Fondamentaux de l'Apprentissage Automatique
Hoeffding Inequality

Lecturer: Liva Ralaivola
Scribe: Lucas Fourest

Lecture n°3 #
18/10/2023

---

1 What do we want to achieve ?

Intuitively : uniform generalization bounds for a family $\mathcal{F}$: with high probability,

$$\forall f \in \mathcal{F}, R_{\text{general}}(f) \leq \text{error}_{\text{empirical}}(f, \mathcal{S}) + \epsilon(n, \mathcal{F}, ...)$$

where $\mathcal{S}$ denotes a dataset of size n.

1.1 Typical setting of ML theory

1. $(X, Y)\tilde{}\mathcal{D}$ where $\mathcal{D}$ is an unknown distribution
2. $\{(x_i, y_i)\}_{i\in[1,n]}$ are i.i.d drawn from $(\mathcal{X}, \mathcal{Y})$

1.2 Ultimate criterion

We want to minimize $R_l(f) = \mathbb{E}_{\mathcal{X},\mathcal{Y}\tilde{}\mathcal{D}}[l(f(X), Y)]$

As we don't know $\mathcal{D}$, we use δ to gather information with samples. The bound we look for is then : with probability $1 - \delta$,

$$R_l(f) \leq \frac{1}{n}\sum_{i=1}^{N} l(f(x_i), y_i) + \epsilon(n, \delta, \mathcal{F})$$

What appears here is the connection between an empirical entity and its expectation concentration.

2 Hoeffding inequality

Theorem 1. *$X_1, ..., X_n$ are n independant R.V, such that $\forall i \in \{1, ..., n\}, \exists a_i, b_i, P(a_i \leq X_i \leq b_i) = 1$. Let $S_n = \sum_{i=1}^n X_i$. Then, $\forall \epsilon > 0$*

$$\begin{cases} P(S_n - E(S_n) \geq \epsilon) \leq exp(\frac{-2\epsilon^2}{\sum_{i=0}^n (a_i - b_i)^2}) \\ P(E(S_n) - S_n \geq \epsilon) \leq exp(\frac{-2\epsilon^2}{\sum_{i=0}^n (a_i - b_i)^2}) \end{cases} \tag{1}$$

From 1 we can deduce that (if we apply the theorem with $\epsilon' = n\epsilon > 0$?) :
If $X_1, ..., X_n$ are n i.i.d R.V and $\forall i \in \{1, ..., n\}, P(0 \leq X_i \leq 1) = 1$, with $\mu_i = \mathbb{E}[X_i] = \mathbb{E}[X_1] = \mu$, then

$$P(\frac{1}{n}S_n - \mu \geq \epsilon) \leq exp(-2\epsilon^2 n)$$

$$P(\mu - \frac{1}{n}S_n \geq \epsilon) \leq exp(-2\epsilon^2 n)$$

hence

$$P(|\frac{1}{n}S_n - \mu \geq \epsilon|) \leq 2exp(-2\epsilon^2 n)$$

<u>Example :</u>

A biased coin for which we try to get an approximation of the expectation with confidence $1-\delta$, having $X = \begin{cases} 1 \text{ if tail} \\ 0 \text{ if head} \end{cases}$. If we denote X_i the outcome of the i^{th} try, from 1

$$P(|\frac{1}{n}S_n - \mu \geq \epsilon|) \leq 2exp(-2\epsilon^2 n)$$

Then to get our estimation of μ, it suffices that

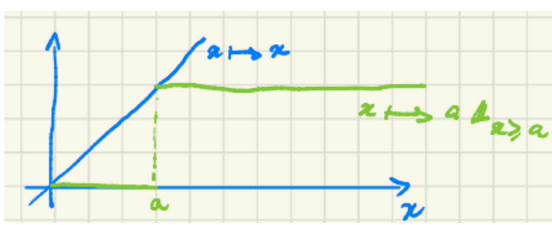
$$2exp(-2n\epsilon^2) \leq \delta \iff \epsilon \geq \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$$

Hence with a level of confidence/probability of $1-\delta, \mu \in [S_n \pm \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}]$

2.1 Proof of the Hoeffding inequality

Lemma 1. *Markov Inequality*
Let X be a RV taking non-negative values ($P(X \geq 0) = 1$) and such that $\mathbb{E}[X] < \infty$

$$\forall a > 0, P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{2}$$



Démonstration.

