5

Finite Hypothesis Sets

In this chapter we assume that the set of classifiers $\mathcal{H} = \{h_1, \ldots, h_K\}$ is finite. Let $\hat{h}_n$ be a minimizer of the empirical risk in $\hat{h}_n$. We are interested in how well this classifier performs in relation to the best possible risk $R(h)$ for $h \in \mathcal{H}$:

$$R(\hat{h}_n) - \min_{h \in \mathcal{H}} R(h).$$

We will derive a probabilistic bound on this difference that depend logarithmically on the size $K = |\mathcal{H}|$ and decays with order $n^{-1/2}$. This bound makes use of Hoeffding's inequality, and serves as a prototype for a whole range of similar risk bounds in statistical learning theory.

Finite Hypothesis Sets

The following theorem gives us an effective bound on the estimation error.

Theorem 5.1. *Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite dictionary. Then for $\delta \in (0, 1)$,*

$$\mathbb{P}\left(R(\hat{h}_n) - \inf_{h \in \mathcal{H}} R(h) \leq \sqrt{\frac{2 \log\left(\frac{2K}{\delta}\right)}{n}} \right) \geq 1 - \delta.$$

Unless otherwise stated, the logarithm refers to the natural logarithm (though this is not important). This important result shows that (with high probability) we can bound the estimation error by a term that is *logarithmic* in the size of $\mathcal{H}$, and proportional to $n^{-1/2}$, where n is the number of samples. For fixed or moderately growing K, this error goes to zero as n goes to infinity.

Recall from Chapter 4 the inequality

$$R(\hat{h}_n) - \min_{h \in \mathcal{H}} R(h) \leq 2 \max_{h \in \mathcal{H}} |\hat{R}(h) - R(h)|, \tag{5.1}$$

where $\hat{R}$ is the empirical risk. As a first step towards bounding the supremum on the right-hand side, we need to bound the difference $|R(h) - \hat{R}(h)|$ of an individual, fixed h. The key ingredient for such a bound Hoeffding's inequality, which we recall here.

Theorem 5.2 (Hoeffding's Inequality)**.** *Let $Z_1, \dots, Z_n$ be independent random variables taking values in $[0,1]$, and let $\overline{Z}_n = (1/n)\sum_{i=1}^n Z_i$ be the average. Then for $t \geq 0$,*

$$\mathbb{P}\left(|\overline{Z}_n - \mathbb{E}[\overline{Z}_n]| > t\right) \leq 2\mathrm{e}^{-2nt^2}.$$

Using Hoeffding's Inequality we obtain the following bound on the difference between the empirical risk and the risk of a classifier.

Lemma 5.3. *For any classifier h and $\delta \in (0,1)$,*

$$|\hat{R}(h) - R(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

holds with probability at least $1-\delta$.

Proof. Set $Z_i = \mathbf{1}\{h(X_i) \neq Y_i\}$. Then

$$\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i = \hat{R}(h),$$
$$\mathbb{E}[\overline{Z}_n] = \mathbb{E}[\hat{R}(h)] = R(h),$$

and the Z_i satisfy the conditions of Hoeffding's inequality. Set $\delta = 2\mathrm{e}^{-2nt^2}$ and resolve for t, which gives

$$t = \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Hence, by Hoeffding's inequality,

$$\mathbb{P}(|\hat{R}(h) - R(h)| > t) = \mathbb{P}(|\overline{Z}_n - \mathbb{E}[\overline{Z}_n]| > t) \leq \delta$$

and therefore, by taking the complement,

$$\mathbb{P}(|\hat{R}(h) - R(h)| \leq t) = 1 - \mathbb{P}(|\hat{R}(h) - R(h)| > t) \geq 1 - \delta,$$

which was claimed. □

Proof of Theorem 5.1. Courtesy of (5.1), the goal is to bound the supremum

$$\max_{h \in \mathcal{H}} |\hat{R}(h) - R(h)|. \tag{5.2}$$

For this, we use the union bound. Indeed, for each h_i we can apply Lemma 5.3 with δ/K to show that

$$\mathbb{P}\left(|\hat{R}(h_i) - R(h_i)| > \frac{t}{2}\right) \leq \frac{\delta}{K},$$

where $t = \sqrt{\frac{2\log(2K/\delta)}{n}}$. The probability of (5.2) being bounded by t can be expressed equivalently as

$$\begin{aligned}&\mathbb{P}\left(\max_{h\in\mathcal{H}}|\hat{R}(h) - R(h)| \leq t/2\right)\\ &\quad = \mathbb{P}(|\hat{R}(h_1) - R(h_1)| \leq t/2, \ldots, |\hat{R}(h_K) - R(h_K)| \leq t/2).\end{aligned}$$

