Machine Learning Notes

This note is based on Machine Learning at DIKU

15 December 2021

Chapter 3. Generalisation Bounds for Classification

3.3 Generalisation Bound for Finite Hypothesis Classes

A hypothesis set $\mathcal{H}$ containing a single hypothesis is a very boring set. In fact, we cannot learn in this case, because we end up with the same single hypothesis no matter what the sample S is. Learning becomes interesting when training sample S helps to improve future predictors or, equivalently, decrease the expected loss $L(h)$. In this section we consider the simplest non-trivial case, where $\mathcal{H}$ consists of a finite number of hypotheses M. There are at least two cases, where we meet a finite $\mathcal{H}$ in real life. The first is when the input space $\mathcal{H}$ is finite. This case is relatively rare. The second and much more frequent case is when $\mathcal{H}$ itself is an outcome of a learning process. For example, this is what happens in a validation procedure, see Figure 3.4. In validation we are using a validation set in order to select the best hypothesis out of a finite number of candidates corresponding to different parameter values and/or different algorithms.

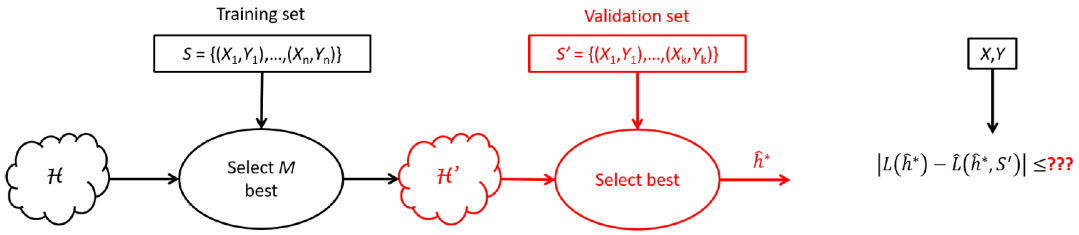

Figure 3.4: Validation (the red part in the figure) is identical to learning with a reduced hypothesis set $\mathcal{H}'$ (most often $\mathcal{H}'$ is finite).

NB. We reorganise a bit the above content, as shown in Table 3.2.

Table 3.2: Situations where we meet a finite $\mathcal{H}$ in real life

	the first case	***the second case***
when	the input space $\mathcal{X}$ is finite	$\mathcal{H}$ itself is an outcome of a learning process
frequency	relatively rare	much more frequent
e.g.,		this is what happens in a validation procedure. In validation we are using a validation set in order to select the best hypothesis out of a finite number of candidates corresponding to different parameter values and/or different algorithms.

And now comes the delicate point. Let $\hat{h}_S^*$ be a hypothesis with minimal empirical risk, $\hat{h}_S^* = \operatorname{argmin}_h \hat{L}(h, S)$ (it is natural to pick the empirical risk minimiser $\hat{h}_S^*$ to make predictions on new samples, but the following discussion equally applies to any other selection rule that takes sample S into account; note that there may be multiple hypotheses that achieve the minimal empirical error and in this case we can pick one arbitrarily). While for each h individually $\mathbf{E}[\hat{L}(h, S)] = L(h)$, this is not true for $\mathbf{E}[\hat{L}(\hat{h}_S^*, S)]$. In other words, $\mathbf{E}[\hat{L}(\hat{h}_S^*, S)] \neq \mathbf{E}[L(\hat{h}_S^*)]$ (we have to put expectation on the right hand side, because $\hat{h}_S^*$ depends on the sample). The reason is that when we pick $\hat{h}_S^*$ that minimises the empirical error on S, from the perspective of $\hat{h}_S^*$ the samples in S no longer look identical to future samples (X, Y). This is because $\hat{h}_S^*$ is selected in a very special way — it is selected to minimise the empirical error on S and, thus, it is tailored to S and most likely does better on S than on new random samples (X, Y). One way to handle this issue is to apply a union bound.

NB. **Question:** When you say "While for each h individually $\mathbf{E}[\hat{L}(h, S)] = L(h)$, this is not true for $\mathbf{E}[\hat{L}(\hat{h}_S^*, S)]$." Why is that? Why does the first one hold? It means that $\mathbf{E}[\hat{L}(h, S)] = L(h)$, right?

Theorem 3.2. *Assume that ℓ is bounded in the $[0, 1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0, 1)$, we have,*

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{M}{\delta}}\right) \leqslant \delta\,. \tag{3.20}$$

Proof.

