机器学习理论研究导引

2019年1月

目录

1 基础不等式

1.1 引言

机器学习研究如何从已有的‘经验数据’中学习得到能有效处理‘未见数据’的模型. 在学习理论中, 如何形式化理解机器学习的定义. 这里以二分类问题为例, 考虑示例空间$\mathcal{X} \subseteq \mathbb{R}^d$以及标记空间$\mathcal{Y} = \{0, 1\}$, 假设$\mathcal{D}$是空间$\mathcal{X} \times \mathcal{Y}$的一个联合分布. 这里的联合分布$\mathcal{D}$是未知的, 用于形式化刻画机器学习定义中的‘未见数据’. 在实际任务中真实数据分布$\mathcal{D}$不可知, 但可以假设从分布$\mathcal{D}$中独立采样得到训练数据集$D_m = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_m, y_m)\}$, 训练数据用于形式化刻画机器学习定义‘经验数据’. 学习模型一般形式化描述为函数$f\colon \mathcal{X} \to \{0, 1\}$, 又被称为‘假设(hypothesis)’或‘分类器’. 为了衡量模型的分类性能, 引入一个指示函数$\mathbb{I}(\cdot)$, 当论断为真时其返回值为1, 否则为0.

当给定一个模型f后, 可以得到此模型在训练数据集D_n的分类错误率为

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

称$\hat{R}(f, D_m)$为‘训练错误率’. 同理可以定义模型f在整个数据分布$\mathcal{D}$的分类错误率为

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

称$R(f)$为‘期望错误率’. 值得注意的是: 数据分布$\mathcal{D}$是潜在未知的, 从而导致期望错误率$R(f)$不可知, 我们可知的信息是训练错误率$\hat{R}(f, D_n)$. 能否从已知的训练错误率$\hat{R}(f, D_n)$中推导出未知期望错误率$R(f)$的一些信息, 或者能将未知期望错误率$R(f)$ 限制在一定范围呢？即

$$|\hat{R}(f, D_m) - R(f)| \leq ?$$

这里不妨假设随机变量$X_i = \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i)$, 因此上述问题可以进一步抽象描述为: 假设有$m$个独立同分布的随机变量$X_1, X_2, X_m$, 需要从这$m$ 个独立同分布的随机变量中获得其期望$E[X]$的一个估计, 即

$$\left| \frac{1}{m} \sum_{i=1}^{m} X_i - E(X_i) \right| \leq ?$$

本章的组织结构如下: 1.2节介绍不等式的一些基本知识,1.3 节介绍Concentration不等式,1.4节介绍Martingale不等式.

1.2 不等式基础知识

首先介绍几个基础的概率不等式:

- 对任意常数a_1, a_2和随机变量X_1, X_2, 有

$$E(a_1 X_1 + a_2 X_2) = a_1 E(X_1) + a_2 E(X_2).$$

- (Union不等式)对任意概率事件A, B, 有
$$P(A \cup B) \leq P(A) + P(B).$$
- (Bayes公式)对任意概率事件A, B, 有
$$P(A|B) = \frac{P(A)}{P(B)} P(B|A).$$
- (全概率公式)假设概率事件$B_1, B_2, \ldots, B_n$构成全概率空间且两两互不相交, 对任意概率事件A, 有
$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i) P(B_i).$$
对于随机变量X, 有
$$E(X) = \sum_{i=1}^{n} E(X|B_i) P(B_i).$$
- 对m个独立的随机变量$X_1, X_2, \ldots, X_m$, 有
$$E\left(\prod_{i=1}^{m} X_i\right) = \prod_{i=1}^{m} E(X_i).$$

下面给出概率统计中常用不等式.

定义1.1 (凸函数). *如果对任何$x_1, x_2 \in \mathbb{R}$以及$t \in [0, 1]$, 有*
$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$
成立, 则称函数$f\colon \mathbb{R} \to \mathbb{R}$是凸函数.

引理1.1 (Jensen不等式). *对于凸函数$f\colon \mathbb{R} \to \mathbb{R}$和随机变量$X$, 有$f(E(X)) \leq E(f(X))$.*

证明. 首先考虑离散型随机变量X有$x_1, x_2, \ldots, x_n$种不同的取值, 其概率分别为$t_1, t_2, \ldots, t_n$, 其中$t_i \in [0, 1]$以及$t_1 + t_2 + \ldots + t_n = 1$. 需要证明
$$f(t_1x_1 + t_2x_2 + \cdots + t_nx_n) \leq t_1f(x_1) + t_2f(x_2) + \cdots + t_nf(x_n). \tag{1}$$
这里将利用数学归纳法证明. 当$n = 2$时, 由凸函数的定义可知
$$f(t_1x_1 + t_2x_2) = f(t_1x_1 + (1-t_1)x_2) \leq t_1f(x_1) + (1-t_1)f(x_2) = t_1f(x_1) + t_2f(x_2)$$
这里利用关系$t_2 = 1 - t_1$. 假设$n = k$时不等式(1)成立, 现证明$n = k + 1$时成立. 令$x' = (t_1x_1 + \cdots + t_kx_k)/(1 - t_{k+1})$, 由凸函数定义有
$$f(t_1x_1 + \cdots + t_kx_k + t_{k+1}x_{k+1}) = f((1-t_{k+1})x' + t_{k+1}x_{k+1}) \leq (1-t_{k+1})f(x') + t_{k+1}f(x_{k+1}),$$
另一方面, 由归纳假设可知$n = k$时成立, 则有
$$f(x') \leq \frac{t_1}{1-t_{k+1}} f(x_1) + \cdots + \frac{t_k}{1-t_{k+1}} f(x_k),$$

从而完成离散情形的证明.

