根据协方差的性质有

$$\mathrm{Cov}(X_i, X_j) = \sum_{k=1}^{m} \mathrm{Cov}(Y_i^k, Y_j^k) + \sum_{k\neq l} \mathrm{Cov}(Y_i^k, Y_j^l) = -mp_ip_j \ .$$

最后得到 X_i 和 X_j 的相关系数为

$$\rho = \frac{\mathrm{Cov}(X_i, X_j)}{\mathrm{Var}(X_i)\mathrm{Var}(X_j)} = \frac{-mp_ip_j}{\sqrt{mp_i(1-p_i)}\sqrt{mp_j(1-p_j)}} = -\frac{\sqrt{p_ip_j}}{\sqrt{(1-p_i)(1-p_j)}} \ .$$

6.4 条件期望

前一章介绍了条件分布, 基于条件分布可以考虑条件期望, 分离散和连续性随机变量两种情况.

定义 6.3 设 (X, Y) 为连续型随机变量, 在 $Y = y$ 条件下 X 的条件密度函数为 $f_{X|Y}(x|y)$, 称

$$E(X|y) = E(X|Y = y) = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx$$

为在 $Y = y$ 条件下 X 的 **条件期望**. 设 (X, Y) 为离散型随机变量, 在 $Y = y$ 条件下 X 的条件分布列为 $P(X = x_i|Y = j)$, 称

$$E(X|y) = E(X|Y = y) = \sum_i x_i P(X = x_i|Y = j)$$

为在 $Y = y$ 条件下 X 的 **条件期望**.

条件期望 $E[X|y]$ 一般都与 y 相关, 是 y 的函数, 而 (无条件) 期望 $E(X)$ 是一个具体的常数. 在上面的定义中, 我们都默认期望存在. 条件期望是条件分布的期望, 具有期望的一切性质:

- 对任意常数 a, b 有 $E(aX_1 + bX_2|Y) = aE(X_1|Y) + bE(X_2|Y)$;
- 对离散型随机变量 (X, Y) 和函数 $g(X)$ 有

  $$E(g(X)|Y) = \sum_i g(x_i)P(X = x_i|Y = y) \ ;$$

  对连续型随机变量 (X, Y) 和函数 $g(X)$ 有

  $$E(g(X)|Y) = \int_{-\infty}^{+\infty} g(x)f(x|Y = y)dx \ ;$$

- 设随机向量 $(X, Y) \sim \mathcal{N}(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, 则在 $Y = y$ 的条件下随机变量 X 服从正太分布 $\mathcal{N}(\mu_x - \rho\sigma_x(y - \mu_y)/\sigma_y, (1 - \rho^2)\sigma_x^2)$, 由此可得 $E(X|y) = \mu_x - \rho\sigma_x(y - \mu_y)/\sigma_y$.

下面给出了计算期望的另一种方法.

定理 6.5 对二维随机变量 (X, Y) 有

$$E(X) = E_Y(E(X|Y)) = \begin{cases} \sum_{y_j} E(X|y_j)P(Y = y_j) & \text{离散型随机变量}, \\ \int_{-\infty}^{\infty} E(X|y)f_Y(y)dy & \text{连续型随机变量}. \end{cases}$$

证明 对连续型随机变量 (X, Y), 不妨假设其联合密度函数为 $f(x, y)$, 根据条件概率有

$$E[X] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x, y)dydx = \int_{-\infty}^{+\infty} f_Y(y)\int_{-\infty}^{+\infty} xf_{X|Y}(x|y)dxdy = \int_{-\infty}^{+\infty} E(X|y)f_Y(y)dy \ .$$

对离散型随机变量 (X, Y), 根据条件概率和全概率公式有

$$\begin{aligned} E[X] &= \sum_i x_i P_X(X = x_i) = \sum_i\sum_j x_i P(X = x_i, Y = y_j) \\ &= \sum_i\sum_j x_i P(X = x_i|Y = y_j)P(Y = y_j) \\ &= \sum_j P(Y = y_j)\sum_i x_i P(X = x_i|Y = y_j) \\ &= \sum_j P(Y = y_j)E[X|Y = y_j] = E_Y[E[X|Y]]. \end{aligned}$$

下面介绍与全概率公式相对于的一个公式: **全期望公式** (law of total expectation), 在期望的计算起到重要作用.