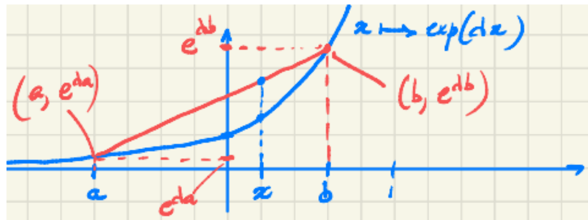
$$\forall x, x \geq a1_{x\geq a} \Rightarrow \mathbb{E}[X] \geq \mathbb{E}[a1_{X\geq a}] = a\mathbb{E}[1_{X\geq a}] = aP(X \geq a)$$

□

Lemma 2. *Hoeffding lemma*

If X a R.V, $\mathbb{E}[X] = 0$ and $\exists a < 0$ and $b > 0, P(a \leq X \leq b) = 1$

$$\forall \lambda \in \mathbb{R}^*, \mathbb{E}[e^{\lambda X}] \leq exp(\frac{\lambda^2(b-a)^2}{8}) \tag{3}$$



Démonstration. We will use the convexity of $x \rightarrow exp(\lambda x)$

$$\forall x \in [a,b], exp(\lambda x) \leq \frac{b-x}{b-a} exp(\lambda a) + \frac{x-a}{b-a} exp(\lambda b)$$

$$\Rightarrow \mathbb{E}[exp(\lambda x)] \leq \mathbb{E}[\frac{b-x}{b-a} exp(\lambda a) + \frac{x-a}{b-a} exp(\lambda b)]$$

$$\iff \mathbb{E}[exp(\lambda x)] \leq exp(L(h))$$

where $L(h) = -ln(p) + ln(1 - p + pexp(h))$, $\begin{cases} p = \frac{a}{b-a} > 0 \\ h = \lambda(b-a) \end{cases}$

Taylor expansion :

$$\exists v, L(h) = L(0) + hL'h(0) + \frac{1}{2}h^2 L''(v)$$

with $L(0) = L'(0) = 0$ and by derivation

$$L''(v) = \frac{pexp(v)(1 - p + pexp(v)) - (pexp(v))^2}{(1 - p + pexp(v))^2} = \frac{(1-p)pexp(v)}{(1 - p + pexp(v))^2}$$

hence

$$L''(v) = t(1-t) \leq \frac{1}{4} \text{ where } t = \frac{1-p}{1 - p + pexp(v)}$$

then

$$L(h) = 0 + 0 + \frac{1}{2}h^2 L''(v) \leq \frac{1}{8}h^2 = \frac{1}{2}\lambda^2(b-a)^2 \tag{4}$$

Lemma 3. *$X_1, ..., X_n$ are n independant R.V*

$$\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i] \tag{5}$$

Full proof of Hoeffding inequality

Finally, by gathering all the previous elements, we demonstrate that $\forall \epsilon > 0$

$$\begin{aligned} \forall \lambda > 0, P(S_n - \mathbb{E}[S_n] \geq \epsilon) &= P(exp(\lambda(S_n - \mathbb{E}[S_n])) \geq exp(\lambda \epsilon)) \\ &= P(exp(\lambda \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])) \geq exp(\lambda \epsilon)) \end{aligned}$$

By growth of $x \rightarrow exp(\lambda x)$. If we note $\mu_i = \mathbb{E}[X_i]$ and $Z_i = X_i - \mu_i$, we have $P(Z_i \in [a_i - \mu_i,\ b_i - \mu_i]) = 1$ as $\forall i, a_i \leq X_i \leq b_i$. Then

$$\begin{aligned}
P(S_n - \mathbb{E}[S_n] \geq \epsilon) &= P(exp(\sum_{i=1}^{n} Z_i) \geq exp(\lambda\epsilon)) \\
&\leq \frac{\mathbb{E}[exp(\lambda \sum_{i=1}^{n} Z_i)]}{exp(\lambda\epsilon)} \text{ with 2} \\
&\leq \mathbb{E}[\prod_{i=1}^{n} exp(\lambda Z_i)]exp(-\lambda\epsilon) \\
&\leq (\prod_{i=1}^{n} \mathbb{E}[exp(\lambda Z_i)])exp(-\lambda\epsilon) \text{ with 5} \\
&\leq (\prod_{i=1}^{n} exp(\frac{\lambda^2}{8}((b_i - \mu_i) - (a_i - \mu_i))^2))exp(-\lambda\epsilon) \text{ with 3}
\end{aligned}$$

$$\text{Then finally } ,\forall \lambda > 0, P(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq exp(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\epsilon) \tag{6}$$

If we note $g(\lambda) = exp(\frac{\lambda^2}{8}S - \lambda\epsilon)$ with $S = \sum_{i=1}^{n}(b_i - a_i)^2 > 0$, we have that :

$$g'(\lambda) = (\frac{\lambda}{4}S - \epsilon)exp(\frac{\lambda^2}{8}S - \lambda\epsilon)$$

then

$$g'(\lambda) = 0 \iff \lambda = \lambda_m = \frac{4\epsilon}{S}$$

Thus g as a single extrema $g(\lambda_m) = exp(\frac{-2\epsilon^2}{S})$. Given 6 is true for all $\lambda > 0$, it is also for λ_m which gives :

$$P(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq g(\lambda_m) = exp(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(a_i - b_i)^2})$$

□