对于连续型随机变量X, 其期望为$E(X)$, 先假设直线$L(x)=a+bx$是函数$f(x)$在$x=E(X)$的切线, 由凸函数的性质可知切线在凸函数的下方, 于是得到

$$E(f(X)) \geq E(L(X)) = a + bE(X) = L(E(X)) = g(E(X)).$$

从而完成证明. ■

由Jensen不等式很容易得到如下推论:

推论1.1. *对有限方差随机变量X, 有$[E(X)]^2 \leq E(X^2)$.*

引理1.2 (Cauchy-Schwartz不等式). *对有限方差随机变量X和Y, 有$E(|XY|) \leq \sqrt{E(X^2)E(Y^2)}$.*

证明. 对随机变量X和Y, 以及任何的实数t, 有

$$(tX+Y)^2 \geq 0 \quad \Rightarrow \quad E((tX+Y)^2) \geq 0 \quad \Leftrightarrow \quad t^2E(X^2) + 2tE(XY) + E(Y^2) \geq 0$$

上述一元二次不等式对任何$t \in \mathbb{R}$恒成立, 因此我们有

$$(2E(XY))^2 - 4E(X^2)E(Y^2) \leq 0$$

从而完成证明. ■

引理1.3 (Hölder不等式). *对任意随机变量X和Y以及实数$p>0, q>0$满足$\frac{1}{p}+\frac{1}{q}=1$, 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.

证明. 对任意实数对任意实数$a>0, b>0$, 根据凸函数性质可以得到Young不等式

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

令$A=(E(|X|^p))^{\frac{1}{p}}$和$B=(E(|Y|^q))^{\frac{1}{q}}$, 根据上面Young不等式有

$$\frac{|X|}{A}\frac{|Y|}{B} \leq \frac{1}{p}\frac{|X|^p}{A^p} + \frac{1}{q}\frac{|Y|^q}{B^q}.$$

对上式两边同时取期望有

$$\frac{E(|XY|)}{AB} \leq \frac{1}{p}\frac{E(|X|^p)}{A^p} + \frac{1}{q}\frac{E(|Y|^q)}{B^q} = \frac{1}{p} + \frac{1}{q} = 1$$

从而完成证明. ■

引理1.4 (Bhatia–Davis不等式). *对任意随机变量$X \in [a,b]$, 假设其期望为$E(X)=\mu$, 则对随机变量X方差的上界为*

$$V(X) \leq (b-\mu)(\mu-a) \leq (b-a)^2/4.$$

证明. 对任意$X \in [a,b]$, 有$(b-X)(X-a) \geq 0$, 从而得到

$$X^2 \leq X(a+b) - ab$$

由此可得

$$V(X) \leq \int_a^b ((a+b)x - ab)p(x)dx - \mu^2 = (a+b)\mu - ab - \mu^2 = (b-\mu)(\mu-a) \leq \frac{(b-a)^2}{4}.$$

从而完成证明. ■

定理1.1 (Etemadi不等式). 假设$X_1, X_2, \ldots, X_n$是n个独立的随机变量, 记$S_k = \sum_{i=1}^k X_i$. 对任何$x > 0$, 有

$$P\left(\max_{k\in[n]} |S_k| \geq 3x\right) \leq 2P(|S_n| \geq x) + \max_{k\in[n]} P(|S_k| \geq x) \leq 3 \max_{k\in[n]} P(|S_k| \geq x)$$

证明. 对于$k \in [n]$, 设随机事件

$$A_k = \left\{\max_{j\in[k-1]} |S_j| < 3x\} \cap \{|S_k| > 3x\right\}$$

以及$A_1 = \{|S_1| > 3x\}$. 随机事件$A_1, A_2, \ldots, A_n$互不相交, 且有

$$A = A_1 \cup A_2 \cup \cdots A_n = \left\{\max_{k\in[n]} |S_k| \geq 3x\right\}.$$

对任意$k \in [n]$有

$$A_k \cap \{S_n < x\} \subseteq A_k \cap \{|S_n - S_k| > 2x\},$$

且事件A_k与$\{|S_n - S_k| > 2x\}$是相互独立的. 由此可得

$$P(A) = P(A \cap \{S_n \geq x\}) + P(A \cap \{S_n < x\}) \leq P(\{S_n \geq x\}) + P(A \cap \{S_n < x\})$$

进一步有

$$\begin{aligned}
P(A \cap \{S_n < x\}) &= P\left(\left(\cup_{k\in[n]} A_k\right) \cap \{S_n < x\}\right) \\
&= P\left(\cup_{k\in[n]} \left(A_k \cap \{S_n < x\}\right)\right) \\
&\leq \sum_{k\in[n]} P\left(A_k \cap \{S_n < x\}\right) \\
&\leq \sum_{k\in[n]} P\left(A_k \cap \{|S_n - S_k| \geq 2x\}\right) \\
&\leq \sum_{k\in[n]} P(A_k) P(|S_n - S_k| \geq 2x) \\
&\leq P(A) \max_{k\in[n]} P(|S_n - S_k| \geq 2x)
\end{aligned}$$

再根据$|a-b| \geq 2x$可以推导出$|a| > x$或$|b| > x$, 由此得到

$$P(A) \leq P(\{S_n \geq x\}) + \max_{k\in[n]} P(|S_n - S_k| \geq 2x) \leq P(\{S_n \geq x\}) + \max_{k\in[n]} \{P(|S_n| \geq x) + P(|S_k| \geq x)\}$$

从而完成证明. ■

定理1.2. 对n个随机变量$X_1, X_2, \ldots, X_n$, 如果存在$\sigma > 0$, 使得对任意$t > 0$都有

$$E(\exp(tX_i)) \leq \exp(t^2\sigma^2/2),$$

那么有

$$E\left(\max_{i\in[n]} X_i\right) \leq \sigma\sqrt{2\log n}$$

证明. 根据Jensen不等式, 对任意$t > 0$, 有

$$\begin{aligned}\exp\left(tE\left(\max_{1\leq i\leq n} X_i\right)\right) &\leq E\left(\exp\left(t\max_{1\leq i\leq n} X_i\right)\right) = E\left(\max_{1\leq i\leq n}\exp\left(tX_i\right)\right)\\ &\leq \sum_{i=1}^{n} E\left(\exp\left(tX_i\right)\right) = n\exp(t^2\sigma^2/2).\end{aligned}$$