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{M}{\delta}}\right) \leqslant \sum_{h \in \mathcal{H}} \mathbf{Pr}\left(L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{M}{\delta}}\right) \leqslant \sum_{h \in \mathcal{H}} \tfrac{\delta}{M} = \delta\,, \tag{3.21}$$

where the first inequality is by the union bound and the second is by Hoeffding's inequality.

NB. We have known that for one single hypothesis h, and for any $\delta \in (0, 1)$, it holds:

$$\mathbf{Pr}\left(\hat{L}(h, S) - L(h) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2)\,, \tag{3.22a}$$

$$\mathbf{Pr}\left(\hat{L}(h, S) - L(h) \leqslant -\varepsilon\right) = \mathbf{Pr}\left(L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2)\,, \tag{3.22b}$$

$$\mathbf{Pr}\left(L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{1}{\delta}}\right) \leqslant \delta\,. \tag{3.22c}$$

Note that ℓ is assumed to be bounded in the $[0, 1]$ interval as well, which means that $\ell(Y', Y) \in [0, 1]$ for all pairs of (Y', Y). Then for a finite hypotheses set $\mathcal{H}$ such that $|\mathcal{H}| = M$, we may have that:

$$\mathbf{Pr}\left(L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2)\,, \quad \forall h \in \mathcal{H}\,, \tag{3.23a}$$

$$\sum_{h \in \mathcal{H}} \mathbf{Pr}\left(L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) \leqslant \sum_{h \in \mathcal{H}} \exp(-2n\varepsilon^2) = |\mathcal{H}| \exp(-2n\varepsilon^2) = M \exp(-2n\varepsilon^2)\,. \tag{3.23b}$$

Let $\delta \stackrel{\text{def}}{=} M e^{-2n\varepsilon^2}$, then we get $n\varepsilon^2 = -\frac{1}{2} \ln \frac{\delta}{M} = \frac{1}{2} \ln \frac{M}{\delta}$, and then

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) \leqslant \sum_{h \in \mathcal{H}} \mathbf{Pr}\left(L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) = M \exp(-2n\varepsilon^2)\,, \tag{3.24a}$$

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : L(h) \geqslant \hat{L}(h, S) + \varepsilon\right) \leqslant \sum_{h \in \mathcal{H}} \mathbf{Pr}\left(L(h) \geqslant \hat{L}(h, S) + \varepsilon\right) = M \exp(-2n\varepsilon^2)\,, \tag{3.24b}$$

$$\mathbf{Pr}\left(\exists h \in \mathcal{H} : L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{M}{\delta}}\right) \leqslant \sum_{h \in \mathcal{H}} \mathbf{Pr}\left(L(h) \geqslant \hat{L}(h, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{M}{\delta}}\right) \stackrel{\text{def}}{=} \delta\,. \tag{3.24c}$$

An alternative way would be like: Let $\delta/M \overset{\text{def}}{=} e^{-2n\varepsilon^2}$ for $\mathbf{Pr}(\hat{L}(h,S) - L(h) \geqslant \varepsilon) \leqslant \exp(-2n\varepsilon^2)$. Then we have $n\varepsilon^2 = -\frac{1}{2}\ln\frac{\delta}{M} = \frac{1}{2}\ln\frac{M}{\delta}$, and then the other derivation thing left would be the same as (3.21). ■

Another way of reading Theorem 3.2 is: with probability at least $(1-\delta)$ for all $h \in \mathcal{H}$,

$$L(h) \leqslant \hat{L}(h,S) + \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\,. \tag{3.25}$$

It means that no matter which h from $\mathcal{H}$ is returned by the algorithm, with high probability we have the guarantee (3.25). In particular, it holds for $\hat{h}^*_S$. Again, remember that the random quantity is actually $\hat{L}(h,S)$ and the right way to read the bound is $\hat{L}(h,S) \geqslant L(h) - \sqrt{\ln(M/\delta)/(2n)}$, see the discussion in the previous section.

The price for considering M hypotheses instead of a single one is $\ln M$. Note that it grows only logarithmically with M. Also note that there is no contradiction between the upper bound and the lower bound we have discussed in Section 3.1. In the construction of the lower bound we took $M = |\mathcal{H}| = 2^{2n}$. If we substitute this value of M into (3.25) we obtain $\sqrt{\ln(M/\delta)/(2n)} \geqslant \sqrt{\ln(2)} \geqslant 0.8$, which has no contradiction with $L(h) \geqslant 0.25$.

