Hoeffding Inequality

IASD.

Oct 12, 2023.

Foundations of Machine Learning. Hoeffding Inequality

Oct 12. 2023
Luca RALAIVOLA l.ralaivola@criteo.com
CRITEO AI LAB.
↳ Created in 2018.

I. What do we want to achieve?

Uniform generalization bounds, i.e. bounds like:

With high probability, $\forall f \in \mathcal{F}$

$$\text{generalization}(f) \leq \text{empirical error}(f, S) + \mathcal{E}(n, \mathcal{F}, \dots)$$

"Capacity" of $\mathcal{F}$

$\mathcal{E}$ supposed to decrease when $n \to +\infty$.

Typical setting of Statistical learning theory.

- $(X, Y) \sim \mathcal{D}$, $\mathcal{D}$ fixed and unknown distribution
- $S = \{(X_i, Y_i)\}_{i=1}^n$ where (X_i, Y_i) are independent copies of (X, Y)

In other words, the training sample is an IID sample where IID stands for Identically and Independently Distributed.

- Ultimate goal / criterion.

We want to minimize

$$R_\ell(f) := \mathbb{E}_{xy \sim D}\left[\ell(f(x), y)\right]$$

R_ℓ: ℓ-risk; f: learned predictor; ℓ: loss function

Examples: binary classification

- $\mathcal{F} \subseteq \{-1,+1\}^{\mathcal{X}}$
- $\ell(f(x), y) := \mathbb{1}_{f(x) \neq y}$

or

- $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$
- $\ell(f(x), y) := \mathbb{1}_{yf(x) \leq 0}$

- As we do not know D, we use S to "gather" information. The bound we look for is then:

With prob $1-\delta$

$$\forall f \in \mathcal{F} \quad R_\ell(f) \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)}_{\text{Empirical } \ell\text{-risk, the traditional quantity you look at when training.}} + \varepsilon(n, \delta, \mathcal{F})$$

- What appears here is the connection between an empirical quantity and its expectation. Concentrat°

II. Hoeffding Inequality 1963.

Thm: $X_1 - X_n$ independent r.v

Such that $\forall i \in \{1-n\}\ \exists a_i, b_i$: $P(a_i \le X_i \le b_i) = 1$.

Let $S_n := \sum_{i=1}^{n} X_i$ then

$$P(S_n - ES_n \ge \varepsilon) \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

$$P(ES_n - S_n \ge \varepsilon) \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2}\right)$$

Proposition. $X_1 - X_n$ IID and $\forall i\ P(0 \le X_i \le 1) = 1$

If $\mu := EX_1$ $(= EX_2 = \dots = EX_n)$

We have. $P\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \ge \varepsilon\right) \le \exp(-2n\varepsilon^2)$

$$P\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \ge \varepsilon\right) \le \exp(-2n\varepsilon^2)$$

Ex: Biased coin toss. How to estimate the bias with prob $1-\delta$.

$X_i := 1$ if tail, 0 if head

(2) $P\left(\left|\frac{1}{n}\sum X_i - \mu\right| \ge \varepsilon\right) \le 2\exp(-2n\varepsilon^2)$

μ: bias

② To get our estimate on μ, it suffices that

$$2\exp(-2n\varepsilon^2) \le \delta \iff \exp(-2n\varepsilon^2) \le \frac{\delta}{2}$$

$$\iff -2n\varepsilon^2 \le \log\frac{\delta}{2}$$

$$\iff 2n\varepsilon^2 \ge \log\frac{2}{\delta}$$

$$\iff \varepsilon^2 \ge \frac{1}{2n}\log\frac{2}{\delta}$$

$$\iff \varepsilon \ge \sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \quad (\varepsilon \ge 0)$$

with proba $1-\delta$ $\mu \in \left[\frac{1}{n}\sum x_i \pm \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}\right]$

Proof of the prop: do it, it's easy.

Remark: there is another concentration inequality, McDiarmid's inequality that is often used in SLT. It generalizes Hoeffding's inequality to the case where the function we look at is more complex than the mere average.

