

第 7 章 集中不等式 (Concentration)

给定一个训练数据集

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

其中 $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ 表示第 i 个训练样本的特征 (feature), $y_i \in \mathcal{Y} = \{0, 1\}$ 表示第 i 个训练样本的标记 (二分类). 假设 $\mathcal{D}$ 是空间 $\mathcal{X} \times \mathcal{Y}$ 的一个未知不可见的联合分布. 机器学习的经典假设是训练数据集 S_n 中每个数据 $(\mathbf{x}_i, y_i)$ 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 $f: \mathcal{X} \rightarrow \{0, 1\}$, 定义函数 f 在训练数据集 S_n 上的分类错误率为

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

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. 在实际应用中我们更关心函数 f 对未见数据的分类性能, 即函数 f 在分布 $\mathcal{D}$ 上的分类错误率, 称之为 ‘泛化错误率’

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

由于分布 $\mathcal{D}$ 不可知, 不能直接计算 $R(f, \mathcal{D})$, 但我们已知训练数据集 S_n 和训练错误率 $\hat{R}(f, S_n)$, 如何基于训练错误率 $\hat{R}(f, S_n)$ 来有效估计 $R(f, \mathcal{D})$? 我们可以将问题归纳为

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

即能否以很大的概率保证

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

从而在理论上保证 $\hat{R}(f, S_n)$ 是 $R(f)$ 的一个有效估计. 上述性质在机器学习被称为 ‘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.

首先来看一种简单的例子:

例 7.1 假设训练数据集 $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ 根据分布 $\mathcal{D}$ 独立采样所得, 分类器 f 在训练集 S_n 的错误率为零 (全部预测正确), 求分类器 f 在分布 $\mathcal{D}$ 上的错误率介于 0 和 ϵ 之间的概率 ($\epsilon > 0$).

解 设随机变量

$$X_i = \mathbb{I}[f(\mathbf{x}_i) \neq y_i] \quad (i \in [n]),$$

根据数据集的独立同分布假设可知 $X_1, X_2, \dots, X_n$ 是独立同分布的随机变量. 令 $p = E[X_i]$, 则有 $X_i \sim \text{Ber}(p)$. 分类器 f 在训练集 S_n 的错误率为零, 且在分布 $\mathcal{D}$ 上的错误率大于 ϵ 的概率为

$$\begin{aligned} \Pr \left[\sum_{i=1}^n X_i = 0, p > \epsilon \right] &\leq \Pr \left[\sum_{i=1}^n X_i = 0 | p > \epsilon \right] \\ &= \Pr [X_1 = 0, X_2 = 0, \dots, X_n = 0 | p > \epsilon] \quad (\text{根据独立性假设}) \\ &= \prod_{i=1}^n \Pr [X_i = 0 | p > \epsilon] \leq (1 - \epsilon)^n \leq \exp(-n\epsilon). \end{aligned}$$

因此当分类器 f 在训练集 S_n 的错误率为零且 $p \in (0, \epsilon)$ 的概率至少以 $1 - \exp(-n\epsilon)$ 成立.

对上例的求解进一步进行归纳, 设随机变量

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

则机器学习问题可通过概率统计抽象描述为: 假设有 n 个独立同分布的随机变量 $X_1, X_2, \dots, X_n$, 如何从 n 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.

7.1 基础不等式

首先给出一些基础的概率或期望不等式. 首先研究著名的 Markov 不等式:

定理 7.1 对任意随机变量 $X \geq 0$ 和 $\epsilon > 0$, 有

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

证明 利用全期望公式考虑随机事件 $X \geq \epsilon$ 有

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

从而完成证明.