对上式两边同时取对数可得

$$E\left(\max_{1\leq i\leq n} X_i\right) \leq \frac{\ln n}{t} + \frac{t\sigma^2}{2}.$$

令$t = \sqrt{2\ln n}/\sigma$, 从而完成证明. ■

对标准正太分布随机变量, 有如下结论:

定理1.3. 假设随机变量X服从标准正太分布, 即$X \sim \mathcal{N}(0, 1)$. 对任意$\epsilon > 0$, 有

$$\frac{1}{3}e^{-(\epsilon+1)^2/2} \leq P(X \geq \epsilon) \leq \frac{1}{2}e^{-\epsilon^2/2}.$$

证明. 标准正太分布的密度函数为

$$p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

于是有上界

$$\begin{aligned}P(X \geq \epsilon) &= \int_{\epsilon}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x+\epsilon)^2/2}dx\\ &\leq \int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x^2+\epsilon^2)/2}dx = \frac{1}{2}e^{-\epsilon^2/2}.\end{aligned}$$

同时得到下界

$$\begin{aligned}P(X \geq \epsilon) &= \int_{\epsilon}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \int_{0}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(x+\epsilon)^2/2}dx\\ &\geq \int_{0}^{1}\frac{1}{\sqrt{2\pi}}e^{-(x+\epsilon)^2/2}dx \geq \frac{1}{3}e^{-(\epsilon+1)^2/2}.\end{aligned}$$

定理得证. ■

对于标准正太分布, 可以得到一个更紧地不等式:

定理1.4 (Mill不等式). 假设随机变量X服从标准正太分布$\mathcal{N}(0, 1)$, 则有

$$\Pr(|X| \geq \epsilon) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-\epsilon^2/2}}{\epsilon}.$$

证明. 随机变量X的密度函数$p(x)=e^{-x^2/2}/\sqrt{2\pi}$, 因此有$p'(x)=-xp(x)$, 从而得到

$$\Pr(X\geq\epsilon)=\int_\epsilon^\infty p(x)dx=\int_\epsilon^\infty\frac{xp(x)}{x}dx\leq\frac{-1}{\epsilon}\int_\epsilon^\infty p'(x)dx=\frac{-1}{\epsilon}[p(x)]_{x=\epsilon}^\infty=\frac{1}{\sqrt{2\pi}}\frac{e^{-\epsilon^2/2}}{\epsilon}$$

从而完成证明. ∎

1.2.1 Markov不等式及其应用

在各种不等式推导证明中，有一个基本但非常有用的工具: Markov不等式, 以俄罗斯著名数学家Andrey Markov命名.

定理1.5 (**Markov不等式**). *假设X是一个非负随机变量. 对任意$\epsilon>0$, 有*

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

证明. 证明一: 根据全概率公式, 有

$$\begin{aligned}E(X)&=E(X|X\geq\epsilon)P(X\geq\epsilon)+E(X|X<\epsilon)P(X<\epsilon)\\&\geq\epsilon P(X\geq\epsilon)+0.\end{aligned}$$

证明二: 对于非负随机变量X, 函数的期望有

$$E(X)=\int_0^\infty\Pr(Z\geq x)dx\geq\int_0^\epsilon\Pr(Z\geq x)dx\geq\int_0^\epsilon\Pr(Z\geq\epsilon)dx=\epsilon\Pr(Z\geq\epsilon).$$

定理证毕. ∎

下面根据Markov不等式给出一般性推导指数不等式的证明技巧, 称为Chernoff方法:

定理1.6 (**Chernoff方法**). *对于随机变量X, 以及任意$t>0$, 有*

$$P(X\geq\epsilon)\leq\inf_{t>0}\left\{e^{-t\epsilon}E(e^{tX})\right\}.$$

证明. 根据Markov不等式有

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

再对t求最小即可完成证明. ∎

由Markov不等式可以直接推导出Chebyshev不等式.

推论1.2 (**Chebyshev不等式**). *假设X是一个均值为$\mu>0$的随机变量, 有*

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

其中$Var(X)$表示随机变量X的方差.

推论1.3. 对n个独立同分布的随机变量$X_1, X_2, \dots, X_n$, 如何满足$E(X_i) = \mu$和$Var(X_i) \le \nu$, 则对任意$\delta > 0$, 至少以$1 - \delta$的概率有

$$\left|\frac{1}{m}\sum_{i=1}^{n} X_i - \mu\right| \le \sqrt{\frac{\nu}{n\delta}}.$$

证明. 根据Chebyshev不等式有

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

而根据方差的性质有

$$Var\left(\frac{1}{m}\sum_{i=1}^{n} X_i\right) = \frac{1}{m^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{m}Var(X_i) \le \frac{\nu}{m}.$$

由此得到

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

令$\delta = \nu/m\epsilon^2$, 求解δ从而完成证明. ∎

由Markov不等式得到比Chebyshev不等式更紧地Cantelli不等式, 又被成为单边Chebyshev不等式.

推论1.4 (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

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

证明. 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. ∎

下面介绍一个与Chebyshev相关的不等式

推论1.5 (Chebyshev Association不等式). 如果函数f和g是两个非递减的函数, 则对随机变量X有

$$E(f(X)g(X)) \geq E(f(X))E(g(X));$$

如果函数f是非单调递增的, 而函数g是非单调递减的, 则对随机变量X有

$$E(f(X)g(X)) \leq E(f(X))E(g(X)).$$

证明. 假设随机变量Y与随机变量X独立同分布, 如果函数f和g是两个非递减的函数, 则有$(f(x)-f(y))(g(x)-g(y)) \geq 0$, 从而得到

$$(f(X)-f(Y))(g(X)-g(Y)) \geq 0$$

两边同时取期望从而完成第一式的证明, 第二式类似证明. ∎

从上式可容易证明

$$E(X^2) \geq (E(X))^2.$$

Chebyshev Association不等式的一个重要推广结论是: 考虑一个在$\mathbb{R}^n$实值函数f, 满足当其他变量保持不变的情形下, 函数f对其中某一个变量非单调递减, 或单独增加.

