Machine Learning Notes

This note is based on Machine Learning at DIKU

15 December 2021

Chapter 3. Generalisation Bounds for Classification

3.2 Generalisation Bound for a Single Hypothesis

We start with the simplest case, where $\mathcal{H}$ consists of a single prediction rule h. We are interested in the quality of h, measured by $L(h)$, but all we can measure is $\hat{L}(h, S)$. What can we say about $L(h)$ based on $\hat{L}(h, S)$? Note that the samples $(X_i, Y_i) \in S$ come from the same distribution as any future samples (X, Y) we will observe. Therefore, $\ell(h(X_i), Y_i)$ has the same distribution as $\ell(h(X), Y)$ for any future sample (X, Y). Let $Z_i = \ell(h(X_i), Y_i)$ be the loss of h on (X_i, Y_i). Then $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^n Z_i$ is an average of n i.i.d. random variables with $\mathbf{E}[Z_i] = \mathbf{E}[\ell(h(X), Y)] = L(h)$. The distance between $\hat{L}(h, S)$ and $L(h)$ can thus be bounded by application of Hoeffding's inequality.

NB. Note that the samples $(X_i, Y_i) \in S$ come from the same distribution as any future samples (X, Y) we will observe. Let $Z_i \overset{\text{def}}{=} \ell\left(h(X_i), Y_i\right)$, then we have

$$\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h(X_i), Y_i\right), \tag{3.6a}$$

$$L(h) = \mathbf{E}[Z_i] = \mathbf{E}\left[\ell\left(h(X), Y\right)\right] = \mathbf{E}\left[\ell\left(h(X_i), Y_i\right)\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell\left(h(X_i), Y_i\right)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[\ell\left(h(X_i), Y_i\right)\right]. \tag{3.6b}$$

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:*

$$\mathbf{Pr}\left(L(h) \geqslant \hat{L}(h, S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leqslant \delta, \tag{3.7}$$

and

$$\mathbf{Pr}\left(\left|L(h) - \hat{L}(h, S)\right| \geqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \leqslant \delta. \tag{3.8}$$

Proof. For (3.7) take $\varepsilon = \sqrt{\ln(\frac{1}{\delta})/(2n)}$ in (2.16) and rearrange the terms. Eq. (3.8) follows in a similar way from the two-sided Hoeffding's inequality. Note that in (3.7) we have $1/\delta$ and in (3.8) we have $2/\delta$.

NB. According to Section 2.3,[28] we know that

$$\mathbf{Pr}\left(\hat{L}(h, S) - L(h) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2), \tag{3.15a}$$

$$\mathbf{Pr}\left(\hat{L}(h, S) - L(h) \leqslant -\varepsilon\right) = \mathbf{Pr}\left(L(h) - \hat{L}(h, S) \geqslant \varepsilon\right) \leqslant \exp(-2n\varepsilon^2), \tag{3.15b}$$

$$\mathbf{Pr}\left(\left|\hat{L}(h, S) - L(h)\right| \geqslant \varepsilon\right) \leqslant 2\exp(-2n\varepsilon^2). \tag{3.15c}$$

[28] We relist the Hoeffding's inequalities here.

Let $\delta \stackrel{\text{def}}{=} e^{-2n\varepsilon^2} \in (0,1)$, then we have $n\varepsilon^2 = -\frac{1}{2}\ln\delta = \frac{1}{2}\ln\frac{1}{\delta}$, and then

$$\mathbf{Pr}\left(\hat{L}(h,S) - L(h) \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{3.16a}$$

$$\mathbf{Pr}\left(L(h) - \hat{L}(h,S) \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{1}{\delta}}\right) \leqslant \delta\,, \tag{3.16b}$$

therefore, we can get (3.7) according to (3.16b).

Let $\delta \stackrel{\text{def}}{=} 2e^{-2n\varepsilon^2} \in (0,2)$, then we have $n\varepsilon^2 = -\frac{1}{2}\ln\frac{\delta}{2} = \frac{1}{2}\ln\frac{2}{\delta}$, and then

$$\mathbf{Pr}\left(\left|L(h) - \hat{L}(h,S)\right| \geqslant \sqrt{\tfrac{1}{2n}\ln\tfrac{2}{\delta}}\right) \leqslant \delta\,, \tag{3.17}$$

that is, (3.8).

Question: There is a bit problem here though. Why the following content keep saying $\delta \in (0,1)$ for the two-sided inequality? Oh I was wrong, thay are talking about the one-sided inequality. ■

There is an alternative way to read Eq. (3.7): with probability at least $(1-\delta)$ we have

$$L(h) \leqslant \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,. \tag{3.18}$$

We remind the reader that the above inequality should actually be interpreted as

$$\hat{L}(h,S) \geqslant L(h) - \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,, \tag{3.19}$$

and it means that with probability at least $(1-\delta)$ the empirical loss $\hat{L}(h,S)$ does not underestimate the expected loss $L(h)$ by more than $\sqrt{\ln(1/\delta)/(2n)}$. However, it is customary to write

---

Theorem (2.3 Hoeffding's inequality). *Let $X_1, ..., X_n$ be independent real-valued random variables, such that for each $i \in \{1, ..., n\}$ there exist $a_i \leqslant b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(-2\varepsilon^2/\sum_{i=1}^n (b_i-a_i)^2\right)\,, \tag{3.9}$$

and

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(-2\varepsilon^2/\sum_{i=1}^n (b_i-a_i)^2\right)\,, \tag{3.10}$$

aka "one-sided Hoeffding's inequalities".

Remark (Corollary 2.4). *Under the assumption of Theorem 2.3, taking a union bound of the events in those two:*

$$\mathbf{Pr}\left(\left|\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(-2\varepsilon^2/\sum_{i=1}^n (b_i-a_i)^2\right)\,, \tag{3.11}$$

aka "two-sided Hoeffding's inequality".

Remark (Corollary 2.5). *Let $X_1, ..., X_n$ be independent random variables, such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,, \tag{3.12}$$

and

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geqslant \varepsilon\right) \leqslant \exp\left(-2n\varepsilon^2\right)\,. \tag{3.13}$$

Remark (Lemma 2.6, Hoeffding's lemma). *Let X be a random variable, such that $X \in [a,b]$. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbf{E}[e^{\lambda X}] \leqslant \exp\left(\lambda\mathbf{E}[X] + \lambda^2(b-a)^2/8\right)\,. \tag{3.14}$$

The function $f(\lambda) = \mathbf{E}[e^{\lambda X}]$ is known as the moment generating function *of X, since generally, $f^{(k)}(0) = \mathbf{E}[X^k]$.*

the inequality in the first form (as an upper bound on $L(h)$) and we follow the tradition (see the discussion at the end of Section 2.3.1).

Theorem 3.1 is analogous to the problem of estimating a bias of a coin based on coin flip outcomes. There is always a small probability that the flip outcomes will not be representative of the coin bias. For example, it may happen that we flip a fair coin $1,000$ times (without knowing that it is a fair coin!) and observe "all heads" or some other misleading outcome. And if this happens we are doomed — there is nothing we can do when the sample doesn't represent the reality faithfully. Fortunately for us, this happens with a small probability that decreases exponentially with the sample size n.

Whether we use the one-sided bound (3.7) or the two-sided bound (3.8) depends on the situation. In most cases we are interested in the upper bound on the expected performance of the prediction rule given by (3.7).