利用 Markov 不等式可得到一系列有用的不等式:

推论 7.1 对任意随机变量 X 和 $\epsilon \geq 0$, 以及单调递增的非负函数 $g(x)$, 有

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

利用 Markov 不等式可以推导 Chebyshev 不等式:

定理 7.2 (Chebyshev 不等式) 设随机变量 X 的均值为 μ , 则有

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

证明 根据 Markov 不等式有

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

例 7.2 设随机变量 X 和 Y 的期望分别为 -1 和 1 , 方差分别为 2 和 8 , 以及 X 和 Y 的相关系数为 $-1/2$, 利用 Chebyshev 不等式估计概率 $P(|X + Y| \geq 6)$ 的上界.

解 根据随机变量 X 和 Y 的相关系数为 -1 可知

$$\text{Cov}(X, Y) = \rho_{XY} \sqrt{\text{Var}(X)\text{Var}(Y)} = -2.$$

由 $E[X + Y] = 0$, 利用 Chebyshev 不等式有

$$\begin{aligned} P(|X + Y| \geq 6) &= P(|X + Y - E[X + Y]| \geq 6) \\ &\leq \text{Var}(X + Y)/36 = (\text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y))/36 = 1/6. \end{aligned}$$

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

引理 7.1 随机变量 X 的均值 $\mu > 0$, 方差 σ^2 , 则对任意 $\epsilon > 0$ 有

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

证明 设随机变量 $Y = X - \mu$, 有 $E(Y) = 0$ 以及 $\text{Var}(Y) = \sigma^2$. 对任意 $t > 0$ 有

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

对 $(\sigma^2 + t^2)/(\epsilon + t)^2$ 求关于 t 的最小值, 求解可得 $t = \sigma^2/\epsilon$, 由此得到

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

另一方面, 对任意 $t > 0$ 有

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

同理完成证明.

下面介绍 Chebyshev 不等式的推论.

推论 7.2 设独立同分布的随机变量 $X_1, X_2, \dots, X_n$ 满足 $E(X_i) = \mu$ 和 $\text{Var}(X_i) \leq \sigma^2$, 对任意实数 $\epsilon > 0$ 有

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

证明 根据 Chebyshev 不等式有

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

而独立同分布的假设有

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

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.

例 7.3 设分类器 f 在训练集 $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ 的错误率为 $\hat{p} > 0$, 求分类器 f 在分布 $\mathcal{D}$ 上的错误率在 $(9\hat{p}/10, 11\hat{p}/10)$ 之间的概率.

解 设 $X_i = \mathbb{I}[f(\mathbf{x}_i) \neq y_i]$ ($i \in [n]$), 则这些随机变量是独立同分布的. 训练错误率

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

设分类器 f 在分布 $\mathcal{D}$ 上的错误率为 p , 则 $X_i \sim \text{Ber}(p)$ 以及

$$p = E[X_i] = E \left[\frac{1}{n} \sum_{i=1}^n X_i \right],$$

根据独立性假设和 Chebyshev 不等式有

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

取 $\epsilon = \hat{p}/10$ 有

$$\Pr[|p - \hat{p}| > \hat{p}/10] \leq \frac{25}{n\hat{p}^2} .$$

引理 7.2 (Young 不等式) 给定正常数 a, b , 对任意满足 $1/p + 1/q = 1$ 的正实数 p, q 有

$$ab \leq a^p/p + b^q/q .$$

证明 根据凸函数性质有

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

引理得证.

根据 Young 不等式可证明著名的 Hölder 不等式.

引理 7.3 (Hölder 不等式) 设 X 和 Y 是随机变量, 若正数 p, q 满足 $1/p + 1/q = 1$, 则有

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

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

证明 设 $c = [E(|X|^p)]^{1/p}$ 和 $d = [E(|Y|^q)]^{1/q}$, 根据 Young 不等式有

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

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.

7.2 Chernoff 不等式

首先给出随机变量的矩生成函数 (Moment Generating Function) 的定义.

定义 7.1 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 7.3 设随机变量 X 的矩生成函数为 $M_X(t)$, 对任意 $n \geq 1$ 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).