定理1.7 (Harris不等式). 假设函数$f, g : \mathbb{R}^n \to \mathbb{R}$是两个对单一变量而言非单调递减的函数, 假设$X_1, X_2, \ldots, X_n$是$n$个独立的随机变量, 以及记随机向量$X = (X_1, X_2, \ldots, X_n)$, 则有

$$E(f(X)g(X)) \geq E(f(X))E(g(X)).$$

类似地, 当单一变量而言, $f : \mathbb{R}^n \to \mathbb{R}$是非单调递减的, 而$g : \mathbb{R}^n \to \mathbb{R}$是非单调递增的, 则有

$$E(f(X)g(X)) \leq E(f(X))E(g(X)).$$

证明. 将利用数学归纳法证明Harris不等式. 由Chebyshev Association不等式可知$n=1$时显然成立. 不妨假设Harris不等式对$n-1$时成立, 下面将证明其对n亦成立.

对任意给定$X_1, X_2, \ldots, X_{n-1}$, 设$f_1(X_n) = f(X)$和$g_1(X_n) = g(X)$. 由Chebyshev Association不等式可得

$$E_{X_n}(f_1(X_n)g_1(X_n)) \geq E_{X_n}(f_1(X_n))E_{X_n}(g_1(X_n)).$$

由此进一步得到

$$\begin{aligned} E(f(X)g(X)) &= E_{X_1,\ldots,X_{n-1}}[E_{X_n}[f(X)g(X)|X_1,\ldots,X_{n-1}]] \\ &\geq E_{X_1,\ldots,X_{n-1}}[E_{X_n}[f(X)]E_{X_n}[g(X)]|X_1,\ldots,X_{n-1}] \end{aligned}$$

可以发现$E_{X_n}[f(X)]E_{X_n}[g(X)]$是一个关于$X_1, X_2, \ldots, X_{n-1}$的函数, 由归纳假设可知

$$E_{X_1,\ldots,X_{n-1}}[E_{X_n}[f(X)]E_{X_n}[g(X)]] \geq E[f(X)]E[g(X)],$$

从而完成证明. ∎

根据$P(|X-\mu|\geq\epsilon)=P(|X-\mu|^2\geq\epsilon^2)$, 以及将Markov不等式应用于随机变量$Y=|X-\mu|^2$, 上述推论得证. 更一般的情况有

推论1.6. 假设X是一个均值为$\mu>0$的随机变量, 以及$g(t)$是一个非负且单调增加的函数, 有

$$P(X\geq\mu+\epsilon)\leq\frac{E(g(X-\mu))}{g(\epsilon)}.$$

特别地, 如果$g(t)=e^t$, 有

$$P(X\geq\mu+\epsilon)\leq e^{-t}E(e^{X-\mu}). \tag{2}$$

1.2.2 信息熵不等式

下面介绍一些与信息熵相关的不等式. 假设X是一个定义在可数集合$\mathcal{X}$的随机变量, 其分布定义为

$$P(X=x)=p(x) \quad \text{for} \quad x\in\mathcal{X}.$$

则随机变量X的信息熵(又称为香农熵, 或简称为熵)定义为

$$H(X)=E(-\log p(X))=-\sum_{x\in\mathcal{X}}p(x)\log p(x).$$

这里log表示自然对数, 且按惯例记$0\log 0=0$. 这里需要注意的是$H(X)$表示随机变量X的信息熵, 而不是关于随机变量X的函数. 同时有信息熵$H(X)\geq 0$.

相对熵是信息论中另外一个重要的概念, 给定在可数集$\mathcal{X}$上的两个分布P和Q, 以及其对应的概率函数为p和q. 关于分布P和Q的相对熵(又称为K-L距离)定义为

$$D(P||Q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}.$$

首先给出相对熵的一些基本性质:

引理1.5. 对任意给定的两个分布P和Q, 有$D(P||Q)\geq 0$成立, 且等号成立的充要条件是$P=Q$.

证明. 对任意$t>0$, 有$\log t\leq t-1$成立, 且等号成立的充要条件是$t=1$. 进一步可得

$$D(P||Q)=-\sum_{x\in\mathcal{X}}p(x)\log\frac{q(x)}{p(x)}\geq-\sum_{x\in\mathcal{X}}p(x)\left(\frac{q(x)}{p(x)}-1\right)=-\sum_{x\in\mathcal{X}}(q(x)-p(x))=0.$$

上式中等号成立的充要条件是对任意$x\in\mathcal{X}$, 有$q(x)=p(x)$ 成立, 即$P=Q$. ■

由上述引理可知两个分布的相对熵非负, 且分布相同时相对熵为零, 因此在一定程度上刻画了两个分布之间的距离. 如果分布Q为可数集$\mathcal{X}$上的均匀分布, 则有

$$D(P||Q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{p(x)}{q(x)}=\sum_{x\in\mathcal{X}}p(x)\log p(x)-p(x)\log q(x)=\log|\mathcal{X}|-H(X)\geq 0$$

这里X是一个服从分布P的随机变量. 因而有

$$H(X) \le \log |\mathcal{X}|,$$

其中等号成立的充要条件是随机变量X在可数集$\mathcal{X}$上服从均匀分布.