定理 6.6 设 $A_1, A_2, \cdots, A_n$ 是样本空间 Ω 的一个分割, 即 $A_iA_j = \emptyset$ 和 $\Omega = \cup_{i=1}^n A_i$. 对任意随机变量 X 有

$$E[X] = E[X|A_1]P(A_1) + E[X|A_2]P(A_2) + \cdots + E[X|A_n]P(A_n) \ ,$$

特别地, 随机事件 A 与其对立事件 $\bar{A}$ 构成样本空间 Ω 的一个划分, 对任意随机变量 X 有

$$E[X] = E[X|A]P(A) + E[X|\bar{A}]P(\bar{A}) \ .$$

证明 对于随机变量 X 和 $A_1, A_2, \cdots, A_n$, 首先引入新的离散随机变量 $Y = 1, 2, \cdots, n$ 满足

$$\text{随机事件 } Y = i \text{ 发生的充要条件是 } X \in A_i \ .$$

根据定理 6.5 可知

$$E(X) = E_Y(E(X|Y)) = \sum_{i=1}^n E(X|Y = i)P(Y = i) = \sum_{i=1}^n E(X|A_i)P(A_i) \ .$$

例 6.8 设 (X,Y) 的联合概率密度为

$$f(x,y)=\begin{cases}\exp(-y) & 0<x<y<+\infty\\ 0 & \text{其它}\ ,\end{cases}$$

求条件期望 $E(X|y)$.

解 首先计算 Y 的边缘密度函数, 当 $y>0$ 时

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx=\int_0^y\exp(-y)dx=y\exp(-y)\ ,$$

由此得到在 $Y=y$ 的条件下 X 的条件分布

$$f_{X|Y}(x|y)=f(x,y)/f_Y(y)=1/y\qquad(0<x<y<+\infty)\ .$$

最后得到条件期望

$$E(X|y)=\int_{-\infty}^{+\infty}xf_{X|Y}(x|y)dx=\int_0^y x/ydx=y/2\ .$$

例 6.9 一矿工被困在有三个门的矿井里, 第一个门通一坑道, 沿此坑道走 3 小时可使他到达安全地点; 第二个门可使他走 5 小时后又回到原处; 第三个门可使他走 7 小时后也回到原地. 如设此矿工在任何时刻都等可能地选定其中一个门, 试问他到达安全地点平均要用多长时间?

解 用 X 为该矿工到达安全地点所需时间, 用 Y 为他所选的门, 根据全期望公式有

$$E(X)=E(X|Y=1)P(Y=1)+E(X|Y=2)P(Y=2)+E(X|Y=3)P(Y=3)\ ,$$

其中 $P(Y=1)=P(Y=2)=P(Y=3)=1/3$, $E(X|Y=1)=3$. 用 $E(X|Y=2)$ 表示矿工从第二个门出去要到达安全地点所需平均时间. 而他沿此坑道走 5 小时又转回原地, 而一旦返回原地, 问题就与当初他还没有进第二个门之前一样, 因此他要到达安全地点平均还需再用 $E(X)$小时. 同理可以考虑 $E(X|Y=2)$, 故有

$$E(X|Y=2)=5+E(X)\quad \text{和}\quad E(X|Y=3)=7+E(X)\ .$$

于是得到

$$E(X)=(3+5+E(X)+7+E(X))/3\ .$$

求解出 $E(X)=15$ (小时), 该矿工到达安全地点平均需要 15 小时.

6.5 随机向量的数学期望与协方差阵

定义 6.4 设随机向量 $X=(X_1,X_2,\cdots,X_n)^\top$, 称

$$E(X)=(E(X_1),E(X_2),\cdots,E(X_n))^\top$$

为随机向量 X 的期望, 以及称

$$\mathrm{Cov}(X)=\Sigma=\begin{pmatrix}\mathrm{Cov}(X_1,X_1) & \cdots & \mathrm{Cov}(X_1,X_n)\\ \mathrm{Cov}(X_2,X_1) & \cdots & \mathrm{Cov}(X_2,X_n)\\ \vdots & & \vdots\\ \mathrm{Cov}(X_n,X_1) & \cdots & \mathrm{Cov}(X_n,X_n)\end{pmatrix}$$

为随机变量 X 的协方差矩阵.

