3 Probabilistic Error Bounds

This section examines the difference between the **empirical error rate** and the **true error rate** of classification decision rules in a class $\mathcal{F}$. In particular, the tail probability of the uniform deviation, $\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$, is studied. Upper bounds of the tail probability lead to probabilistic error (or risk) bounds of any $f \in \mathcal{F}$ in the PAC learning framework. For a finite class $\mathcal{F}$, simple error bounds are derived in this section.

Recall that

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n} I(f(X_i) \neq Y_i) \text{ and } R(f) = P(f(X) \neq Y).$$

$R_n(f)$ is the expectation of $I(f(X) \neq Y)$ with respect to the empirical distribution $\mathcal{P}_n$, which puts a probability mass $1/n$ at (x_i, y_i) while $R(f)$ is with respect to the true distribution $\mathcal{P}_{X,Y}$. One can regard $\{R_n(f) - R(f) : \ f \in \mathcal{F}\}$ as a stochastic process indexed by $\mathcal{F}$. Such a process is called the empirical process.

3.1 Hoeffding's Inequality

First, the tail probability of $R_n(f) - R(f)$ is examined for a fixed f. As it concerns the deviation of an average $R_n(f)$ from its mean $R(f)$, the following Hoeffding's inequality (Hoeffding, JASA 1963) is relevant. It deals with concentration of the sum of independent bounded random variables around its expectation.

Hoeffding's Inequality Let $X_1, \ldots, X_n$ be independent and bounded random variables such that $X_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^n X_i$. Then for $\epsilon > 0$, we have

$$\begin{aligned} P(S_n - ES_n \geq \epsilon) &\leq \exp\left\{-\frac{2\,\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}, \\ \text{and} \quad P(S_n - ES_n \leq -\epsilon) &\leq \exp\left\{-\frac{2\,\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}. \end{aligned} \tag{3.1}$$

Remark 3. The inequality implies

$$P(|S_n - ES_n| \geq \epsilon) \leq 2\exp\left\{-\frac{2\,\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}. \tag{3.2}$$

With $0 \leq I(f(X_i) \neq Y_i) \leq 1$, it gives

$$P(|R_n(f) - R(f)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2).$$

Thus, the tail probability of the deviation decays exponentially to zero.

In order to prove Hoeffding's inequality we need the following lemma:

Lemma 1. *Let X be a random variable with $EX = 0$ and $a \leq X \leq b$. Then for any $s \in \mathbb{R}$,*

$$Ee^{sX} \leq e^{s^2(b-a)^2/8}.$$

Proof. By convexity of $g(x) := \exp(sx)$,

$$g(x) = g\left(\frac{x-a}{b-a}b + \frac{b-x}{b-a}a\right) \le \frac{x-a}{b-a}g(b) + \frac{b-x}{b-a}g(a) = \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

Therefore, with $p := -a/(b-a)$,

$$\begin{aligned} Ee^{sX} &\le p\,e^{sb} + (1-p)e^{sa} = e^{sa}\big[(1-p) + p\,e^{s(b-a)}\big] \\ &= e^{-sp(b-a)}\big[(1-p) + p\,e^{s(b-a)}\big] \\ &= e^{-pu}\big[(1-p) + p\,e^{u}\big] \quad \text{letting } u := s(b-a) \text{ and } \phi(u) := -pu + \log(1-p+p\,e^u) \\ &= e^{\phi(u)}. \end{aligned}$$

Taylor's expansion of $\phi(u)$ at $u = 0$ gives

$$\phi(u) = \phi(0) + \phi'(0)u + \frac{1}{2}\phi''(\theta)u^2 \quad \text{for some } \theta \text{ between 0 and } u.$$

Here

$$\begin{aligned} \phi(0) &= 0, \\ \phi'(u) &= -p + \frac{p\,e^u}{1-p+p\,e^u} = -p + \frac{p}{(1-p)e^{-u}+p} \Rightarrow \phi'(0) = 0, \\ \phi''(u) &= \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \le \frac{1}{4} \quad \left(\because (y+z)^2 \ge 4yz\right). \end{aligned}$$

Therefore, $\phi(u) = \frac{1}{2}\phi''(\theta)u^2 \le u^2/8$, which implies

$$Ee^{sX} \le e^{u^2/8} = e^{s^2(b-a)^2/8}.$$

□

Proof of Hoeffding's inequality: For any $s > 0$, and any random variable X, Markov's inequality states that

$$P(X \ge \epsilon) = P(e^{sX} \ge e^{s\epsilon}) \le Ee^{sX}/e^{s\epsilon}.$$

Thus for any $s > 0$,

$$P(S_n - ES_n \ge \epsilon) \le \frac{Ee^{s(S_n - ES_n)}}{e^{s\epsilon}} = e^{-s\epsilon}\prod_{i=1}^{n} E\,e^{s(X_i - EX_i)}.$$

