A second look at the theory of generalization

In these notes, we give (most of) a careful proof of one of the most important results in the theory of machine learning[1]. We are given (randomly generated, independent) data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ labeled with classes $y_1, \ldots, y_n$. For each classifier h in our (now infinite) set of candidates $\mathcal{H}$, the empirical risk is given by

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad Z_i = \begin{cases} 1, & h(\boldsymbol{x}_i) \neq y_i, \\ 0, & h(\boldsymbol{x}_i) = y_i \end{cases}.$$

Recall that the (unknown) true risk of h is simply

$$R(h) = \mathbb{P}\left[h(\boldsymbol{x}) \neq y\right].$$

This is also called the *generalization error* or *out of sample error* as it is precisely the probability that h makes a mistake when given a randomly generated feature vector and label $(\boldsymbol{x}, y)$.

The following result, formulated and proved by Vapnik and Chervonenkis in 1971, gives a bound on the difference between the empirical risk and the true risk that holds **uniformly** over all $h \in \mathcal{H}$. It of course depends on the number of data points n and some measure of the complexity of the set $\mathcal{H}$ — the latter is captured by the *growth function* $m_{\mathcal{H}}(2n)$, which we recall is given by:

$$m_{\mathcal{H}}(2n) = \begin{array}{c}\text{most dichotomies that } \mathcal{H} \text{ can} \\ \text{generate on any set of } 2n \text{ points.}\end{array}$$

Here is the theorem:

[1] This material comes from the Appendix of *Learning from Data* by Abu-Mostafa et al.

Theorem 1 *Let $(\boldsymbol{x}_i, y_i)$ be a training set consisting of independent samples, and $\mathcal{H}$ be an arbitrary ensemble of classifiers. Then*

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \;\leq\; 4\, m_{\mathcal{H}}(2n)\, e^{-\frac{\epsilon^2 n}{8}}.$$

Just as in the finite case, this uniform bound allows us to compare the classifier we chose from h using empirical risk minimization to the one we would have chosen given oracle knowledge of the true risks $R(h)$. If

$$h^\sharp = \arg\inf_{h\in\mathcal{H}} R(h),$$
$$h^* = \arg\inf_{h\in\mathcal{H}} \widehat{R}_n(h),$$

then following what we did in Notes 2,

$$|R(h^*) - R(h^\sharp)| \;\leq\; 2\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)|.$$

Proving Theorem 1 moves in three steps; we provide a lemma below for each of these steps:

Step 1. Replace the true risk with the empirical risk of *another* independently drawn data set of size n. Lemma 1 below shows that for the relevant range of ϵ and n,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \;\leq\; 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\right],$$

where $\widehat{R}'_n(h)$ is the empirical risk for another independent data set of size n. This allows us to work with the complexity of $\mathcal{H}$ relative to a finite number of data points, allowing us to use things like dichotomies, shattering numbers, and growth functions.

Step 2. Replace the two randomly drawn sets of size n with a fixed set $\mathcal{S}$ of size that is randomly partitioned into two sets of size n. We then look at the worst case $\mathcal{S}$ among all choices of $2n$ vectors in $\mathbb{R}^d$. Lemma 2 below establishes that

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|\widehat{R}_n(h)-\widehat{R}'_n(h)|>\frac{\epsilon}{2}\right] \;\leq\; m_{\mathcal{H}}(2n)\cdot\sup_{\mathcal{S}}\sup_{h\in\mathcal{H}}\mathbb{P}\left[|\widehat{R}_n(h)-\widehat{R}'_n(h)|>\frac{\epsilon}{2}\mid\mathcal{S}\right]$$

The probabilities on the left and right hand sides above use different models for generating the random data samples $(\boldsymbol{x}_i, y_i)$. On the left, we draw two sets independently of size n, compute the empirical risks for each, then compare these. On the right, we fix a set $\mathcal{S}$ of $2n$ points, then randomly divide it in two subsets of size n, and then compare the empirical risks. We end up considering the worst (largest) such probability over subsets $\mathcal{S}$ of size $2n$.

The important part of this step is that we have moved the supremum over $\mathcal{H}$ from inside the probability to outside the probability, while incurring a cost of only $m_{\mathcal{H}}(2n)$.

Step 3. Developing a Hoeffding-like bound for comparing the two empirical risks above for a fixed h. Lemma 3 tells us that

$$\mathbb{P}\left[|\widehat{R}_n(h)-\widehat{R}'_n(h)|>\frac{\epsilon}{2}\mid\mathcal{S}\right] \;\leq\; 2e^{-\epsilon^2 n/8},$$

uniformly, no matter what set $\mathcal{S}$ or classifier h we choose.

Theorem 1 just combines these three results. By moving the terms around, we can state the result as an upper bound on the worst case generalization error; with probability at least $1-\delta$, we have

$$R(h) \;\leq\; \widehat{R}_n(h)+\sqrt{\frac{8}{n}\left(\log(m_{\mathcal{H}}(2n))+\log\left(\frac{4}{\delta}\right)\right)}$$

for all $h\in\mathcal{H}$ simultaneously.