NB. In Section 3.1,[29] the expected loss of $\hat{h}^*_S$ is

$$L(\hat{h}^*_S) \geqslant \frac{1}{2}\times 0 + \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}\,, \tag{3.26}$$

where the first term is an upper bound on the probability of observing an already seen student 0 times *the expected error that $\hat{h}^*_S$ makes in this case*, and the second term is a lower bound on the probability of observing a new student $1/2$ times *the expected error that $\hat{h}^*_S$ makes in this case*. Note that the expected error (that) $\hat{h}^*_S$ makes is $1/2$.

Because $M = |\mathcal{H}| = 2^{2n}$, then we get that

$$\sqrt{\frac{1}{2n}\ln\frac{M}{\delta}} = \sqrt{\frac{1}{2n}\ln\frac{2^{2n}}{\delta}} = \sqrt{\frac{1}{2n}\left(\ln\left(2^{2n}\right) - \ln(\delta)\right)} = \sqrt{\frac{1}{2n}\left(2n\ln(2) - \ln(\delta)\right)} = \sqrt{\ln(2) - \frac{\ln(\delta)}{2n}} \tag{3.27a}$$

$$\geqslant \sqrt{\ln(2)} \approx 0.8325546111576977 \geqslant 0.8 \geqslant 0.25\,, \tag{3.27b}$$

because $\delta \in (0,1)$ and $n \geqslant 1$. (Note that: $\ln(\delta) < 0$ and $-\ln(\delta)/(2n) \leqslant -\ln(\delta)/2$.)

Question: There are some problems with the probability. Note that we are talking about Theorem 3.2, that is to say, for any $\delta \in (0,1)$ we have: $\exists\, h \in \mathcal{H}$ such that

$$L(h) \geqslant \hat{L}(h,S) + \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\,, \tag{3.28}$$

[29] **Informal Lower Bound** Imagine that we want to learn a classifier that predicts whether a student's birthday is on an even or odd day based on student's id. Assume that the total number of students is $2n$, that the hypothesis class $\mathcal{H}$ includes all possible mappings from students id to even/odd, so that $\mathcal{H} = 2^{2n}$, and that we observe a sample of n uniformly sampled students (potentially with repetitions). Since all possible mappings are within $\mathcal{H}$, we have $\hat{h}^*_S \in \mathcal{H}$ for which $\hat{L}(\hat{h}^*_S, S) = 0$. However, $\hat{h}^*_S$ is guaranteed to make zero error only on the samples that were observed, which constitute at most half of the total number of students. For the remaining students $\hat{h}^*_S$ can, at the best, make a random guess which will succeed with probability $1/2$. Therefore, the expected loss of $\hat{h}^*_S$ is $L(\hat{h}^*_S) \geqslant \frac{1}{2}\times 0 + \frac{1}{2}\times\frac{1}{2} = \frac{1}{4}$, where the first term is an upper bound on the probability of observing an already seen student times the expected error $\hat{h}^*_S$ makes in this case and the second term is a lower bound on the probability of observing a new student times the expected error $\hat{h}^*_S$ makes in this case. For a more formal treatment see the lower bounds in Chapter 3.7.

(see the difference between it and (3.25)); therefore, the sign of the inequality is correct.

Similar to Theorem 3.1 it is possible to derive a two-sided bound on the error. It is also possible to derive a lower bound by using the other side of Hoeffding's inequality (2.15): with probability at least $(1-\delta)$, for all $h \in \mathcal{H}$ we have $L(h) \geqslant \hat{L}(h,S) - \sqrt{\ln(M/\delta)/(2n)}$. Typically we want the upper bound for one and a lower bound for the other. The "lazy" approach is to take the two-sided bound for everything, but sometimes it is possible to save the factor of $\ln(2)$ by carefully considering which hypotheses require the lower bound and which require the upper bound and applying the union bound correspondingly (we are not getting into the details).

NB. In Section 3.2,[30] we can get the upper bound (3.25) by (3.29b), with probability at least $(1-\delta)$. With the same proability, we may get the lower bound by (3.29a), that is,

$$L(h) \geqslant \hat{L}(h,S) - \sqrt{\frac{1}{2n}\ln\frac{M}{\delta}}\,, \tag{3.32}$$

and the two-sided bound by (3.29c).

Question: A more detailed illustration would be? (Solved, I guess.)

