第 6 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 6.1　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.

6.1 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 6.1 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 6.1 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 6.2 (Chebyshev 不等式) 设随机变量 X 的均值为 μ, 则有

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

证明 根据 Markov 不等式有

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

例 6.2 设随机变量 $X \sim N(-1,2)$ 和 $Y \sim N(1,8)$, 且 X 和 Y 的相关系数为 -1, 利用 Chebyshev 不等式求 $P(|X+Y| \geqslant 6) \leqslant$?

解 根据随机变量 X 和 Y 的相关系数为 -1 可知

$$\mathrm{Cov}(X,Y) = -\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} = -4.$$

由 $E[X+Y]=0$, 利用 Chebyshev 不等式有

$$\begin{aligned} P(|X+Y| \geqslant 6) &= P(|X+Y-E[X+Y]| \geqslant 6) \\ &\leqslant \mathrm{Var}(X+Y)/36 = (\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y))/36 = 1/18. \end{aligned}$$

下面补充一个Chebysheve不等式的应用例子:

例 6.3 设随机变量X和Y满足$E(X)=-2$, $E(Y)=2$, $\mathrm{Var}(X)=1$, $Var(Y)=4$, $\rho_{XY}=-1/2$. 利用Chebyshev不等式估计$\Pr(|X+Y| \geqslant 6)$的上界.

解 根据期望的线性关系有$E[X+Y]=0$, 根据相关系数的定义有

$$\rho_{XY} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{Var(X)}\sqrt{Var(Y)}} = -\frac{1}{2}.$$

由此可得$Var(X+Y) = Var(X)+Var(Y)+2E_{XY}[X-E(X)][Y-E(Y)] = 3$. 根据Chebysheve不等式有$\Pr\{|X+Y| \geqslant 6\} \leqslant \mathrm{Var}(X+Y)/36 = 1/12$.

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 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 $Var(Y) = \sigma^2$. 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 6.2 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 $Var(X_i) \leqslant \sigma^2$, 对任意 $\epsilon > 0$ 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

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

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 6.4 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

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

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

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

取 $\epsilon = \hat{p}/10$ 有

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

引理 6.2 (Young 不等式) 给定常数 $a > 0, b > 0$, 对满足 $1/p + 1/q = 1$ 的实数 $p > 0, q > 0$ 有

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

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 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

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

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

证明 设 $c = (E(|X|^p))^{\frac{1}{p}}$ 和 $d = (E(|Y|^q))^{\frac{1}{q}}$, 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.

6.2 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 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 $M_X(t)$, 对任意 $n \geqslant 1$ 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 Tayler 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 6.4 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 6.3 对任意独立的随机变量 X 和 Y 有 $M_{X+Y}(t) = M_X(t)M_Y(t)$.