下面介绍协方差矩阵的一些性质:

定理 6.7 随机向量 $X=(X_1,X_2,\cdots,X_n)$ 的协方差矩阵是对称半正定的矩阵.

证明 对任意 $i\neq j$, 根据协方差的性质

$$\mathrm{Cov}(X_i,X_j)=\mathrm{Cov}(X_j,X_i)\ ,$$

可知协方差矩阵是对称的. 对于半正定性的证明, 首先引入新的函数

$$f(t_1,t_2,\ldots,t_n)=(t_1,t_2,\ldots,t_n)\begin{pmatrix}\mathrm{Cov}(X_1,X_1) & \cdots & \mathrm{Cov}(X_1,X_n)\\ \mathrm{Cov}(X_2,X_1) & \cdots & \mathrm{Cov}(X_2,X_n)\\ \vdots & & \vdots\\ \mathrm{Cov}(X_n,X_1) & \cdots & \mathrm{Cov}(X_n,X_n)\end{pmatrix}(t_1,t_2,\ldots,t_n)^\top\ ,$$

由此得到

$$f(t_1,t_2,\ldots,t_n)=E\left(\left(\sum_{i=1}^n t_i(X_i-E(X_i))\right)^2\right)\geqslant 0\ ,$$

由此完成证明.

定理 6.8 设多维正态分布 $X=(X_1,X_2,\ldots,X_n)^\top\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 则有

$$\boldsymbol{\mu}=(E[X_1],E[X_2],\ldots,E[X_n])^\top \quad 和 \quad \boldsymbol{\Sigma}=[\mathrm{Cov}(X_i,X_j)]_{n\times n}.$$

6.6 应用案例

有时我们能观察到随机变量 X 的值, 需要对随机变量 Y 的值进行预测, 即选择一个函数 $g(x)$, 使得 $g(X)$ 接近预测值 Y. 选择函数 $g(x)$ 的一个准则是最小化 $E[(Y-g(x))^2]$. 关于最优的函数 $g(X)$, 有如下结论:

定理 6.9 对任意函数 $g(x)$ 和随机变量 X 和 Y, 有

$$E\left([Y-g(X)]^2\right) \geqslant E\left([Y-E(Y|X)]^2\right) .$$

证明 根据定理 6.5 只需证明对任意给定 X 有

$$E\left([Y-g(X)]^2|X\right) \geqslant E\left([Y-E(Y|X)]^2|X\right) , \tag{6.1}$$

对上式两边分别对 X 求期望可完成证明. 下面考虑如何上面的条件期望不等式, 首先有

$$\begin{aligned}&E\left([Y-g(X)]^2|X\right)=E\left([Y-E(Y|X)+E(Y|X)-g(X)]^2|X\right)\\&=E\left([Y-E(Y|X)]^2|X\right)+E\left([E(Y|X)-g(X)]^2|X\right)+2E\left([Y-E[Y|X]][E[Y|X]-g(X)]|X\right),\end{aligned}$$

当给定 X 后, $[E[Y|X]-g(X)]^2$ 和 $E[Y|X]$ 都是常数, 因此有

$$E\left([Y-E[Y|X]][E[Y|X]-g(X)]|X\right)=[E[Y|X]-g(X)]E\left([Y-E[Y|X]]|X\right)=0 ,$$

结合上面两式完成 (6.1) 的证明.

很多情况下很难知道随机变量 X 和 Y 的联合分布, 有些情况下即使知道联合分布计算 $E(Y|X)$ 也非常复杂. 若已知随机变量 X 和 Y 的一些统计量, 依然可以很好地估计出 X 和 Y 的最优线性预测, 例如,

例 6.10 设随机变量 X 和 Y 的期望、方差、相关系数分别为 $\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho$, 其中 $\sigma_x>0,\sigma_y>0,\rho\in[-1,+1]$, 求解最优的线性预测 $Y=aX+b$ 使得 $E((Y-aX-b)^2)$ 最小化.