⇒ It provides an inequality in the form

$$P\left(\Phi(x_1,\dots,x_n) - E\Phi(x_1,\dots,x_n) \ge \varepsilon\right) \le \exp\left(-\frac{2\varepsilon^2}{\sum c_i^2}\right)$$

where $\forall i$

$$\sup_{x_i, x_i'} \left|\Phi(x_1,\dots,x_i,\dots,x_n) - \Phi(x_1,\dots,x_i',\dots,x_n)\right| \le c_i$$

Proof of Hoeffding's inequality:

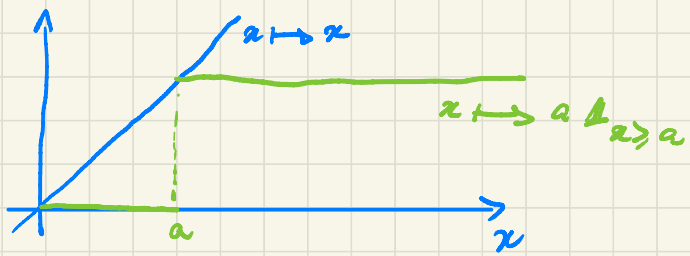
- Markov inequality
- Laplace transform / exp transform
- Hoeffding lemma
- Close out (easy).

Lemma: Markov Inequality

Let X be a rv taking nonnegative values ($\mathbb{P}(X \geq 0) = 1$) such that $\mathbb{E}X < +\infty$

$$\forall a > 0 \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}$$

Proof:



$\forall x \quad x \geq a \mathbb{1}_{x \geq a}$

$$\Rightarrow \quad \mathbb{E}X \geq \mathbb{E}[a \mathbb{1}_{X \geq a}] = a\, \mathbb{E}[\mathbb{1}_{X \geq a}] = a\, \mathbb{P}(X \geq a)$$ ∎

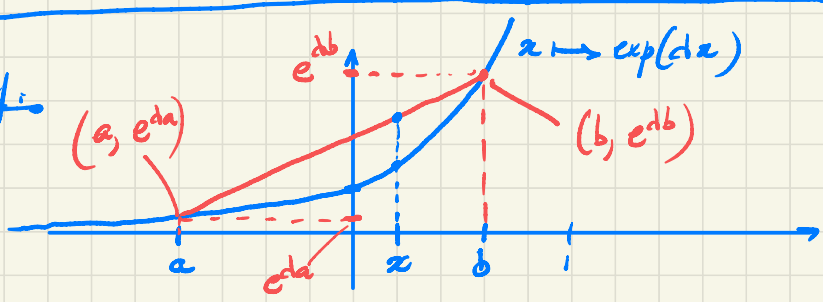
Hoeffding lemma

X rv / $\mathbb{E}X = 0$, $\exists a, b$, $\mathbb{P}(a \leq X \leq b) = 1$

($a \leq 0$, $b \geq 0$)

$$\forall d \in \mathbb{R}^* \quad \mathbb{E}e^{dX} \leq \exp\left(\frac{d^2 (b-a)^2}{8}\right)$$

Proof:



$$\forall x \in [a, b] \quad e^{dx} \leq \frac{b-x}{b-a} e^{da} + \frac{x-a}{b-a} e^{db}$$

$$\Rightarrow \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}\left[\frac{b-X}{b-a} e^{\lambda a} + \frac{X-a}{b-a} e^{\lambda b}\right]$$

$$\Leftrightarrow \mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}$$

$$\Leftrightarrow \mathbb{E}[e^{\lambda x}] \leq e^{\lambda(b-a)\frac{a}{b-a}}\left(1 + \frac{a}{b-a} - \frac{a}{b-a} e^{\lambda(b-a)}\right)$$

$$= e^{hp}(1+p-pe^{h}) \quad \text{where} \quad \begin{cases} h := \lambda(b-a) \\ p := \frac{a}{b-a} \end{cases}$$

$$= e^{L(h)} \quad \text{where } L(h) := -hp + \log(1-p+pe^{h})$$

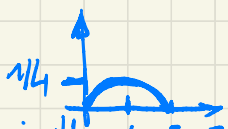
Taylor expansion: $\exists v$ / $L(h) = L(0) + hL'(0) + \frac{1}{2} h^2 L''(v)$

with $L(0) = 0$

$L'(0) = 0$

$L''(v) = \dots = t(1-t)$ with $t \in [0,1]$

$\leq \frac{1}{4}$.