证明 由 Taylor 公式有

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

两边同时取期望有

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

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 7.4 对随机变量 X 和 Y , 如果存在常数 $\delta > 0$, 使得当 $t \in (-\delta, \delta)$ 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 7.3 对任意独立的随机变量 X 和 Y 有 $M_{X+Y}(t) = M_X(t)M_Y(t)$.

下面将利用矩生成函数来证明一系列不等式. 给定任意随机变量 X 和任意 $t > 0$ 和 $\epsilon > 0$, 利用 Markov 不等式有

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

特别地, 有

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

类似地, 对任意 $\epsilon > 0$ 和 $t < 0$ 有

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

同理有

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

上述方法称为‘**Chernoff 方法**’, 是证明集中不等式一种最根本最重要的方法. 下面将针对特定的分布或特定的条件, 先求解矩生成函数 $E[e^{tX}]$, 然后求解最小值 t 的取值.

7.2.1 二值随机变量的 Chernoff 不等式

定理 7.5 设随机变量 $X_1, X_2, \dots, X_n$ 相互独立且满足 $X_i \sim \text{Ber}(p_i)$, 令 $\mu = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$. 对任意 $\epsilon > 0$ 有

$$\Pr \left[\sum_{i=1}^n X_i \geq (1 + \epsilon)\mu \right] \leq \left(\frac{e^\epsilon}{(1 + \epsilon)^{(1 + \epsilon)}} \right)^\mu;$$

对任意 $\epsilon \in (0, 1)$ 有

$$\Pr \left[\sum_{i=1}^n X_i \geq (1 + \epsilon)\mu \right] \leq e^{-\mu\epsilon^2/3}.$$

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.

证明 令 $\bar{X} = \sum_{i=1}^n X_i$. 对任意 $t > 0$, 根据 Chernoff 方法有

$$\Pr[\bar{X} \geq (1 + \epsilon)\mu] = \Pr[e^{t\bar{X}} \geq e^{t(1 + \epsilon)\mu}] \leq e^{-t(1 + \epsilon)\mu} E[e^{t\bar{X}}].$$

利用随机变量的独立性以及 $1 + x \leq e^x$ 有

$$\begin{aligned} E[e^{t\bar{X}}] &= E[e^{\sum_{i=1}^n tX_i}] = \prod_{i=1}^n E[e^{tX_i}] \\ &= \prod_{i=1}^n [(1 - p_i) + p_i e^t] = \prod_{i=1}^n [1 + p_i(e^t - 1)] \\ &\leq \exp \left(\sum_{i=1}^n p_i(e^t - 1) \right) = \exp(\mu(e^t - 1)). \end{aligned}$$

由此可得

$$\Pr[\bar{X} \geq (1 + \epsilon)\mu] \leq \exp(-t(1 + \epsilon)\mu + \mu(e^t - 1)).$$

对上式求最小值得 $t_{\min} = \ln(1 + \epsilon)$, 代入可得

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

对第二个不等式, 只需证明当 $\epsilon \in (0, 1)$ 有

$$f(\epsilon) = \ln \left(\frac{e^\epsilon}{(1+\epsilon)^{(1+\epsilon)}} \right) + \frac{\epsilon^2}{3} = \epsilon - (1+\epsilon) \ln(1+\epsilon) + \frac{\epsilon^2}{3} \leq 0.$$

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 $\Pr[\sum_{i=1}^n X_i \leq (1-\epsilon)\mu]$ 的估计, 证明作为练习题留给大家完成.

定理 7.6 设随机变量 $X_1, X_2, \dots, X_n$ 相互独立且满足 $X_i \sim \text{Ber}(p_i)$, 令 $\mu = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n p_i$. 对任意 $\epsilon \in (0, 1)$ 有

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

定义 7.2 若随机变量 $X \in \{+1, -1\}$ 满足

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

则称 X 为 Rademacher 随机变量.

