第 7 章　集中不等式 (Concentration)

泛化性是机器学习中最重要的研究问题之一, 研究从训练数据中得到的模型能否很好地处理未见过的新数据, 相关研究与概率统计中的 Concentration 密切相关, 也与大数定律和中心极限定理相关. 本章将介绍相关知识, 以及应用这些知识解决实际问题. 在机器学习中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 (label). 假设 $\mathcal{D}$ 是空间 $\mathcal{X} \times \mathcal{Y}$ 的一个未知不可见的数据分布, 机器学习的经典假设是训练数据集 S_n 中每个样本 $(\boldsymbol{x}_i, y_i)$ 是根据数据分布 $\mathcal{D}$ 独立同分布采样所得.

给定分类器 $f\colon \mathcal{X} \to \{0, 1\}$, 考虑分类器 f 在一个样本点 $(\boldsymbol{x}_i, y_i)$ 分类情况, 设

$$X_i = \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i) = \begin{cases} 0 & \text{当 } f(\boldsymbol{x}_i) = y_i \text{ (分类正确)} \\ 1 & \text{当 } f(\boldsymbol{x}_i) \neq y_i \text{ (分类错误)}, \end{cases} \tag{7.1}$$

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. 根据独立性假设可知 $X_1, X_2, \cdots, X_n$ 是相互独立的. 分类器 f 在训练数据集 S_n 的分类错误率 (称为 ‘训练错误率’) 为

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i).$$

在应用中更关心分类器 f 对未见的数据分布 $\mathcal{D}$ 上的分类错误率, 称之为 ‘泛化错误率’, 即

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

由于数据分布 $\mathcal{D}$ 未见不可知, 不能直接利用数据分布计算期望 $E[X]$, 而训练错误率 $\sum_{i=1}^{n} X_i/n$ 已知, 因而问题转变为如何基于训练错误率来估计期望 $E[X]$? 在概率统计中可表述为

$$P_{S_n\sim\mathcal{D}^n}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| \geqslant \epsilon\right) \text{ 是否足够小?}$$

即能否以很大的概率给 $E[X]$ 的一个有效估计

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| < \epsilon.$$

例 7.1 假设训练集 $S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ 中每个元素根据分布 $\mathcal{D}$ 独立采样所得, 若分类器 f 在训练集 S_n 上全分类正确, 求分类器 f 在分布 $\mathcal{D}$ 上的正确率.

解 设随机变量 $X_i = \mathbb{I}[f(\boldsymbol{x}_i) \neq y_i] \quad (i \in [n])$, 根据训练数据集 S_n 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. 分类器 f 在分布 $\mathcal{D}$ 上的错误率

$$p = E[X] = E_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathbb{I}[f(\boldsymbol{x}) \neq y]] = P_{(\boldsymbol{x},y)\sim\mathcal{D}}(f(\boldsymbol{x}) \neq y).$$

不妨设 $p > \epsilon$, 则分类器 f 在训练集 S_n 分类错误率为零的概率为

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

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. 设 $\delta = \exp(-n\epsilon)$ 求解出 $\epsilon = \ln(1/\delta)/n$, 于是至少以 $1-\delta$ 的概率有分类器 f 在训练集 S_n 分类正确且

$$p = E[X] = E_{(\boldsymbol{x},y)\sim\mathcal{D}}[\mathbb{I}[f(\boldsymbol{x}) \neq y]] \leqslant \frac{\ln(1/\delta)}{n}.$$

7.1 基础不等式

本节给出一些基础不等式, 他们是机器学习和计算机科学研究的基础分析工具.

定理 7.1 (Markov 不等式) 对任意非负的随机变量 X 和常数 $\epsilon > 0$ 有

$$P(X \geqslant \epsilon) \leqslant E[X]/\epsilon.$$

证明 考虑随机事件 $A = \{X \geqslant \epsilon\}$ 和其对立事件 $\bar{A} = \{X < \epsilon\}$, 根据全期望公式有

$$\begin{aligned}
E[X] &= E[X|A]P(A) + E[X|\bar{A}]P(\bar{A}) \\
&= E[X|X \geqslant \epsilon]P(X \geqslant \epsilon) + E[X|X < \epsilon]P(X < \epsilon) \\
&\geqslant P(X \geqslant \epsilon)\epsilon,
\end{aligned}$$

这里利用了 $E[X|X \geqslant \epsilon] \geqslant \epsilon$ 和 $E[X|X < \epsilon]P(X < \epsilon) \geqslant 0$.

根据 Markov 不等式, 可以推导出一系列有用的不等式.

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

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

最常用的函数包括 $g(t) = e^t$ 和 $g(t) = t^2$, 例如考虑 $g(t) = t^2$ 很容易得到 Chebyshev 不等式.

