Measure concentration and statistics in high-dimension: an introduction
Exercise sheet

October 21st, 2021

Exercise: randomized algorithm boosting

Imagine we have an algorithm for solving some decision problem (e.g., is a given number p a prime?). Suppose the algorithm makes a decision at random and returns the correct answer with probability $1/2 + \delta$, for some $\delta > 0$, which is just a bit better than a random guess. To improve the performance, we run the algorithm N times and take the majority vote. Show that, for any $\epsilon \in (0, 1)$, the answer is correct with probability $> 1 - \epsilon$ as soon as

$$N > \frac{\log(1/\epsilon)}{2\delta^2} .$$

Exercise: multiplicative concentration for Bernoulli variables

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{B}(\mu)$. What do the classical inequalities give to bound

$$\mathbb{P}\left(\bar{X}_n \leq \frac{\mu}{2}\right) \quad \text{and} \quad \mathbb{P}\left(\bar{X}_n \leq 2\mu\right) \quad ?$$

Exercise: characterizing sub-Gaussianity

Let X be a centered variable. Show that the following assertions are equivalent:

1. There exists $\sigma^2 > 0$ such that $\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 \sigma^2/2}$,

2. There exists $c > 0$ such that $\forall t \geq 0, \mathbb{P}\big(|X| > t\big) \leq 2e^{-ct^2}$,

3. There exists $a > 0$ such that $\mathbb{E}\left[e^{aX^2}\right] \leq 2$.

Exercise: VC dimension

What is the VC-dimension of the class

$$\mathcal{H}_{\text{seg}} = \left\{\mathbb{R} \ni x \mapsto \mathbb{1}_{[a,b]}(x) : a \leq b\right\} \ ?$$

and

$$\mathcal{H}^2_{\text{rec}} = \left\{\mathbb{R}^2 \ni x \mapsto \mathbb{1}_{[a_1,b_1]}(x_1)\mathbb{1}_{[a_2,b_2]}(x_2) : a_1 \leq b_1 \text{ and } a_2 \leq b_2\right\} \ ?$$

What about

$$\mathcal{H}_{\text{conv}} = \left\{\mathbb{R}^d \ni x \mapsto \mathbb{1}_K : K \text{convex}\right\} \ ?$$

Problem: Median of Means

In this problem, we denote by $\mathcal{B}(n, p)$ the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ and by $\mathbb{1}$ the indicator function. We assume that k and m are integers, and that $n = m \times (2k-1)$. We assume that $X_1, \dots, X_n$ are i.i.d. random variables on $\mathbb{R}$ with expectation μ and finite variance σ^2, but we do not assume that X_1 has finite exponential moments.

Given a fixed risk δ (for example $\delta = 1\%$), we want to construct a confidence interval I_n for μ, that is a $\sigma(X_1, \dots, X_n)$-measurable interval $I_n = [L_n, U_n]$ such that $\mathbb{P}(\mu \in I_n) \geq 1 - \delta$.

1. What confidence interval can you propose using the deviation inequalities you already know? How does its width depend on δ?

2. If you know that there exists $s > 0$ such that $\mathbb{P}(-s \leq X_1 \leq s) = 1$, what better confidence interval can you propose? How does its width depend on δ?

3. Let ℓ be a positive integer, let $0 \leq p \leq q \leq 1$, let $Y \sim \mathcal{B}(\ell, p)$ and $Z \sim \mathcal{B}(\ell, q)$. Show that for every $x \geq 0$, $\mathbb{P}(Y \geq x) \leq \mathbb{P}(Z \geq x)$.

4. Let k be a positive integer and let $0 \leq p \leq 1/4$. Show that if $T \sim \mathcal{B}(2k-1, p)$,
$$P(T \geq k) \leq \left(\frac{3}{4}\right)^k .$$

For every $j \in \{1, \dots, 2k-1\}$, we define $M_j = \dfrac{X_{(j-1)m+1} + X_{(j-1)m+2} + \cdots + X_{jm}}{m}$.
Let $\left(M_{(j)}\right)_{1 \leq j \leq 2k-1}$ be an order statistics of $\left(M_{(j)}\right)_{1 \leq j \leq 2k-1}$, that is a $2k-1$-uple of random variables such that

$$\left\{M_{(j)} : 1 \leq j \leq 2k-1\right\} = \left\{M_j : 1 \leq j \leq 2k-1\right\} \quad \text{and} \quad M_{(1)} \leq M_{(2)} \leq \cdots \leq M_{(2k-1)} .$$

Finally, let $\hat{\mu}_{k,m} = M_{(k)}$.

5. Show that for every $j \in \{0, \dots, 2k-2\}$,
$$\mathbb{P}\left(\left|M_j - \mu\right| \geq \frac{2\sigma}{\sqrt{m}}\right) \leq \frac{1}{4} .$$

6. Show that
$$|\hat{\mu}_{k,m} - \mu| \geq \frac{2\sigma}{\sqrt{m}} \implies \sum_{j=1}^{2k-1} \mathbb{1}\left\{|M_j - \mu| \geq \frac{2\sigma}{\sqrt{m}}\right\} \geq k .$$

UNIVERSITÉ DE LYON | ENS DE LYON

7. Show that
$$\mathbb{P}\left(|\hat{\mu}_{k,m} - \mu| \geq \frac{2\sigma}{\sqrt{m}}\right) \leq \left(\frac{3}{4}\right)^k .$$

8. Show that for every $\delta \leq e^{-2}$ and every $n \geq 16\ln(1/\delta)$, one can find integers k and m such that $n \geq m \times (2k-1)$ and
$$\mathbb{P}\left(|\hat{\mu}_{k,m} - \mu| \geq 8\sigma\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right) \leq \delta .$$

9. Deduce from the last question a confidence interval I_n for μ. How does it compare with the one proposed in Question 1? and with the one proposed in Question 2?

10. Is it possible to improve the result obtained in Question 8?