We have known that for one single hypothesis h, and for any $\varepsilon > 0$, it holds:

$$\mathbf{Pr}\left(\hat{L}(h,S) - L(h) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta\,,\ \text{where } \varepsilon = \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\,, \tag{3.33a}$$

$$\mathbf{Pr}\left(\hat{L}(h,S) - L(h) \leqslant -\varepsilon\right) = \mathbf{Pr}\left(L(h) - \hat{L}(h,S) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta\,,\ \text{where } \varepsilon = \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\,, \tag{3.33b}$$

$$\mathbf{Pr}\left(\left|\hat{L}(h,S) - L(h)\right| \geqslant \varepsilon\right) \leqslant 2\exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta\,,\ \text{where } \varepsilon = \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\,. \tag{3.33c}$$

Note that ℓ is assumed to be bounded in the $[0,1]$ interval as well, which means that $\ell(Y',Y) \in [0,1]$ for all pairs of (Y',Y). Then for a finite hypotheses set $\mathcal{H}$ such that $|\mathcal{H}| = M$, we may have that:

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : \hat{L}(h,S) - L(h) \geqslant \varepsilon\right) \leqslant \sum_{h\in\mathcal{H}} \mathbf{Pr}\left(\hat{L}(h,S) - L(h) \geqslant \varepsilon\right) = M\exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta \in (0,M)\,, \tag{3.34a}$$

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geqslant \varepsilon\right) \leqslant \sum_{h\in\mathcal{H}} \mathbf{Pr}\left(L(h) - \hat{L}(h,S) \geqslant \varepsilon\right) = M\exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta \in (0,M)\,, \tag{3.34b}$$

where $\varepsilon = \sqrt{1/(2n)\ln(M/\delta)}$ because $n\varepsilon^2 = -1/2\ln\delta/M$. Additionally,

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : \left|\hat{L}(h,S) - L(h)\right| \geqslant \varepsilon\right) \leqslant \sum_{h\in\mathcal{H}} \mathbf{Pr}\left(\left|\hat{L}(h,S) - L(h)\right| \geqslant \varepsilon\right) \leqslant 2M\exp(-2n\varepsilon^2) \stackrel{\text{def}}{=} \delta \in (0,2M)\,, \tag{3.35}$$

[30] According to Section 2.3, we know that:

$$\mathbf{Pr}\left(\hat{L}(h,S) - L(h) \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta \stackrel{\text{def}}{=} e^{-2n\varepsilon^2}\,, \tag{3.29a}$$

$$\mathbf{Pr}\left(\hat{L}(h,S) - L(h) \leqslant -\sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) = \mathbf{Pr}\left(L(h) - \hat{L}(h,S) \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta \stackrel{\text{def}}{=} e^{-2n\varepsilon^2}\,, \tag{3.29b}$$

$$\mathbf{Pr}\left(\left|L(h) - \hat{L}(h,S)\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\right) \leqslant \delta \stackrel{\text{def}}{=} 2e^{-2n\varepsilon^2}\,. \tag{3.29c}$$

Theorem (3.1). *Assume that ℓ is bounded in the $[0,1]$ interval (i.e., $\ell(Y',Y) \in [0,1]$ for all Y',Y), then for a single h and any $\delta \in (0,1)$ we have:*

$$\mathbf{Pr}\left(L(h) \geqslant \hat{L}(h,S) + \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{3.30}$$

and

$$\mathbf{Pr}\left(\left|L(h) - \hat{L}(h,S)\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\right) \leqslant \delta\,. \tag{3.31}$$

where $\varepsilon = \sqrt{1/(2n)\ln(2M/\delta)}$ because $n\varepsilon^2 = -1/2 \ln \delta/(2M)$.

1) The lower bound

- Assume that ℓ is bounded in the $[0,1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0,1)$, we have,

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : L(h) \leqslant \hat{L}(h,S) - \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\right) \leqslant \delta\,. \tag{3.36}$$

- With probability at least $(1-\delta)$ for all $h \in \mathcal{H}$,

$$L(h) \geqslant \hat{L}(h,S) - \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\,. \tag{3.37}$$

2) The upper bound

- Assume that ℓ is bounded in the $[0,1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0,1)$, we have,

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : L(h) \geqslant \hat{L}(h,S) + \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\right) \leqslant \delta\,. \tag{3.38}$$

- With probability at least $(1-\delta)$ for all $h \in \mathcal{H}$,

$$L(h) \leqslant \hat{L}(h,S) + \sqrt{\tfrac{1}{2n}\ln\tfrac{M}{\delta}}\,. \tag{3.39}$$

3) The two-sided bound

- Assume that ℓ is bounded in the $[0,1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0,1)$, we have,

$$\mathbf{Pr}\left(\exists\, h \in \mathcal{H} : \left|L(h) - \hat{L}(h,S)\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2M}{\delta}}\right) \leqslant \delta\,. \tag{3.40}$$

- With probability at least $(1-\delta)$ for all $h \in \mathcal{H}$,

$$\left|L(h) - \hat{L}(h,S)\right| \leqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2M}{\delta}}\,. \tag{3.41}$$