解 首先设函数

$$\begin{aligned}F(a,b) &= E\left((Y-aX-b)^2\right)\\ &= E(Y^2)-2aE(Y)-2bE(XY)+a^2+2abE(X)+b^2E(X^2) .\end{aligned}$$

求函数 $F(a,b)$ 的最小值, 可以考虑令 a 和 b 的偏导等于零, 即

$$\begin{cases}\partial F(a,b)/\partial a=2a+2bE(X)-2E(Y)=0\\ \partial F(a,b)/\partial b=2bE(X^2)+2aE(X)-2E(XY)=0 .\end{cases}$$

求解上面的方程组可得

$$\begin{cases} a = E(Y) + bE(X) = \mu_y - \rho\sigma_y\mu_x/\sigma_x \\ b = \frac{E(XY)-E(X)E(Y)}{E(X^2)-(E(X))^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{\rho\sigma_x\sigma_y}{\sigma_x^2} = \rho\sigma_y/\sigma_x \ . \end{cases}$$

由此给出 Y 的最优线性预测为

$$Y = \rho\sigma_y(X-\mu_x)/\sigma_x + \mu_y \ ,$$

在最优线性预测下预测的均分误差

$$\begin{aligned} & E\left((Y-\rho\sigma_y(X-\mu_x)/\sigma_x-\mu_y)^2\right) \\ = \ & E\left((Y-\mu_y)^2\right) + \rho^2\sigma_y^2 E\left((X-\mu_x)^2\right)/\sigma_x^2 - 2\rho\sigma_y E\left((X-\mu_x)(Y-\mu_y)\right)/\sigma_x \\ = \ & \sigma_y^2 + \rho^2\sigma_y^2 - 2\rho^2\sigma_y^2 = \sigma_y^2(1-\rho^2) \ . \end{aligned}$$

由此可以看出, 当 $\rho^2 \to 1$ 时最优线性预测的均方误差接近零.

第 7 章　集中不等式 (Concentration)

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 $(\boldsymbol{x}_i, y_i)$ 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 $f\colon \mathcal{X} \to \{0, 1\}$, 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. 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’

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

由于分布 $\mathcal{D}$ 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

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

即能否以很大的概率保证

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

从而在理论上保证 $\hat{R}(f, D_n)$ 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.

首先来看一种简单的例子:

例 7.1　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, 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

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

根据数据集的独立同分布假设可知 $X_1, X_2, \ldots, X_n$ 是独立同分布的随机变量. 令 $p = E[X_i]$, 则有 $X_i \sim \mathrm{Ber}(p)$. 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

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

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

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

则机器学习问题可通过概率统计抽象描述为: 假设有 n 个独立同分布的随机变量 $X_1, X_2, \ldots, X_n$, 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 $X \geqslant 0$ 和 $\epsilon > 0$, 有

$$P(X \geqslant \epsilon) \leqslant \frac{E(X)}{\epsilon}\ .$$

证明 利用全期望公式考虑随机事件 $X \geqslant \epsilon$ 有

$$E[X] = E[X|X \geqslant \epsilon]P(X \geqslant \epsilon) + E[X|X \leqslant \epsilon]P(X \leqslant \epsilon) \geqslant P(X \geqslant \epsilon)\epsilon\ ,$$

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 $\epsilon \geqslant 0$, 以及单调递增的非负函数 $g(x)$, 有

$$P(X \geqslant \epsilon) \leqslant \frac{E[g(X)]}{g(\epsilon)} \ .$$

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

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

证明 根据 Markov 不等式有

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

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 $P(|X+Y| \geqslant 6)$ 的上界.

解 根据随机变量 X 和 Y 的相关系数为 -1 可知

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

由 $E[X+Y]=0$, 利用 Chebyshev 不等式有

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

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

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

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

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

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

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

另一方面, 对任意 $t>0$ 有

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

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 $X_1, X_2, \ldots, X_n$ 满足 $E(X_i)=\mu$ 和 $\mathrm{Var}(X_i) \leqslant \sigma^2$, 对任意实数 $\epsilon>0$ 有

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

证明 根据 Chebyshev 不等式有

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

而独立同分布的假设有

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

由此得到

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

从而完成证明.

例 7.3 设分类器 f 在训练集 $S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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 $X_i = \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i]$ $(i \in [n])$, 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 $X_i \sim \mathrm{Ber}(p)$ 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

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

取 $\epsilon=\hat{p}/10$ 有

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

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

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

证明 根据凸函数性质有

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

引理得证.

根据 Young 不等式可证明著名的 Hölder 不等式.

引理 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

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

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

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

对上式两边同时取期望有

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

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.