The method we use below is generally known as the *Chernoff's bounding technique*. For $i = 1, \ldots, n$ and $Y_i = X_i - EX_i$, we have

$$EY_i = 0 \quad \text{and} \quad a_i - EX_i \le Y_i \le b_i - EX_i.$$

Thus, using Lemma 1 we get

$$\begin{aligned} P(S_n - ES_n \ge \epsilon) &\le e^{-s\epsilon}\prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8} \\ &= e^{-s\epsilon}\exp\left\{s^2\sum_{i=1}^{n}(b_i-a_i)^2/8\right\} \\ &= \exp\left\{-\epsilon s + cs^2\right\} \quad \text{with } c = \sum_{i=1}^{n}(b_i-a_i)^2/8 \ge 0. \end{aligned}$$

Now letting $\phi(s) := -\epsilon s + cs^2$, find the minimizer of $\phi(s)$. Since $\phi'(s) = -\epsilon + 2cs$, $s = \epsilon/2c(> 0)$ minimizes ϕ. At the minimizer, $\phi(\epsilon/2c) = -\epsilon^2/(4c)$. Therefore,

$$P(S_n - ES_n \geq \epsilon) \leq \exp\big(-\frac{\epsilon^2}{4c}\big) = \exp\Big\{-\frac{2\,\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big\}.$$

□

3.2 Uniform Deviation for a Finite Class

Theorem 4. *Assume that $\mathcal{F}$ contains finitely many classifiers, $|\mathcal{F}| = N$. Then for any $\epsilon > 0$,*

$$P\left(\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \geq \epsilon\right) \leq 2N\exp(-2n\epsilon^2).$$

Proof. Since $\mathcal{F}$ is finite, we can use the union bound to obtain

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \geq \epsilon\right) &\leq \sum_{i=1}^N P\big(|R_n(f_i) - R(f)| \geq \epsilon\big) \\ &\leq \sum_{i=1}^N 2\exp(-2n\epsilon^2) = 2N\exp(-2n\epsilon^2). \end{aligned}$$

□

Remark 4. Setting the upper bound to δ, we have

$$2N\exp(-2n\epsilon^2) = \delta \Rightarrow \epsilon = \sqrt{\frac{1}{2n}\log(\frac{2N}{\delta})} = \sqrt{\frac{1}{2n}\Big(\log N + \log\frac{2}{\delta}\Big)}.$$

That is, with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \leq \sqrt{\frac{1}{2n}\Big(\log N + \log(2/\delta)\Big)}.$$

The extra $\log N$ term accounts for the fact that N bounds hold simultaneously. This provides a simple description of how the size of $\mathcal{F}$ affects the uniform error bound. Also, note that the actual distribution of the data did not play a role in the derivation, i.e. the probabilistic bound is distribution-free.

Lemma 2. *If a non-negative random variable Z satisfies*

$$P(Z > t) \leq c\exp(-2nt^2) \text{ for every } t > 0 \text{ and some } c \geq 1,$$

then

$$EZ^2 \leq \frac{1}{2n}\log(ce).$$

Furthermore,

$$EZ \leq \sqrt{EZ^2} \leq \sqrt{\frac{1}{2n}\log(ce)}.$$

Proof. For any $0 < u < \infty$,

$$\begin{aligned} EZ^2 &= \int_0^\infty P(Z^2 > t)dt = \int_0^u P(Z^2 > t)dt + \int_u^\infty P(Z > \sqrt{t})dt \\ &\le u + \int_u^\infty c\exp(-2nt)dt = u + \frac{c}{2n}\exp(-2nu). \end{aligned}$$

Define $\phi(u) := u + \frac{c}{2n}\exp(-2nu)$. Since the above inequality is valid for any $u > 0$, we would obtain the best bound by using the minimum value of $\phi(u)$.

$$\phi'(u) = 1 - c\exp(-2nu) = 0 \Rightarrow u = \frac{\log c}{2n} \quad \text{and } \phi''(u) = 2nc\exp(-2nu) > 0.$$

Thus,

$$\min_{u>0}\phi(u) = \phi\left(\frac{\log c}{2n}\right) = \frac{1}{2n}\log(ce).$$

The second inequality is easily obtained by Jensen's inequality. □

Theorem 4 states an exponential inequality for the uniform deviation $\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|$. By the above lemma, we have its convergence to zero in the mean.

$$\begin{aligned} & P\left(\sup_{f\in\mathcal{F}}|R_n(f) - R(f)| \ge \epsilon\right) \le 2N\exp\{-2n\epsilon^2\} \\ \Rightarrow \qquad & E\left(\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|\right) \le \sqrt{\frac{1}{2n}\log(2Ne)} = O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

So, the rate of convergence of the deviation of the empirical error from the true error to zero is $n^{-1/2}$ when the class of decision rules is finite. The exponential inequality can be used to further establish almost sure convergence as well.

$$\begin{aligned} & P\left(\sup_{f\in\mathcal{F}}|R_n(f) - R(f)| \ge \epsilon\right) \le 2N\exp(-2n\epsilon^2) \\ \Rightarrow \qquad & \sum_{n=1}^\infty P\left(\sup_{f\in\mathcal{F}}|R_n(f) - R(f)| \ge \epsilon\right) \le 2N\sum_{n=1}^\infty\{\exp(-2\epsilon^2)\}^n < \infty. \end{aligned}$$

Then by the Borel-Cantelli lemma,

$$P\left(\sup_{f\in\mathcal{F}}|R_n(f) - R(f)| \ge \epsilon \text{ i.o.}\right) = 0.$$

Therefore $\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|$ converges to 0 almost surely.