定理 7.2 (Chebyshev 不等式) 对任意随机变量 X 和常数 $\epsilon \geqslant 0$ 有

$$P\left(|X - E[X]| > \epsilon\right) \leqslant \mathrm{Var}(X)/\epsilon^2.$$

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

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

Cantelli 不等式是一种比 Chebyshev 更紧的不等式, 又被成为单边 Chebyshev 不等式.

定理 7.3 (Cantelli 不等式) 对任意随机变量 X 和常数 $\epsilon > 0$ 有

$$P(X - E[X] \geqslant \epsilon) \leqslant \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \epsilon^2} \quad \text{和} \quad P(X - E[X] \leqslant -\epsilon) \leqslant \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \epsilon^2}.$$

证明 设随机变量 $Y = X - \mu$, 容易得到 $E[Y] = 0$ 以及 $\mathrm{Var}(Y) = \mathrm{Var}(X)$. 对任意 $t > 0$ 有

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

上面的不等式对任意的实数 $t > 0$ 都成立, 因此对不等式的右边求最小值, 即

$$P(X - \mu \geqslant \epsilon) \leqslant \min_{t>0}\left\{\frac{\mathrm{Var}(X) + t^2}{(\epsilon+t)^2}\right\} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \epsilon^2},$$

其中当 $t=\mathrm{Var}(X)/\epsilon$ 时取得最小值. 另一方面, 对任意 $t>0$ 有

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

再求解 t 使得上式达到最小, 从而完成证明.

基于 Chebyshev 不等式可以发现多个随机变量的均值与期望之间的关系.

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 $\mathrm{Var}(X_i)\leqslant\sigma^2$, 则对任意常数 $\epsilon>0$ 有

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

证明 根据 Chebyshev 不等式有

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

根据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

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

从而完成证明.

下面给出著名的 Hölder 不等式.

定理 7.4 (Hölder 不等式) 若正实数 p,q 满足 $1/p+1/q=1$, 则对任意随机变量 X 和 Y 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 根据凸函数的性质和 $1/p+1/q=1$, 对任意实数 $a>0$ 和 $b>0$ 有

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

设 $a = (E[|X|^p])^{1/p}$ 和 $b = (E[|Y|^q])^{1/q}$, 根据上述不等式有

$$\frac{|XY|}{ab} = \frac{|X|}{a}\frac{|Y|}{b} \leqslant \frac{1}{p}\frac{|X|^p}{a^p} + \frac{1}{q}\frac{|Y|^q}{b^q}.$$

对上式两边同时取期望有

$$\frac{E[|XY|]}{ab} \leqslant \frac{1}{p}\frac{E[|X|^p]}{c^p} + \frac{1}{q}\frac{E[|Y|^q]}{d^q} = \frac{1}{p} + \frac{1}{q} = 1,$$

从而完成证明.

7.2 Chernoff 不等式

Chernoff 方法 是证明集中不等式一种非常基本有效的方法, 这里首先介绍基本原理: 设 Y 是一个随机变量, 对给定任意 $t > 0$ 和 $\epsilon > 0$, 利用 Markov 不等式有

$$\begin{aligned} P\left[Y \geqslant E[Y] + \epsilon\right] &= P\left[tY \geqslant tE[Y] + t\epsilon\right] \\ &= P\left[\exp(tY) \geqslant \exp(tE[Y] + t\epsilon)\right] \leqslant \exp(-t\epsilon - tE[Y])E[\exp(tY)], \end{aligned}$$

上面的不等式对任意 $t > 0$ 都成立, 于是有

$$P\left[Y \geqslant E[Y] + \epsilon\right] \leqslant \min_{t>0}\left\{\exp(-t\epsilon - tE[Y])E[\exp(tY)]\right\}.$$

另一方面, 对任意 $\epsilon > 0$ 和 $t < 0$ 有

$$\begin{aligned} P\left[Y \leqslant E[Y] - \epsilon\right] &= P\left[tY \geqslant tE[Y] - t\epsilon\right] \\ &= P\left[\exp(tY) \geqslant \exp(tE[Y] - t\epsilon)\right] \leqslant \exp(t\epsilon - tE[Y])E[\exp(tY)], \end{aligned}$$

同理因为上面的不等式对任意 $t > 0$ 都成立, 于是有

$$P[Y \leqslant E[Y] - \epsilon] \leqslant \min_{t>0}\left\{\exp(t\epsilon - tE[Y])E[\exp(tY)]\right\}.$$

可以发现在此过程中还需要计算 $E[e^{tX_i}]$, 此时将根据不同的随机变量给出不同的计算方法. 下面来看一些具体的随机变量.

定理 7.5 设独立同分布的随机变量 $X_1, \cdots, X_n$ 满足 $P(X_i = +1) = P(X_i = -1) = 1/2$, 则有

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