The first lemma replaces the true risk with the risk on an independent data set of size n.

Lemma 1 *Suppose that* $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ *and that* $\mathcal{D}' = \{(\boldsymbol{x}_{n+1}, y_{n+1}), \dots, (\boldsymbol{x}_{2n}, y_{2n})\}$ *are independent sets of labeled training data, and let* $\widehat{R}_n(h)$ *be the empirical risk for a classifier* h *computed on* $\mathcal{D}$*, and* $\widehat{R}'_n(h)$ *be the same computed on* $\mathcal{D}'$*. Then for* ϵ *and* n *such that* $e^{-\epsilon^2 n/2} < \frac{1}{4}$,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| > \epsilon\right] \ \le\ 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2}\right].$$

Note that the condition on ϵ and n is what we need for the bound in Theorem 1 to be meaningful anyway (i.e., what we need to make the bound on the probability less than 1).

We will not give a full proof of this lemma, as it is more technical than enlightening. If you are interested, see *Learning from Data* (by Abu-Mostafa et al), Appendix A.1.

We will, however, give some quick reasoning about why it works. It is clear that $\widehat{R}_n(h)$ and $\widehat{R}'_n(h)$ are independent identically distributed random variables with mean $R(h)$. If this distribution was symmetric around its mean, then $R(h)$ would also be a median:

$$\mathbb{P}\left[\widehat{R}_n(h) < \mu\right] = \frac{1}{2}, \quad \text{and} \quad \mathbb{P}\left[\widehat{R}_n(h) \ge \mu\right] = \frac{1}{2},$$

and similarly for $\widehat{R}'_n(h)$. Suppose now that $\widehat{R}_n(h)$ is fixed, and it happens that $|\widehat{R}_n(h) - R(h)| > \epsilon$. Under this condition, what is the probability that $|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \epsilon$? Well, it is at least $\frac{1}{2}$, since that is the probability that $\widehat{R}'_n(h)$ falls on the other side of the mean

$R(h)$ as $\widehat{R}_n(h)$. Thus

$$\begin{aligned}\mathbb{P}\left[|\widehat{R}_n(h) - \widehat{R}_n'(h)| > \epsilon\right] &= \mathbb{P}\left[|\widehat{R}_n(h) - R(h)| > \epsilon\right] \\ &\quad \cdot \mathbb{P}\left[|\widehat{R}_n(h) - \widehat{R}_n'(h)| > \epsilon \;\mid\; |\widehat{R}_n(h) - R(h)| > \epsilon\right] \\ &\geq \frac{1}{2}\,\mathbb{P}\left[|\widehat{R}_n(h) - R(h)| > \epsilon\right]\end{aligned}$$

and the above would hold for all $h \in \mathcal{H}$ simultaneously.

In general, the distribution of $\widehat{R}_n(h)$ and $\widehat{R}_n'(h)$ will not in general be symmetric around its mean. But it almost is; these are binomial random variables, whose means and medians are not too far from one another. Proving the lemma above in full is basically about cleaning up this detail.

The second lemma relates the two different probability models; the one on the left is over two independent draws of size n, the one on the right is for a fixed data set of size $2n$ that gets randomly partitioned into two sets of size n.

Lemma 2 *Let $\mathcal{S}$ denote a set of (random) samples of size $2n$:*

$$\mathcal{S} = \{(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, 2n\}.$$

Then

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2}\right] \ \le\ m_{\mathcal{H}}(2n)\sup_{\mathcal{S}}\sup_{h\in\mathcal{H}}\mathbb{P}\left[\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right],$$

where the probability on the right-hand side is over a random partition of $\mathcal{S}$ into two training sets of size n, one of which is used to compute $\widehat{R}_n(h)$ and the other is used to compute $\widehat{R}'_n(h)$.

Proof First, we condition on the random draw of $\mathcal{S} \subset \mathbb{R}^d$:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2}\right] = \int f_X(\mathcal{S})\,\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right] d\mathcal{S}$$

Above, we are integrating over the $2n$ realizations of the $\boldsymbol{x}_i$; the notation means

$$f_X(\mathcal{S}) = f_X(\boldsymbol{x}_1), \ldots, f_X(\boldsymbol{x}_{2n}), \quad d\mathcal{S} = d\boldsymbol{x}_1 \cdots d\boldsymbol{x}_{2n},$$

where $f_X(\cdot)$ is the marginal density for the feature on $\mathbb{R}^d$. Since this density integrates to 1,

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2}\right] \ \le\ \sup_{\mathcal{S}}\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right],$$

where $\sup_{\mathcal{S}}$ is a supremum over all points sets of size $2n$ in $\mathbb{R}^d$.