我们有如下定理:

定理 7.7 对 n 个独立的 Rademacher 随机变量 $X_1, X_2, \dots, X_n$, 有

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

证明 根据 Taylor 展开式有

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

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

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

对任意 $t > 0$, 根据 Chernoff 方法有

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

通过对上式右边求最小值解得 $t = \epsilon$, 带入上式得到

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

同理证明另一个不等式.

推论 7.4 对独立同分布的随机变量 $X_1, X_2, \dots, X_n$ 满足 $P(X_1 = 0) = P(X_1 = 1) = 1/2$, 有

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

7.2.2 有界随机变量的 Chernoff 不等式

本节研究有界的随机变量 $X_i \in [a, b]$ 的 Chernoff 不等式. 首先介绍著名的 Chernoff 引理.

引理 7.4 设随机变量 $X \in [0, 1]$ 的期望 $\mu = E[X]$. 对任意 $t > 0$ 有

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

证明 由凸函数的性质可知

$$e^{tX} = e^{tX+(1-X)0} \leq Xe^t + (1-X)e^0,$$

两边再同时取期望有

$$E(e^{tX}) \leq 1 - \mu + \mu e^t = \exp(\ln(1 - \mu + \mu e^t)).$$

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

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

进一步有

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

根据泰勒中值定理有

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

引理得证.

由上面的 Chernoff 引理进一步推导出

推论 7.5 设随机变量 $X \in [a, b]$ 的期望 $\mu = E[x]$. 对任意 $t > 0$ 有

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

根据上述推论, 我们得到有界随机变量的 Chernoff 不等式:

定理 7.8 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 $\epsilon > 0$ 有

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

证明 这里给出第一个不等式的证明, 第二个不等式证明作为习题. 对任意 $t > 0$, 根据 Chernoff 方法有

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

根据 Chernoff 引理, 对任意 $X_i \in [a, b]$ 有

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

由此得到

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

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.

7.2.3 Gaussian 和 Sub-Gaussian 随机变量不等式

首先考虑独立同分布的 Gaussian 随机变量:

定理 7.9 设随机变量 $X_1, \dots, X_n$ 相互独立、且服从 $X_i \sim \mathcal{N}(\mu, \sigma)$, 对任意 $\epsilon > 0$ 有

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

证明 对随机变量 $X_i \sim \mathcal{N}(\mu, \sigma)$, 根据正太分布的性质有

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

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

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

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.

下面定义 Sub-Gaussian 随机变量, 将有界随机变量和 Gaussian 随机变量统一起来:

定义 7.3 对任意 $t \in (-\infty, +\infty)$, 若随机变量 X 满足

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

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

亚高斯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 7.4 对任意有界的随机变量 $X \in [a, b]$, 根据 Chernoff 引理有

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

即有界的随机变量是参数为 $(b-a)^2/4$ 的亚高斯随机变量.

例 7.5 如果随机变量 X 服从高斯分布 $\mathcal{N}(\mu, \sigma^2)$, 则有

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

Gaussian 随机变量是参数为 σ^2 的亚高斯随机变量.

由前面的例子可知高斯随机变量和有界的随机变量都是亚高斯随机变量. 根据 Chernoff 方法有

定理 7.10 设 $X_1, \dots, X_n$ 是 n 个独立的且参数为 b 的亚高斯随机变量, 对任意 $\epsilon > 0$ 有

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

证明 对任意 $t > 0$, 根据 Chernoff 方法有

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

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

对亚高斯型随机变量, 还可以给出最大值期望的估计:

定理 7.11 设 $X_1, \dots, X_n$ 是 n 个相互独立的、参数为 b 的亚高斯随机变量, 且满足 $E[X_i] = 0$, 我们有

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

证明 对任意 $t > 0$, 根据 Jensen 不等式有

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

对上式两边同时取对数整理可得

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

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

前面所讲的概率不等式, 可以用另外一种表达形式给出, 这里以定理 7.10 为例: 假设 $X_1, \dots, X_n$ 是独立的、且参数为 b 的亚高斯随机变量, 对任意 $\epsilon > 0$ 有

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

令 $\delta = \exp(-n\epsilon^2/2b)$, 求解出

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

代入整理可得: 至少以 $1 - \delta$ 的概率有下面的不等式成立

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

前面讲的所有不等式都可以采用 $1 - \delta$ 的形式描述.