现在研究两个随机变量的信息熵. 假设(X,Y)是可数集$\mathcal{X}\times\mathcal{Y}$的两个随机变量, 服从联合分布$P$, 其联合概率质量函数为$p(x,y)_{(x,y)\in\mathcal{X}\times\mathcal{Y}}$. 用$p_X$和$p_Y$表示随机变量$X$ 和Y所对应的边缘分布概率质量函数, 下面给出各种熵的定义:

定义1.2. 随机变量X和Y的联合熵(*joint entropy*)定义为

$$H(X,Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log p(x,y);$$

随机变量X和Y的熵定义为

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x) \quad 和 \quad H(Y) = -\sum_{y\in\mathcal{Y}} p_Y(y)\log p_Y(y);$$

随机变量X和Y的互信息(*mutual infonmation*)定义为

$$I(X,Y) = H(X) + H(Y) - H(X,Y);$$

随机变量X在变量Y下的条件熵(*conditional entropy*) 定义为

$$H(X|Y) = H(X,Y) - H(Y).$$

基于这些定义可以推导如下性质:

引理1.6. 有如下关系成立:

$$\begin{aligned} H(X,Y) &\le H(X)+H(Y), \\ H(X|Y) &\le H(X), \\ H(X_1,X_2,\ldots,X_n) &= H(X_1)+H(X_2|X_1)+\cdots+H(X_n|X_1,\ldots,X_{n-1}). \end{aligned}$$

证明. 根据引理1.5以及定义可得

$$I(X,Y) = H(X)+H(Y)-H(X,Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(Y)} = D(P||Q) \ge 0 \tag{3}$$

这里分布Q为两个边缘分布P_X和P_Y的直积$P_X\otimes P_Y$. 由条件熵的定义得到

$$H(X|Y) = H(X,Y) - H(Y) \le H(X),$$

其中不等式成立是因为$H(X,Y) \le H(Y)+H(X)$. 再次利用条件熵的定义$H(Y|X) = H(X,Y) -$

$H(X)$可得到

$$\begin{aligned} H(X,Y) &= H(X)+H(Y|X) \\ H(X,Y,Z) &= H(X)+H(Y,Z|X)=H(X)+H(Y|X)+H(Z|X,Y) \\ H(X_1,X_2,\ldots,X_n) &= H(X_1)+H(X_2,\ldots,X_n|X_1) \\ &= H(X_1)+H(X_2|X_1)+H(X_3,\ldots,X_n|X_1,X_2)=\cdots \\ &= H(X_1)+H(X_2|X_1)+\cdots+H(X_n|X_1,\ldots,X_{n-1}). \end{aligned}$$

引理证毕. ■

由式(3)可知互信息非负, 即$I(X,Y)\geq 0$, 且互信息$I(X,Y)=0$ 的充要条件是随机变量X和Y相互独立, 因此互信息反映了两个随机变量的独立程度. 下面介绍信息论中非常重要的Han不等式:

定理1.8. 假设$X_1,X_2,\ldots,X_n$为n个离散型随机变量, 有

$$H(X_1,X_2,\ldots,X_n)\leq \frac{1}{n-1}\sum_{i=1}^{n} H(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n).$$

证明. 根据条件熵的定义和引理1.6, 对任意给定$i\in(n)$, 可以得到

$$\begin{aligned} H(X_1,X_2,\ldots,X_n) &= H(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)+H(X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) \\ &\leq H(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)+H(X_i|X_1,\ldots,X_{i-1}) \end{aligned}$$

将上式中i从1到n求和整理得到

$$nH(X_1,X_2,\ldots,X_n)\leq \sum_{i=1}^{n} H(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)+\sum_{i=1}^{n} H(X_i|X_1,\ldots,X_{i-1}).$$

定理证毕. ■

1.3 Concentration不等式

机器学习非常关注多个独立随机变量的随机波动, 假设$X,X_1,X_2,\ldots,X_n$是$n+1$个独立同分布的随机变量, 其期望为μ. 机器学习通常关注随机变量

$$\bar{X}_n=\sum_{i=1}^{n}\frac{X_i}{n}$$

与μ之间的距离. 在机器学习中一种最基本的分析方法是指数不等式, 即

$$P(|\mu-\bar{X}_n|\geq\epsilon)\leq\exp(-r_n(\epsilon))$$

其中$r_n(\epsilon)$是一个依赖于n和ϵ的函数, 且当$n\to\infty$ 时$r_n(\epsilon)\to\infty$. 研究指数不等式有利于分析机器学习方法的泛化性, 其基本构造方法是Markov不等式, 即对任意$\lambda>0$, 有

$$P(\bar{X}\geq\mu+\epsilon)=P(\lambda n\bar{X}-n\lambda\mu\geq n\lambda\epsilon)\leq\frac{E(e^{\lambda n(\bar{X}_n-\mu)})}{e^{n\lambda\epsilon}}=\frac{E(e^{\lambda(\sum_{i=1}^{n}X_i-\mu)})}{e^{n\lambda\epsilon}}=e^{-n\lambda\epsilon}\left(E(e^{\lambda(X_1-\mu)})\right)^n.$$

因此只需要估计$E(e^{\lambda X_1})$, 通常假设$E(e^{\lambda(X_1-\mu)})<\infty$. 对上式两边同时去对数, 有

定理1.9. 对任意$n \geq 0$和$\epsilon > 0$, 有

$$n^{-1} \ln P(\bar{X}_n \geq \mu + \epsilon) \leq \inf_{\lambda > 0} (-\lambda\epsilon + \ln E(e^{\lambda(X_1 - \mu)})).$$

类似有

$$n^{-1} \ln P(\bar{X}_n \leq \mu - \epsilon) \leq \inf_{\lambda < 0} (\lambda\epsilon + \ln E(e^{\lambda(X_1 - \mu)})).$$

函数$\Gamma(\lambda) = \ln E(e^{\lambda X_1})$称为随机变量$X_1$的对数势生成函数(logarithmic moment generating function).

定义1.3. 一个随机变量X称为亚高斯*(sub-Gaussian)*的, 是指其对数势生成函数为一个二次函数, 即

$$\ln E(e^{\lambda(X-\mu)}) \leq \lambda^2 b/2.$$

Sub-Gaussian means that the tail is no heavier than Gaussian.

如果一个随机变量X是亚高斯的, 即$\ln E(e^{\lambda(X-\mu)}) \leq \lambda^2 b/2$, 那么有

$$\inf_{\lambda > 0} (-\lambda\epsilon + \ln E(e^{\lambda(X_1 - \mu)})) = \inf_{\lambda > 0} (-\lambda\epsilon + \lambda^2 b/2) = -\epsilon^2/2b. \tag{4}$$

进一步可得

$$P(\bar{X}_n \geq \mu + \epsilon) \leq e^{-n\epsilon^2/2b} \quad \text{和} \quad P(\bar{X}_n \leq \mu - \epsilon) \leq e^{-n\epsilon^2/2b}.$$

随后研究两种亚高斯随机变量, 即高斯随机变量和任何有界随机变量都是亚高斯的.