Hence: $\forall h, \; L(h) = 0 + 0 \times h + \frac{1}{2} h^2 \underbrace{L''(v)}_{\leq \frac{1}{4}}$

$$\leq \frac{1}{8} h^2$$

$$= \frac{1}{8} \lambda^2 (b-a)^2$$

■

Lemma, $X_1 - X_n$ independent r.v

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E} X_i$$

• Proof of Hoeffding inequality.

$$\forall \lambda \quad P(S_n - \mathbb{E}[S_n] \geq \varepsilon) = P\left(\exp\left(\lambda(S_n - \mathbb{E}S_n)\right) \geq \exp(\lambda\varepsilon)\right)$$

$$= P\left(\exp\left(\lambda\left(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\right)\right) \geq \exp(\lambda\varepsilon)\right)$$

$$= P\left(\exp\left[\lambda \sum_{i=1}^n (X_i - \mathbb{E}X_i)\right] \geq \exp(\lambda\varepsilon)\right)$$

$$\left| \begin{array}{l} Z_i := X_i - \mathbb{E}X_i \\ \mu_i := \mathbb{E}X_i \end{array} \right. \qquad \left. P(Z_i \in [\underbrace{a_i - \mu_i}_{<0}, \underbrace{b_i - \mu_i}_{>0}]) = 1 \right\}$$

$$= P\left(\exp\left(\lambda \sum_{i=1}^n Z_i\right) \geq \exp(\lambda\varepsilon)\right)$$

$$\leq \mathbb{E}\left[\exp\left[\lambda \sum_{i=1}^n Z_i\right]\right] / \exp(\lambda\varepsilon) \quad \text{(Markov)}$$

$$= \mathbb{E}\left[\underbrace{\prod_{i=1}^n \exp(\lambda Z_i)}_{\text{independent}}\right] \times \exp(-\lambda\varepsilon)$$

$$= \prod_{i=1}^n \mathbb{E}\exp(\lambda Z_i) \cdot \exp(-\lambda\varepsilon)$$

$$\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2}{8}\left((b_i - \mu_i) - (a_i - \mu_i)\right)^2\right) \exp(-\lambda\varepsilon)$$

$$= \exp\left(\frac{\lambda^2}{8}\sum (b_i - a_i)^2 - \lambda\varepsilon\right)$$

minimizing wrt λ:

$$g(\lambda) = \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2 - \lambda\varepsilon$$

$$g'(\lambda) = \frac{1}{4}\lambda \sum_{i=1}^n (b_i - a_i)^2 - \varepsilon$$

$$g'(\lambda) = 0 \iff \lambda = \frac{4\varepsilon}{\sum (b_i - a_i)^2}$$

$$= \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

In the case of McDiarmid's Inequality we use:

$$\mathbb{E}[\phi(X_1 - X_n)] = \mathbb{E}_{X_1 - X_{n-1}} \mathbb{E}_{X_n}[\phi(\,) | X_1 - X_{n-1}]$$

or, equivalently:

$$Z_i := \mathbb{E}[\phi(\,) | X_1 - X_i] - \mathbb{E}[\phi(\cdot) | X_1 \cdots X_{i-1}]$$

such that

$$\sum_{i=1}^n Z_i = \underbrace{\mathbb{E}[\phi(\,) | X_1 - X_n]}_{\phi(X_1 - X_n)} - \mathbb{E}\phi(X_1 \cdots X_n)$$