7.3 Bennet 和 Bernstein 不等式

通过考虑随机变量的方差, 可能推导出更紧地集中不等式, 下面介绍两个基于方差的不等式.

定理 7.12 (Bennet不等式) 设 $X_1, \dots, X_n$ 是独立同分布的随机变量且满足 $X_i - E[X_i] \leq 1$, 其均值为 μ 和方差为 σ^2 , 我们有

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

证明 对任意 $t > 0$, 根据 Chernoff 方法有

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

设 $Y = X_1 - \mu$, 利用公式 $\ln z \leq z - 1$ 得到

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

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

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

因此可得

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

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

对于 Bennet 不等式, 令

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

可以给出不等式的另外一种表述: 至少以 $1 - \delta$ 的概率有以下不等式成立

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

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

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矩:

定理 7.13 (Bernstein不等式) 设 $X_1, \dots, X_n$ 是独立同分布的随机变量, 其均值为 μ 和方差为 σ^2 , 若存在常数 $b > 0$, 使得对任意正整数 $m \geq 2$ 有 $E[X_i^m] \leq m!b^{m-2}\sigma^2/2$, 那么我们有

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

证明 对任意 $t > 0$, 根据 Chernoff 方法有

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

利用公式 $\ln z \leq z - 1$ 有

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

由此可得

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

取 $t = \epsilon/(\sigma^2 + b\epsilon)$ 完成证明.

例 7.6 给出 Bernstein 不等式的 $1 - \delta$ 表述.

7.4 应用: 随机投影 (Random Projection)

设高维空间 $\mathbb{R}^d$ 有 n 个点 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ (d 非常大, 如 100 万或 1 亿). 处理这样一个高维的问题很难, 实际中的一种解决方案是能否找到一个保距变换: $f: \mathbb{R}^d \rightarrow \mathbb{R}^k$ ($k \ll d$), 使得以较大概率有

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

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

随机投影可以简单的表示为

$$f(\mathbf{x}) = \mathbf{x}P/c,$$

其中 P 是一个 $d \times k$ 的随机矩阵, 其每个元素之间相互独立, c 为一常数 (根据随机矩阵 P 确定). 下面介绍三种常见的随机矩阵:

- $P = (p_{ij})_{d \times k} \in \mathbb{R}^{d \times k}$, $p_{ij} \sim \mathcal{N}(0, 1)$, 此时 $c = \sqrt{k}$;
- $P = (p_{ij})_{d \times k} \in \{-1, 1\}^{d \times k}$, p_{ij} 为 Rademacher 随机变量, 即 $\Pr(p_{ij} = 1) = \Pr(p_{ij} = -1) = 1/2$, 此时 $c = \sqrt{k}$;
- $P = (p_{ij})_{d \times k} \in \{-1, 0, 1\}^{d \times k}$, 满足 $\Pr(p_{ij} = 1) = \Pr(p_{ij} = -1) = 1/6$ 和 $\Pr(p_{ij} = 0) = 2/3$, 此时 $c = \sqrt{k/3}$. 【主要用于 sparse 投影, 减少计算量】

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 7.5 设 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 为 $\mathbb{R}^d$ 空间的 n 个点, 随机矩阵 $P = (p_{ij})_{d \times k} \in \mathbb{R}^{d \times k}$, $p_{ij} \sim \mathcal{N}(0, 1)$ 且每个元素相互独立, 令

$$\mathbf{y}_i = f(\mathbf{x}_i) = \mathbf{x}_i P / \sqrt{k}, \quad i \in [n]$$

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

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

证明 下面分三步证明 J-L 引理.