1.3.1 高斯随机变量

定理1.10. 如果随机变量X_1服从高斯分布$\mathcal{N}(\mu, \sigma^2)$, 那么它是一个亚高斯随机变量, 且满足

$$\ln E(e^{\lambda(X-\mu)}) \leq \lambda^2\sigma^2/2.$$

证明. 如果随机变量X_1服从高斯分布$\mathcal{N}(\mu, \sigma^2)$, 那么有

$$E(e^{\lambda(X_1-\mu)}) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\lambda x} e^{-x^2/2\sigma^2} dx = e^{\sigma^2\lambda^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(\lambda\sigma - x/\sigma)^2/2} d(x/\sigma) = e^{\sigma^2\lambda^2/2}.$$

定理得证. ■

假设$X_1, \ldots, X_n$是n独立同分布的随机变量, 即$X_i \sim \mathcal{N}(\mu, \sigma^2)$. 设$\bar{X}_n = \sum_{i=1}^n X_i/n$, 根据定理1.10和式 (4)有

$$P(\bar{X} - \mu \geq \epsilon) \leq e^{-n\epsilon^2/2\sigma^2}.$$

习题1.1. 如何利用高斯函数的性质推导$P(\bar{X} - \mu \geq \epsilon)$上界.

1.3.2 Chernoff不等式

定理1.11. *如果随机变量$X \in (\alpha, \beta)$, 那么它是一个亚高斯随机变量且满足$b = (\beta-\alpha)^2/4$, 即*

$$\ln E(e^{\lambda(X-\mu)}) \le \lambda^2(\beta-\alpha)^2/8.$$

此定理的证明依赖于著名的Chernoff引理:

引理1.7. ***Chernoff*引理***: 如果随机变量$X \in (0,1)$, 那么有*

$$\ln E(e^{\lambda X}) \le \lambda\mu + \lambda^2/8.$$

证明. 由凸函数的性质可知

$$e^{\lambda X} \le Xe^{\lambda} + (1-X)e^0 \Rightarrow E(e^{\lambda X}) \le 1-\mu+\mu e^{\lambda} \tag{5}$$

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

$$f'(\lambda) = \frac{\mu e^{\lambda}}{1-\mu+\mu e^{\lambda}} \Rightarrow f'(0) = \mu.$$

Further, we have

$$f''(\lambda) = \frac{\mu e^{\lambda}}{1-\mu+\mu e^{\lambda}} - \frac{\mu^2 e^{2\lambda}}{(1-\mu+\mu e^{\lambda})^2} \le 1/4.$$

根据泰勒展式有

$$f(\lambda) = f(0) + \lambda f'(0) + f''(\xi)\lambda^2/2 \le \lambda\mu + \lambda^2/2/8.$$

引理得证. ■

习题1.2. *如何利用引理1.7证明定理1.11.*

根据定理1.10和式(4)可以推导Chernoff不等式.

定理1.12. ***Chernoff*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$X_i \in (\alpha, \beta)$. 设$\bar{X}_n = \sum_{i=1}^n X_i/n$, 有*

$$P(\bar{X} - \mu \ge \epsilon) \le e^{-2n\epsilon^2/(\beta-\alpha)^2} \quad \text{和} \quad P(\bar{X} - \mu \le -\epsilon) \le e^{-2n\epsilon^2/(\beta-\alpha)^2}.$$

1.3.3 Bennet不等式

当研究有上界的随机变量时, 如果考虑随机变量的方差, 能够推导出更紧地指数不等式.

定理1.13. ***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$\mu = E(X_i)$, 方差为$V = E(X_i - E(X_i))^2$, 以及$\bar{X}_n = \sum_{i=1}^n X_i/n$, 有*

$$P(\bar{X}_n \ge \mu + \epsilon) \le \exp(-n\epsilon^2/(2V + 2\epsilon/3)).$$

证明. 设$X' = X - \mu$, 利用公式$\ln z \le z - 1$得到

$$\begin{aligned}\ln E(e^{\lambda(X-\mu)}) &= \ln E(e^{\lambda X'}) \le E(e^{\lambda X'}) - 1\\ &= \lambda^2 E\left(\frac{e^{\lambda X'} - \lambda X' - 1}{\lambda^2 X'^2}(X')^2\right)\\ &\le \lambda^2 E\left(\frac{e^{\lambda} - \lambda - 1}{\lambda^2}(X')^2\right) = (e^{\lambda} - \lambda - 1)V\end{aligned}$$

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

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

设$\lambda = \epsilon/(V + \epsilon/3)$, 有

$$\inf_{\lambda>0}(-\lambda\epsilon + \frac{\lambda^2}{2(1-\lambda/3)}V) \le -\epsilon^2/(2V(X) + 2\epsilon/3).$$

定理得证. ■

1.3.4 Bernstein不等式

Bennet不等式研究有上界的随机变量基于方差指数不等式, 本节考虑另一种基于方差的不等式. 与Bennet不等式不同之处在于随机变量的势函数具有约束条件.

