Machine Learning Notes

This note is based on Machine Learning at DIKU

December 5, 2021

Chapter 2. Concentration of Measure Inequalities

2.3.1 Understanding Hoeffding's Inequality

Hoeffding's inequality involves three interconnected terms: n, ε, and $\delta = 2e^{-2n\varepsilon^2}$, which is the bound on the probability that the event under $\mathbf{Pr}(\cdot)$ holds (for the purpose of the discussion we consider two-sided Hoeffding's inequality for random variables bounded in $[0,1]$). We can fix any two of the three terms n, ε, and δ and then the relation $\delta = e^{-2n\varepsilon^2}$ provides the value of the third. Thus, we have

$$\delta = 2\exp\left(-2n\varepsilon^2\right), \tag{2.53a}$$

$$\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}, \tag{2.53b}$$

$$n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}. \tag{2.53c}$$

Overall, Hoeffding's inequality tells by how much the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$ can deviate from its expectation μ, but the interplay between the three parameters provides several ways of seeing and using Hoeffding's inequality. For example, if the number of samples n is fixed (we have made a fixed number of experiments and now analyse what we can get from them), there is an interplay between the precision ε and confidence δ. We can request higher precision ε, but then we have to compromise on the confidence δ that the desired bound $\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leqslant \varepsilon$ holds. And the other way around: we can request higher confidence δ, but then we have to compromise on precision ε, i.e., we have to increase the allowed range $\pm\varepsilon$ around μ, where we expect to find the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$.

NB. According to Theorem 2.3–Lemma 2.6 [1], we have: let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, then $-2n\varepsilon^2 = \ln(\delta/2)$ and $2n\varepsilon^2 = -\ln(\delta/2) = \ln(\delta/2)^{-1} = \ln(2/\delta)$. Thus $\varepsilon = \left(\frac{1}{2n}\ln\frac{2}{\delta}\right)^{1/2}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$. As ε increases,

[1] We re-list them here.

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

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{2.54}$$

and

$$\mathbf{Pr}\left(\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.55}$$

Remark (Corollary 2.4). *Under the assumptions of Theorem 2.3: "two-sided Hoeffding's inequality" holds, that is,*

$$\mathbf{Pr}\left(\left|\sum_{i=1}^n X_i - \mathbf{E}\left[\sum_{i=1}^n X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.56}$$

confidence δ will decrease; as δ increases, precision ε will decrease.

As another example, we may have target precision ε and confidence δ and then the inequality provides us the number of experiments n that we have to perform in order to achieve the target. It is often convenient to write the inequalities (2.57) and (2.58) with a fixed confidence in mind, thus we have

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leqslant \delta\,, \tag{2.60a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leqslant \delta\,, \tag{2.60b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \leqslant \delta\,. \tag{2.60c}$$

(Put attention that the $\ln 2$ factor in the last inequality comes from the union bound over the first two inequalities: if we want to keep the same confidence we have to compromise on precision.)

NB. Let $\delta \overset{\text{def}}{=} 2e^{-2n\varepsilon^2}$, then $\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$. If we use them as substitutes in Eqs. (2.57–2.58), we could obtain that

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \frac{1}{2}\delta\,, \tag{2.61a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \frac{1}{2}\delta\,, \tag{2.61b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}\right) \leqslant \delta\,. \tag{2.61c}$$

If let $\delta \overset{\text{def}}{=} e^{-2n\varepsilon^2}$, then $n\varepsilon^2 = -\frac{1}{2}\ln\delta = \ln(\frac{1}{\delta})^{1/2} = \ln\sqrt{\frac{1}{\delta}}$. Thus $\varepsilon = \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$ and $n = \frac{1}{2\varepsilon^2}\ln\frac{1}{\delta}$. Then we could obtain that

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.62a}$$

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant \delta\,, \tag{2.62b}$$

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geqslant \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\right) \leqslant 2\delta\,. \tag{2.62c}$$

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

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} = \exp\left(-2n\varepsilon^2\right)\,, \tag{2.57}$$

and

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \leqslant -\varepsilon\right) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2} = \exp\left(-2n\varepsilon^2\right)\,. \tag{2.58}$$

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

$$\mathbf{E}\left[e^{\lambda X}\right] \leqslant e^{\lambda\mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}} = e^{\lambda\mathbf{E}[X]+\lambda^2(b-a)^2/8} = \exp\left(\lambda\mathbf{E}[X] + \lambda^2(b-a)^2/8\right)\,. \tag{2.59}$$

The function $f(\lambda) = \mathbf{E}\left[e^{\lambda X}\right]$ is known as the *moment generating function* of X, since $f'(0) = \mathbf{E}[X]$, $f''(0) = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(0) = \mathbf{E}[X^k]$. *NB.* Since $f(\lambda) = \mathbf{E}\left[e^{\lambda X}\right]$ and $e^{\lambda X}|_{\lambda=0} = e^0 = 1$, then $f'(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X\right]\big|_{\lambda=0} = \mathbf{E}[X]$, and $f''(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X^2\right]\big|_{\lambda=0} = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(\lambda)|_{\lambda=0} = \mathbf{E}\left[e^{\lambda X}X^k\right]\big|_{\lambda=0} = \mathbf{E}\left[X^k\right]$.

Then the complimentary events of the inequality (2.62b), that is, (2.60b), will become the inequality (2.63).

In many situations we are interested in the complimentary events. Thus, for example, we have

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geqslant (1-\delta)\,. \tag{2.63}$$

Careful reader may point out that the inequalities above should be strict ("<" and ">"). This is true, but if it holds for strict inequalities it also holds for non-strict inequalities ("⩽" and "⩾"). Since strict inequalities provide no practical advantage we will use the non-strict inequalities to avoid the headache of remembering which inequalities should be strict and which should not.

The last inequality essentially says that with probability at least $(1-\delta)$, we have

$$\mu \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,, \tag{2.64}$$

and this is how we will occasionally use it. Note that the random variable is $\frac{1}{n}\sum_{i=1}^{n} X_i$ and the right way of interpreting the above inequality is actually that with probability at least $(1-\delta)$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \mu - \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\,, \tag{2.65}$$

i.e., the probability is over $\frac{1}{n}\sum_{i=1}^{n} X_i$ and not over μ. However, many generalisation bounds that we study in Chapter 3 are written in the first form in the literature and we follow the tradition.

NB. According to the equality (2.63) we will get the inequality (2.64) directly, with the equivalent (2.65). The complimentary events of the inequality (2.61c/2.60c) will become the inequality

$$\mathbf{Pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \leqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) \geqslant (1-\delta)\,, \tag{2.66}$$

then with probability at least $(1-\delta)$, we have

$$-\sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,, \tag{2.67a}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \mu \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,, \tag{2.67b}$$

$$\mu - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \leqslant \frac{1}{n}\sum_{i=1}^{n} X_i \leqslant \mu + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\,. \tag{2.67c}$$