第一步: 对任意非零 $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$, 首先证明

$$E \left[\left\| \mathbf{x}P / \sqrt{k} \right\|_2^2 \right] = \|\mathbf{x}\|_2^2,$$

即在期望的情况下, 随机投影变换前后的点到原点的距离相同. 根据 $P = (p_{ij})_{d \times k}$ ($p_{ij} \sim \mathcal{N}(0, 1)$) 有

$$\begin{aligned} E \left[\left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \right] &= E \left[\sum_{j=1}^k \left(\sum_{i=1}^d \frac{x_i p_{ij}}{\sqrt{k}} \right)^2 \right] = \sum_{j=1}^k \frac{1}{k} E \left[\left(\sum_{i=1}^d x_i p_{ij} \right)^2 \right] \\ &= \sum_{j=1}^k \frac{1}{k} \sum_{i=1}^d x_i^2 = \frac{1}{k} \sum_{j=1}^k \|\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2. \end{aligned}$$

第二步: 对任意非零 $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$, 证明

$$\Pr \left[\left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq \exp(-(\epsilon^2 - \epsilon^3)k/4).$$

将矩阵 P 表示为 $P = (P_1, P_2, \dots, P_k)$, 其中 P_i ($i \in [d]$) 是一个 $d \times 1$ 的列向量, 令 $v_j = \mathbf{x}P_j / \|\mathbf{x}\|_2$, 即

$$(v_1, v_2, \dots, v_k) = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_1, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_2, \dots, \frac{\mathbf{x}}{\|\mathbf{x}\|_2} P_k \right).$$

根据 Gaussian 分布的性质有 $v_j \sim \mathcal{N}(0, 1)$, 且 $v_1, v_2, \dots, v_k$ 是 k 个独立的随机变量. 对任意 $t \in (0, 1/2)$, 根据 Chernoff 方法有

$$\begin{aligned} \Pr \left[\left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] &= \Pr \left[\left\| \frac{\mathbf{x}P}{\|\mathbf{x}\|_2} \right\|_2^2 \geq (1 + \epsilon)k \right] \\ &= \Pr \left[\sum_{j=1}^k v_j^2 \geq (1 + \epsilon)k \right] \leq e^{-(1+\epsilon)kt} \left(E[e^{t \sum_{j=1}^k v_j^2}] \right)^k = e^{-(1+\epsilon)kt} \left(E[e^{tv_1^2}] \right)^k. \end{aligned}$$

对标准 Gaussian 分布有

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

代入可得

$$\Pr \left[\left\| \mathbf{x}P/\sqrt{k} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq \left(\frac{e^{-2(1+\epsilon)t}}{1-2t} \right)^{k/2}.$$

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

$$\Pr \left[\left\| \mathbf{x}P/\sqrt{k} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq ((1 + \epsilon)e^{-\epsilon})^{k/2}.$$

设 $f(\epsilon) = \ln(1 + \epsilon)$, 根据 $\epsilon \in (0, 1/2)$ 有

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

根据泰勒中值定理有

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

于是得到

$$\Pr \left[\left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}\|_2^2 \right] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

同理可证

$$\Pr \left[\left\| \frac{\mathbf{x}P}{\sqrt{k}} \right\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}\|_2^2 \right] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

第三步: 对任意给定 $i \neq j$, 根据第二步的结论可知

$$\Pr[\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2] \leq e^{-k(\epsilon^2 - \epsilon^3)/4},$$

$$\Pr[\|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2] \leq e^{-k(\epsilon^2 - \epsilon^3)/4}.$$

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

$$\Pr \left[\exists i \neq j: \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \geq (1 + \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \quad \text{或} \quad \|\mathbf{y}_i - \mathbf{y}_j\|_2^2 \leq (1 - \epsilon) \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \right] \leq 2n^2 e^{-k(\epsilon^2 - \epsilon^3)/4},$$

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

习题

7.1 设随机变量 X 的期望 $E[X] = \mu > 0$, 方差为 σ^2 , 证明对任意 $\epsilon > 0$ 有

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

7.2 设随机变量 X 和 Y 满足 $E(X) = -2$, $E(Y) = 2$, $\text{Var}(X) = 1$, $\text{Var}(Y) = 4$, $\rho_{XY} = -1/2$. 利用 Chebyshev 不等式估计 $\Pr(|X + Y| \geq 6)$ 的上界.