定理1.14. ***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$\mu = E(X_i)$, 方差为$V = E(X_i - E(X_i))^2$, $\bar{X}_n = \sum_{i=1}^n X_i/n$. 如果存在$b > 0$, 使得对任意$m \ge 2$, 有$E(X_i^m) \le m!b^{m-2}V/2$成立, 那么有*

$$P(\bar{X}_n \ge \mu + \epsilon) \le \exp(-n\epsilon^2/(2V + 2b\epsilon)).$$

证明. 首先有

$$\ln E(e^{\lambda X}) \le E(e^{\lambda X}) - 1 = \sum_{m=1}^{\infty} E(X^m)\lambda^m/m! \le \lambda\mu + \lambda^2 V/2\sum_{m=2}^{\infty}(b\lambda)^{m-2} = \lambda\mu + \frac{\lambda^2 V}{2(1-b\lambda)}.$$

设$\lambda = \epsilon/(V + b\epsilon)$, 进一步可得

$$\inf_{\lambda>0}(-\lambda\epsilon + \frac{\lambda^2 V}{2(1-b\lambda)}) \le -\epsilon^2/2(V + b\epsilon).$$

定理得证. ■

1.3.5 指数不等式的另一种表达方式

对Chernoff指数不等式而言, 如果令

$$P(\bar{X}_n \ge \mu + \epsilon) \le e^{-2n\epsilon^2} = \delta$$

则可以得到Chernoff不等式另一种表述形式, 即至少以$1-\delta$的概率有

$$\bar{X}_n \le \mu + \sqrt{\ln(1/\delta)/2n}.$$

于是得到收敛率$\bar{X}_n \le \mu + O(1/\sqrt{n})$.

而对Bennet指数不等式,

$$P(\bar{X}_n \ge \mu+\epsilon) \le \exp(-n\epsilon^2/(2Var(X)+2\epsilon/3)) = \delta$$

其另外一种表述为: 至少以$1-\delta$的概率有

$$\bar{X}_n \le \mu+\epsilon \le \mu + \frac{2\ln 1/\delta}{3n} + \sqrt{\frac{2V}{n}\ln\frac{1}{\delta}}.$$

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

习题1.3. *给出Bernstein不等式的另一种表述.*

1.4 Martingale不等式

假设$X_1, X_2, \ldots, X_n$为空间$\mathcal{X}$上的n个随机变量, 函数$f: \mathcal{X}^n \to \mathbb{R}$, 设

$$Z = f(X_1, X_2, \ldots, X_n).$$

符号$E_i(\cdot) = E(\cdot|X_1,\ldots,X_i)$表示基于$X_1,\ldots,X_i$ 的条件期望, 于是有

$$E_0 Z = E(Z) \quad \text{和} \quad E_n(Z) = Z.$$

定义变量Z的Doob鞅表示(Doob martingale representation)为

$$Z - E(Z) = \sum_{i=1}^n E_i(Z) - E_{i-1}(Z) = \sum_{i=1}^n \Delta_i$$

其中$\Delta_i = E_i(Z) - E_{i-1}(Z)$称为鞅差(martingale difference). 根据上面的关系, 可以得到变量Z的鞅差表达形式

$$\mathrm{Var}(Z) = E((Z-E(Z))^2) = E\Big(\sum_{i=1}^n \Delta_i\Big)^2 = \sum_{i=1}^n E(\Delta_i^2) + 2\sum_{j>i} E(\Delta_i\Delta_j).$$

如果$j>i$, 有$E_i(\Delta_j)=0$成立, 以及$E_i(\Delta_i\Delta_j) = \Delta_i E_i(\Delta_j) = 0$. 于是得到

$$\mathrm{Var}(Z) = E\Big(\sum_{i=1}^n \Delta_i\Big)^2 = \sum_{i=1}^n E(\Delta_i^2).$$

基于前面的分析, 可以得到如下Efron-Stein不等式定理:

定理1.15 (**Efron-Stein不等式**)**.** *假设$X_1, X_2, \ldots, X_n$是在空间$\mathcal{X}$上n个独立的随机变量. 设$f: \mathcal{X}^n \to \mathbb{R}$和$Z = f(X_1, X_2, \ldots, X_n)$, 有*

$$Var(Z) \le \sum_{i=1}^n E(Z - E^i(Z))^2 = \nu, \tag{6}$$

其中$E^i(Z)$表示仅对第i个变量X_i的期望, 即

$$E^i(Z) = E_{X_i}(Z|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n).$$

假设随机变量$X_1',X_2',\ldots,X_n'$分别与$X_1,X_2,\ldots,X_n$中每个相对应的随机变量独立同分布, 且令$Z_i' = f(X_1,\ldots,X_{i-1},X_i',X_{i+1},\ldots,X_n)$, 则有

$$\nu = \frac{1}{2}\sum_{i=1}^n E((Z-Z_i')^2) = \sum_{i=1}^n E\left((Z-Z_i')_+^2\right) = \sum_{i=1}^n E\left((Z-Z_i')_-^2\right) \tag{7}$$

其中$x_+ = \max(0,x)$和$x_- = \max(-x,0)$.

证明. 证明主要依赖于条件期望的性质. 假设X和Y是任意有界的两个随机变量, 则有$E(XY) = E_Y(E(XY|Y)) = E_X(E(XY|X))$. 由前面的分析可知$\mathrm{Var}(Z) = \sum_{i=1}^n E(\Delta_i^2)$, 以及

$$\Delta_i^2 = \left(E_i(Z) - E_{i-1}(Z)\right)^2 = \left(E_i(Z - E^i(Z))\right)^2$$

其中$E_i(E^i(Z)) = E_{i-1}(Z)$. 对上式利用Jensen不等式有

$$\Delta_i^2 \le E_i(Z - E^i(Z))^2,$$

从而完成了式 (6)的证明. 如果随机变量X与X'是独立同分布的, 则有

$$\begin{aligned}
E((X-X')^2) &= E((X-E(X))^2) + E((X'-E(X'))^2) - 2E(X-E(X))E(X'-E(X')) \\
&= E((X-E(X))^2) + E((X'-E(X'))^2) = 2\mathrm{Var}(X),
\end{aligned}$$

利用上式可得

$$\nu = \sum_{i=1}^n E(Z-E^i(Z))^2 = \frac{1}{2}\sum_{i=1}^n E((Z-Z_i')^2).$$

另一方面有

$$X^2 = X_-^2 + X_+^2 \quad \text{和} \quad E(X_-^2) = E(X_+^2)$$

从而完成式 (7)的证明. ∎

下面研究一个与Efron-Stein不等式密切相关的结论. 对任意随机变量X和a有

$$Var(X) \le E((X-a)^2).$$

基于此性质, 对任意$i \in [n]$, 任何函数$g_i : \mathbb{R}^{n-1} \to \mathbb{R}$有

$$E_i((Z - E_i(Z))^2) \le E_i((Z - Z_i)^2)$$

其中$Z_i = g_i(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$, 结合Efron-Stein不等式有

