i\in[d])$是一个$d\times 1$列向量的列向量. 因此我们有

$$\bar{y}=(\bar{x}P^1,\bar{x}P^2,\ldots,\bar{x}P^k)/\sqrt{k}.$$

设

$$v=k\frac{\|\bar{y}\|_2^2}{\|\bar{x}\|_2^2}=\frac{1}{\|\bar{x}\|_2^2}\sum_{j=1}^{k}(\bar{x}P^j)^2=\sum_{j=1}^{k}\left(\frac{\bar{x}}{\|\bar{x}\|_2}P^j\right)^2=\sum_{j=1}^{k}v_j^2$$

这里我们设$v_j=\bar{x}P^j/\|\bar{x}\|_2$, 由于$P^j$中每个元素服从$\mathcal{N}(0,1)$, 由高斯分布的性质有$v_j\sim\mathcal{N}(0,1)$. 对任意$t\in(0,1/2)$, 根据 Chernoff 方法有

$$\Pr[\|\bar{y}\|_2^2\ge(1+\epsilon)\|\bar{x}\|_2^2]=\Pr[v\ge(1+\epsilon)k]=\Pr[e^{tv}\ge e^{(1+\epsilon)kt}]\le e^{-(1+\epsilon)kt}E[e^{tv}]=e^{-(1+\epsilon)kt}\left(E[e^{tv_1^2}]\right)^k.$$

进一步有

$$E[e^{tv_1^2}]=\int_{-\infty}^{+\infty}\frac{e^{tu^2}}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du=\int_{-\infty}^{+\infty}\frac{e^{-\frac{u^2}{2}(1-2t)}}{\sqrt{2\pi}}du=\frac{1}{\sqrt{1-2t}}.$$

于是得到

$$\Pr\left[\|\bar{y}\|_2^2\ge(1+\epsilon)\|\bar{x}\|_2^2\right]\le\left(e^{-2(1+\epsilon)t}/(1-2t)\right)^{k/2}.$$

对上式右边t求最小解得$t_{\min}=\frac{\epsilon}{2(1+\epsilon)}$, 代入可得

$$\Pr[\|\bar{y}\|_2^2\ge(1+\epsilon)\|\bar{x}\|_2^2]\le[(1+\epsilon)e^{-\epsilon}]^{\frac{k}{2}}.$$

由$\epsilon\in(0,1/2)$, 我们设$f(\epsilon)=\ln(1+\epsilon)$, 并有

$$f^{'}(\epsilon)=\frac{1}{1+\epsilon},f^{''}(\epsilon)=-\frac{1}{(1+\epsilon)^2},f^{'''}(\epsilon)=\frac{2}{(1+\epsilon)^3}\le 2.$$

根据泰勒中值定理有

$$f(\epsilon)=f(0)+f^{'}(0)\epsilon+\frac{f^{''}(0)\epsilon^2}{2!}+\frac{f^{'''}(\xi)\epsilon^3}{3!}=\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^2}{3}\le-\frac{\epsilon^2-\epsilon^3}{2}.$$

于是得到

$$\Pr[\|\bar{y}\|_2^2\ge(1+\epsilon)\|\bar{x}\|_2^2]\le e^{-k(\epsilon^2-\epsilon^3)/4}.$$

同理可证明$\Pr[\|\bar{y}\|_2^2\le(1-\epsilon)\|\bar{x}\|_2^2]\le e^{-2k(\epsilon^2-\epsilon^3)/4}$.

第三步： 对于任意给定的$i \neq j$, 由第二步可知

$$\Pr[\|y_i - y_j\|_2^2 \geq (1+\epsilon)\|x_i - x_j\|_2^2] \leq e^{-k(\epsilon^2-\epsilon^3)/4}, \quad \Pr[\|y_i - y_j\|_2^2 \leq (1-\epsilon)\|x_i - x_j\|_2^2] \leq e^{-k(\epsilon^2-\epsilon^3)/4}.$$

由于$i \in [n], j \in [n]$, 因此共有$n(n-1)$对(i,j), 根据 Union 不等式有,

$$\Pr\left[\exists (i,j) \colon \|y_i - y_j\|_2^2 \geq (1+\epsilon)\|x_i - x_j\|_2^2 \quad \text{或} \quad \|y_i - y_j\|_2^2 \leq (1-\epsilon)\|x_i - x_j\|_2^2\right] \leq 2n^2 e^{-k(\epsilon^2-\epsilon^3)/4} \leq \frac{1}{2}.$$

根据$2n^2 e^{-k(\epsilon^2-\epsilon^3)/4} \leq 1/2$, 求解出$k \geq 8 \log 2n/(\epsilon^2 - \epsilon^3)$. 引理得证. □

chernoff 方法

- $x_i \in \{0,1\}$ $\{-1,1\}$
- $x_i \in [a,b]$ chernoff 引理.
- Gaussian
- 小方差情况下: Bennet, Bernstein

应用:

- $x \in R^d \longrightarrow y \in R^k$
- 保距降维: $y = x \cdot P/c$