7.3 独立同分布随机变量 $X_1, X_2, \dots, X_n$ 满足 $E[X_i] = \mu$ 和 $\text{Var}(X_i) \leq v$. 证明对任意 $\epsilon > 0$ 有

$$\left[\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right] \leq \frac{v}{n\epsilon^2}.$$

7.4 阐述什么是 Chernoff 方法。

7.5 随机变量 $X_1, X_2, \dots, X_n$ 相互独立且满足 $X_i \sim \text{Ber}(p_i)$ ($p_i > 0$). 利用 Chernoff 方法给出下列概率的上界

$$P \left[\frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \geq \epsilon \right] \quad \text{和} \quad P \left[\frac{1}{n} \sum_{i=1}^n (X_i - E[X_i]) \leq -\epsilon \right].$$

7.6 若独立同分布随机变量 $X_1, X_2, \dots, X_n$ 满足 $X_i \in \{a, b\}$ ($b > a$) 且 $P(X_i = a) = P(X_i = b) = 1/2$. 求下列概率的上界

$$P \left[\frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{a+b}{2} \right) \geq \epsilon \right] \quad \text{和} \quad \Pr \left[\frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{a+b}{2} \right) \leq -\epsilon \right].$$

7.7 随机变量 $X_1, X_2, \dots, X_n$ 相互独立且满足 $X_i \sim \text{Ber}(p_i)$ ($p_i > 0$). 证明对任意 $0 < \epsilon < 1$ 有不等式

$$P \left[\sum_{i=1}^n X_i \geq (1 + \epsilon) \sum_{i=1}^n p_i \right] \leq e^{-\mu\epsilon^2/3}.$$

7.8 随机变量 $X \in [a, b]$ 且期望 $\mu = \mathbb{E}[x]$, 证明对任意 $t > 0$ 有

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

7.9 利用 Chernoff 方法证明: 设 $X_1, X_2, \dots, X_k$ 是 k 个独立的随机变量, 且 $X_i \sim N(0, 1)$, 则有

$$\Pr \left(\sum_{i=1}^k X_i^2 \geq (1 + \epsilon)k \right) \leq \exp (-k(\epsilon^2 - \epsilon^3)/4)$$

7.10 证明 Bennet 不等式.

7.11 证明 Bernstein 不等式.

7.12 已知 Bernstein 不等式

$$P \left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu) \geq \epsilon \right] \leq \exp \left(\frac{-n\epsilon^2}{2\sigma^2 + 2b\epsilon} \right),$$

给出其等价 $1 - \delta$ 描述。

7.13 已知独立同分布随机变量 $X_1, X_2, \dots, X_n$ 满足 $X_i \sim N(\mu, \sigma^2)$, 给出 $E[\max_{i \in [n]} \{X_i\}]$ 的上界, 并给出严格证明。

7.14 假设训练数据集 $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ 根据分布 $\mathcal{D}$ 独立采样所得, 分类器 f 在训练集 S_n 的错误率为零 (全部预测正确), 求分类器 f 在分布 $\mathcal{D}$ 上的错误率介于 0 和 ϵ 之间的概率 ($\epsilon > 0$).

7.15 假设训练数据集 $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ 根据分布 $\mathcal{D}$ 独立采样所得, 分类器 f 在训练集 S_n 的错误率为 $\hat{p}$, 求分类器 f 在分布 $\mathcal{D}$ 上的错误率介于 $\hat{p} - \epsilon$ 和 $\hat{p} + \epsilon$ 之间的概率 ($\epsilon > 0$).