定理1.16. 对任意$i \in [n]$, 假设$Z_i = g_i(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$, 有

$$Var(Z) \le \sum_{i=1}^n E(Z-Z_i)^2$$

下面研究与鞅差密切相关的McDiarmid不等式:

定理1.17. ***McDiarmid*不等式***:* 假设$X_1, X_2, \ldots, X_m$ 为空间$\mathcal{X}$的m独立随机变量, 假设$f\colon \mathcal{X}^m \to \mathbb{R}$是一个关于$X_1, \ldots, X_m$的实值函数, 并满足对任意的$x_1, x_2, \ldots, x_m, x_i' \in \mathcal{X}$, 都有

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \le c_i$$

成立. 那么对任意$\epsilon > 0$, 有

$$P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m) \ge \epsilon) \le e^{-2\epsilon^2 / \sum_{i=1}^m c_i^2}.$$

证明. 对任意$\lambda > 0$, 由Markov不等式有

$$\begin{aligned} & P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m) \ge \epsilon) \\ = \; & P(e^{\lambda f(X_1,\ldots,X_m) - \lambda E(f(X_1,\ldots,X_m)} \ge e^{\lambda\epsilon}) \\ \le \; & e^{-\lambda\epsilon} E(e^{\lambda f(X_1,\ldots,X_m) - \lambda E(f(X_1,\ldots,X_m)}) \end{aligned}$$

令$Z_i = E_i(f(X_1, \ldots, X_m)) = E_i(f(X_1, \ldots, X_m)|X_1, X_2, \ldots, Z_i)$, 则有$Z_m = f(X_1, \ldots, X_m)$和$Z_0 = E(f(X_1, \ldots, X_m))$, 以及

$$f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m)) = \sum_{i=1}^m Z_i - Z_{i-1}.$$

于是得到

$$P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m) \ge \epsilon) \le e^{-\lambda\epsilon} E(e^{\lambda \sum_{i=1}^m (Z_i - Z_{i-1})}) \tag{8}$$

根据公式$E(E(X|Y,Z)|Z) = E(X|Z)$可得到

$$E(e^{\lambda \sum_{i=1}^m (Z_i - Z_{i-1})}) = E(e^{\lambda \sum_{i=1}^{m-1} (Z_i - Z_{i-1})} E(e^{\lambda(Z_m - Z_{m-1})}|X_1^{m-1}))$$

根据Chernoff引理1.7可得

$$E(e^{\lambda(Z_m - Z_{m-1})}|X_1^{m-1}) \le e^{\lambda^2 c_m^2/8},$$

于是得到

$$E(e^{\lambda \sum_{i=1}^m (Z_i - Z_{i-1})}) \le e^{\lambda^2 c_m^2/8} E(e^{\lambda \sum_{i=1}^{m-1} (Z_i - Z_{i-1})}),$$

对$i = m-1, m-2, \ldots, 1$重复上述过程可得

$$E(e^{\lambda \sum_{i=1}^m (Z_i - Z_{i-1})}) \le e^{\lambda^2 \sum_{i=1}^m c_i^2/8}.$$

根据式 (8)可得

$$P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m) \ge \epsilon) \le e^{-\lambda\epsilon} e^{\lambda^2 \sum_{i=1}^m c_i^2/8}.$$

设$\lambda = 4\epsilon / \sum_{i=1}^m c_i^2$, 代入上式, 通过简单计算即可完成证明. ■

由McDiarmid不等式可以直接得到Hoeffding不等式:

推论1.7. 在集合$(0,1)$上的n个随机变量$X_1, X_2, \ldots, X_n$, 设$f(X_1, \ldots, X_n) = \sum_{i=1}^m X_i/m$, 则有$c_i = 1/n$, 由*McDiarmid*不等式有

$$P(f(X_1, \ldots, X_n) - E(f(X_1, \ldots, X_n) \ge \epsilon) \le e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2} = e^{-2n\epsilon^2}.$$

参考文献

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320.

Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. (2008a). Optimal stragies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423.

Abernethy, J., Hazan, E., and Rakhlin, A. (2008b). Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 263–274.

Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Dani, V., Hayes, T. P., Kakade, S. M., Rakhlin, A., and Tewari, A. (2008). High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 335–341.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Cesa-bianchi, N., Conconi, A., and Gentile, C. (2002). On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008a). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, pages 345–352.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008b). Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. (2015). Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations, 3rd Edition.* Johns Hopkins University Press.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Hazan, E. and Kale, S. (2011). Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436.

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512.

Hazan, E. and Seshadhri, C. (2007). Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity*, 88.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 393–400.

Hou, B.-J., Zhang, L., and Zhou, Z.-H. (2017). Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems 30*.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 99–109.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 440–447.

Kakade, S. M. and Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, pages 801–808.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition.

Mahdavi, M., Zhang, L., and Jin, R. (2015). Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Annual Conference on Learning Theory*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A., editors (2012). *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Core discussion papers.

Nesterov, Y. (2011). Random gradient-free minimization of convex functions. Core discussion papers.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 636–642.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory - Series A*, 13(1):145–147.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009a). Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Shalev-Shwartz, S., Shamir, O., Sridharan, K., and Srebro, N. (2009b). Learnability and stability in the general learning setting.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261.

Srebro, N., Sridharan, K., and Tewari, A. (2010). Optimistic rates for learning with a smooth loss. *ArXiv e-prints*, arXiv:1009.3896.

Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Tseng, P. (2008). On acclerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vapnik, V. N. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280.

Yang, T., Zhang, L., Jin, R., and Yi, J. (2016). Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 449–457.

Zhang, L., Yang, T., and Jin, R. (2017). Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 1954–1979.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-H. (2016). Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*.

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. (2018). Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.