Since the right-hand side is an *intersection* of events, the *complement* of this event is the *union* of the events $|\hat{R}(h_i) - R(h_i)| > t$, and we can apply the union bound:

$$\begin{aligned}\mathbb{P}\left(\max_{h\in\mathcal{H}}|\hat{R}(h) - R(h)| > t/2\right) &\leq \sum_{i=1}^{K}\mathbb{P}(|\hat{R}(h_i) - R(h_i)| > t/2)\\ &\leq K\cdot\frac{\delta}{K} = \delta.\end{aligned}$$

Therefore, with probability at least $1-\delta$ we have

$$2\max_{h\in\mathcal{H}}|\hat{R}(h) - R(h)| \leq \sqrt{\frac{2\log(2K/\delta)}{n}},$$

and using (5.1) the claim follows. □

Generalization bounds and noise

One drawback of the bound in Theorem 5.1 is that it does not take into account properties of the underlying distribution. In particular, it is the same in the case where Y is completely determined by X as it is in the case in which Y is completely independent on X. Intuitively, in the first situation we would hope to get better rates of convergence than in the second. Using more refined concentration inequalities such as the Bernstein inequality, that take into account the variance of the random variables, we can get better rates of convergence in some situation.

Recall the definition of the regression function $f(X) = \mathbb{E}[Y|X]$ and the associated error $\epsilon = Y - f(X)$. Let $\gamma \in (0, 1/2]$ and assume that

$$|f(X) - 1/2| \geq \gamma$$

almost surely. This condition is known as **Massart's noise condition**. If $\gamma = 1/2$, then $f(X)$ is either 1 or 0 and we are in the deterministic case, where Y is completely determined by X. If, on the other hand, $\gamma \approx 0$, then we are barely placing any restrictions on $f(X)$, and we are allowing for the case where $f(X)$ is close to 0, and hence where Y is almost independent of X.

Theorem 5.4. *Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite dictionary and assume that $h^* \in \mathcal{H}$, where h^* is the Bayes classifier. Then for $\delta \in (0, 1)$,*

$$\mathbb{P}\left(R(\hat{h}_n) - R(h^*) \leq \frac{\log\left(\frac{K}{\delta}\right)}{\gamma n}\right) \geq 1 - \delta.$$

Note that the bound decays with order n^{-1}, rather than $n^{-1/2}$ as with the bound derived from Hoeffding's inequality. The proof of this result relies on a concentration of measure result similar to Hoeffding's inequality, called Bernstein's inequality.

Theorem 5.5 (Bernstein Inequality)**.** *Let $\{Z_i\}_{i=1}^n$ be centred random variables (that is, $\mathbb{E}[Z_i] = 0$) with $|Z_i| \leq c$ and set $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Z_i)$. Then for $t > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i > t\right) \leq \exp\left(-\frac{nt^2}{2(\sigma^2 + ct/3)}\right).$$

We outline the idea of the proof. The proof proceeds by defining the random variables

$$Z_i(h) = \mathbf{1}\{h^*(X_i) \neq Y_i\} - \mathbf{1}\{h(X_i) \neq Y_i\}$$

for each $h \in \mathcal{H}$. The average and expectation of these random variables is then

$$\begin{aligned}\hat{R}(h^*) - \hat{R}(h) &= \frac{1}{n}\sum_{i=1}^n Z_i(h),\\ R(h^*) - R(h) &= \mathbb{E}[Z_i(h)].\end{aligned}$$

Based on this, one gets a bound

$$\begin{aligned}R(\hat{h}_n) - R(h^*) &\leq \frac{1}{n}\sum_{i=1}^n (\underbrace{Z_i(\hat{h}_n) - \mathbb{E}[Z_i(\hat{h}_n)]}_{\overline{Z}_i(\hat{h}_n)})\\ &\leq \max_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \overline{Z}_i(h).\end{aligned} \tag{5.3}$$

The random variables $\overline{Z}_i(h)$ are centred, satisfy $|Z_i(h)| \leq 2 =: c$ and we can bound the variance by

$$\mathrm{Var}(\overline{Z}_i(h)) = \mathrm{Var}(Z_i(h)) \leq \mathbb{P}(h(X_i) \neq h^*(X_i)) =: \sigma^2(h).$$

We can now apply Bernstein's inequality to the probability that the sum (5.3) exceeds a certain value for each individual h, and use a union bound to get a corresponding bound for the maximum that involves the variance $\sigma^2(\hat{h}_n)$. Using the property that $h^* \in \mathcal{H}$, one can also derive a lower bound on the excess risk in terms of the variance, and hence combine both bounds to get the desired result.

Notes

The application of Hoeffding's inequality is a standard tool in statistical learning, and references are [1, Section 8.2] and [2, Chapter 4]. The bounds on derived in this chapter are bounds on the difference between the empirical risk and the generalization risk over the whole class $\mathcal{H}$, a setting usually also referred to as **uniform convergence**.

[1] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[2] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.