Inside the supremum above, the set $\mathcal{S}$ is fixed. This means that there are only a finite number of dichotomies that $\mathcal{H}$ can implement on $\mathcal{S}$. For a particular $\mathcal{S}$, suppose that there are $m(\mathcal{S})$ different dichotomies possible, realized by $h_1, \ldots, h_m$. Then applying the union bound,

$$\begin{aligned}\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right] &= \mathbb{P}\left[\max_{h\in\{h_1,\ldots,h_m\}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right] \\ &\leq \sum_{m=1}^{m(\mathcal{S})} \mathbb{P}\left[\left|\widehat{R}_n(h_m) - \widehat{R}'_n(h_m)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right] \\ &\leq m(\mathcal{S}) \max_{h\in\{h_1,\ldots,h_m\}} \mathbb{P}\left[\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right]\end{aligned}$$

To make this bound uniform over all $\mathcal{S}$, we use the facts that $m(\mathcal{S}) \leq m_{\mathcal{H}}(2n)$, and that taking the maximum over $h_1, \ldots, h_m$ will of course give you something no larger than taking the supremum over all $h \in \mathcal{H}$. Thus

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right] \leq m_{\mathcal{H}}(2n) \cdot \sup_{\mathcal{S}} \sup_{h\in\mathcal{H}} \mathbb{P}\left[\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \mid \mathcal{S}\right].$$

This establishes the lemma. ■

The third and final is very similar to the generalization bound we developed (in the notes several weeks ago) for a single classifier using the Hoeffding inequality. The difference here is that we are comparing the risks between two different random sets of data (instead of one set of data and its mean), and the randomness is different — the data points are fixed here and then divided in half at random.

Lemma 3 *For any fixed classifier $h \in \mathcal{H}$ and any fixed set of $2N$ data points $\mathcal{S}$,*

$$\mathbb{P}\left[\left|\widehat{R}_n(h) - \widehat{R}'_n(h)\right| > \frac{\epsilon}{2} \;\mid\; \mathcal{S}\right] \;\leq\; 2\exp\left(-\frac{\epsilon^2 n}{8}\right),$$

where again the probability is with respect to a random partition of $\mathcal{S}$ into two training sets of size n.

Proof The results follows almost immediately from another classical lemma by Hoeffding. This Lemma 4 also describes how closely a sum of random variables is concentrated around its mean, but in this case the random variables are sampled *without replacement* from a finite set. This causes the random variables to be (weakly) dependent on one another. Nevertheless, the proof of this lemma[2] shows that this probability can be bounded by the independent case.

[2] See Hoeffding, *Probability inequalities for sums of bounded random variables*, 1967.

Lemma 4 *Let $\mathcal{A} = \{a_1, \ldots, a_{2n}\}$ be a fixed set of binary-valued points, $a_i \in \{0,1\}$, and let $\mu = \frac{1}{2n}\sum_{i=1}^{2n} a_i$ be their center of mass*[3]. *Let $\{Z_1, \ldots, Z_n\}$ be a set of n points chosen from $\mathcal{A}$ uniformly at random* without replacement *— this guarantees that the Z_i are unique. Then*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \geq \epsilon\right] \;\leq\; 2e^{-2\epsilon^2 n}.$$

With this lemma in place, the result follows quickly. With h fixed, we take the numbers $\{a_i\}$ above as indicators on whether or not h makes an error on data point $\boldsymbol{x}_i$:

$$a_i = \begin{cases} 1, & h(\boldsymbol{x}_i) \neq y_i \\ 0, & h(\boldsymbol{x}_i) = y_i, \end{cases} \qquad i = 1, \ldots, 2n.$$

We randomly divide the indexes $\{1, \ldots, 2n\}$ into two equal size sets, $\mathcal{I}$ and $\mathcal{I}'$, and then compute the empirical risk of h over the data in both sets of these index sets:

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i\in\mathcal{I}} a_n, \quad \widehat{R}'_n(h) = \frac{1}{n}\sum_{i\in\mathcal{I}'} a_n.$$

Since these are sums over the two partitions of the data, they are equivalent to the "sampling without replacement" described in the statement of Lemma 4. Since

$$\mu = \frac{1}{2n}\sum_{i=1}^{2n} a_n = \frac{\widehat{R}_n(h) + \widehat{R}'_n(h)}{2},$$

[3] We are using the term "center of mass" instead of "mean" since these points are fixed and not random.

we know that

$$\widehat{R}_n(h) - \mu = \mu - \widehat{R}'_n(h),$$

which means

$$|\widehat{R}_n(h) - \widehat{R}'_n(h)| = 2|\widehat{R}_n(h) - \mu|.$$

Then applying Lemma 4 yields

$$\begin{aligned}\mathbb{P}\left[|\widehat{R}_n(h) - \widehat{R}'_n(h)| > \frac{\epsilon}{2} \mid \mathcal{S}\right] &= \mathbb{P}\left[|\widehat{R}_n(h) - \mu| > \frac{\epsilon}{4} \mid \mathcal{S}\right] \\ &= 2\exp\left(-\frac{\epsilon^2}{8}\right),\end{aligned}$$

which establishes the lemma. ■