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

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

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

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

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

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

Proof. 根据Markov不等式有

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

□

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

$$P(X-\mu\geq\epsilon)\leq\frac{\sigma^2}{\sigma^2+\epsilon^2}\quad\text{和}\quad P(X-\mu\leq-\epsilon)\leq\frac{\sigma^2}{\sigma^2+\epsilon^2}.$$

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

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

设$u=\sigma^2/\epsilon$, 由此得到

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

另一方面, 对任意$u>0$, 有

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

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$X_1,X_2,\ldots,X_n$, 如果满足$E(X_i)=\mu$和$Var(X_i)\leq\sigma^2$, 则对任意$\epsilon>0$, 有

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

Proof. 根据Chebyshev不等式有

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

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:

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

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

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$M_X^{(n)}(t) = E[X^n]$. □

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

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

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

类似地, 对$\forall \epsilon > 0$, $t < 0$, 有

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

同样有

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

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

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

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
$$\begin{aligned}
E[e^{tX}] &= E[e^{\sum_{i=1}^n tX_i}] = \prod_{i=1}^n E[e^{tX_i}] \\
&= \prod_{i=1}^n [(1-p_i) + p_i e^t] = \prod_{i=1}^n [1 + p_i(e^t - 1)] \leq \exp\left(\sum_{i=1}^n p_i(e^t - 1)\right) = \exp(\mu(e^t - 1)).
\end{aligned}$$
由此可得
$$\Pr[X \geq (1+\epsilon)\mu] \leq \exp\left(-t(1+\epsilon)\mu + \mu(e^t - 1)\right).$$
对上式求最小值解得$t_{\min} = \ln(1+\epsilon)$, 代入后得到
$$\Pr[X \geq (1+\epsilon)\mu] \leq \left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}}\right)^\mu.$$

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$\Pr[X \geq (1-\epsilon)\mu]$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.

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$X = \sum_{i=1}^n X_i$, $\mu = \sum_{i=1}^n p_i$, 对$\forall \epsilon \in (0,1)$, 有

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

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

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

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

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

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

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

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

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$X_1,\dots,X_n$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. □