M2A 2023-2024
M2 Statistique & Learning

Exercises 1: Generalization bounds

Exercise 1. *The goal of this exercise is to show the following result.*

Theorem 1 (Contraction principle, Ledoux & Talagrand 1991)**.** *If φ is B-Lipshitz and $\epsilon_1^n = \{\epsilon_i\}_{i=1}^n$ is a sequence of i.i.d. random variables* $\mathrm{Rad}(1/2)$*, then*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varphi\left(\theta^\top x_i\right)\right] \leq B\mathbb{E}\left[\sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\theta^\top x_i\right].$$

We are going to prove by induction that for $k \in \{1,\dots,n\}$, for any functions $b:\Theta\to\mathbb{R}$, $a_i:\Theta\to\mathbb{R}$, $i\in\{1,\dots,k\}$ and any 1-Lipschitz functions $\varphi_i:\mathbb{R}\to\mathbb{R}$, $i=1,\dots,k$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{k}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right] \leqslant \mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{k}\varepsilon_i a_i(\theta)\right] \tag{1}$$

1. *For any functions $\varphi,\psi:\Theta\to\mathbb{R}$ and $\epsilon\sim\mathrm{Rad}(1/2)$, show that*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\{\psi(\theta)+\epsilon_i\varphi(\theta)\}\right] = \frac{1}{2}\left[\sup_{\theta,\theta'\in\Theta^2}\{\psi(\theta)+\psi(\theta')+|\varphi(\theta)-\varphi(\theta')|\}\right].$$

2. *Assume that (1) is satisfied for some $k\in\mathbb{N}$. Show that*

$$\mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{k+1}\varepsilon_i\varphi_i\left(a_i(\theta)\right)\right] \leq \mathbb{E}\left[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{k+1}\varepsilon_i a_i(\theta)\right]$$

3. *Conclude.*

Exercise 2 (Bounded differences/McDiarmid's inequality)**.** *Let $f:\mathcal{Z}^n\to\mathbb{R}$ be a function of bounded variation, that is, a function such that for any $i\in\{1,\dots,n\}$, and any $z_1,\dots,z_n,z_i'\in\mathcal{Z}$, we have*

$$|f(z_1,\dots,z_{i-1},z_i,z_{i+1},\dots,z_n)-f(z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_n)|\leq c.$$

Let $Z_1,\dots,Z_n$ be independent (and not necessarily identically distributed) random variables on $\mathcal{Z}$. Then, the random variable $U=f(Z_1,\dots,Z_n)$ is sub-Gaussian with variance proxy $nc^2/4$.

Hint: *The idea is to decompose U into a telescoping sum of* conditionally *independent random variables:*

$$U-\mathbf{E}U=(U_1-U_0)+(U_2-U_1)+(U_n-U_{n-1}),$$

where $U_i=\mathbf{E}_{X_{i+1},\dots,X_n}[U|X_1,\dots,X_i]$ (interpreting U_0 as $\mathbf{E}[U]$). Conclude the proof by a repeated application of Hoeffding's lemma. You may begin the proof by writing $U-\mathbf{E}U=U_n-U_0=(U_n-U_{n-1})+(U_{n-1}-U_0)$ and conditioning on the values of $X_1,\dots,X_{n-1}$.

In the following exercise, we demonstrate how the bounded differences inequality can be applied to obtain generalization error guarantees (applicable for any algorithm) and excess risk guarantees (applicable for empirical risk minimizers) that *hold with high probability.*

Exercise 3 (High-probability generalization and excess risk guarantees)**.**

Let $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ be i.i.d. random variables. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function such that for any y, y' we have $|\ell(y, y')| \le \ell_\infty$. For any function f, let $\mathcal{R}(f) = \mathbf{E}\ell(Y, f(X))$ and $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i))$. Let $\mathcal{F}$ be an arbitrary class of predictors and let $\mathcal{L} = \{\ell_f : (x, y) \mapsto \ell(y, f(x)) : f \in \mathcal{F}\}$.

Using the bounded differences inequality proved in Exercise 2, show that for any $\delta \in (0, 1)$ the following deviation inequality holds:

$$\mathbf{P}\left(\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} \ge \mathbf{E}\sup_{f\in\mathcal{F}}\left\{\mathcal{R}(f) - \hat{\mathcal{R}}(f)\right\} + \ell_\infty\sqrt{\frac{2\log(1/\delta)}{n}}\right) \le \delta.$$

Deduce the following two inequalities:

1. *For any statistical estimator that selects a predictor $\hat{f} = \hat{f}(Z_1, \dots, Z_n)$ from the class $\mathcal{F}$ it holds, with probability at least $1 - \delta$, that*

$$\mathcal{R}(\hat{f}) \le \hat{\mathcal{R}}(\hat{f}) + 2\mathrm{Rad}_n(\mathcal{L}) + \ell_\infty\sqrt{2\frac{\log(1/\delta)}{n}}, \tag{2}$$

   where recall that

$$\mathrm{Rad}_n(\mathcal{L}) = \mathbf{E}_{Z_1,\dots,Z_n}\mathbf{E}_{\varepsilon_1,\dots,\varepsilon_n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i\ell(Y_i, f(X_i))|Z_1, \dots, Z_n\right].$$

2. *Let $\hat{f}^{(erm)} \in \mathrm{argmin}_{f\in\mathcal{F}}\hat{\mathcal{R}}(f)$ be any empirical risk minimizer among the functions in the class $\mathcal{F}$. Let $f^* \in \mathrm{argmin}_{f\in\mathcal{F}}\mathcal{R}(f)$ (for simplicity, we assume that such f^* exists). Prove that*

$$\mathcal{R}(\hat{f}^{(erm)}) \le \hat{\mathcal{R}}(f^*) + 2\mathrm{Rad}_n(\mathcal{L}) + 2\ell_\infty\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{3}$$

Exercise 4 (Rademacher complexity of a set of points)**.** *For a subset $S \subset \mathbb{R}^n$, we define the unnormalized Rademacher complexity as*

$$\mathrm{URad}(S) := \mathbf{E}_\varepsilon \sup_{u\in S} \varepsilon^\top u$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is a vector of independent Rademacher random variables (each taking value $+1$ or -1 with probability $1/2$).

1. *What is the link between the definition in class of* $\mathrm{Rad}(\mathcal{F})$ *for a hypothesis class $\mathcal{F}$ and* URad*?*
2. *Compute* $\mathrm{URad}(\{u\})$ *for an arbitrary $u \in \mathbb{R}^n$*
3. *Compute* $\mathrm{URad}(HC)$ *where $HC = \{-1, +1\}^n$ is the unit hypercube*
4. *Give an upper bound on* $\mathrm{URad}(\{\mathbf{1}, -\mathbf{1}\})$*, where $\mathbf{1} \in \mathbb{R}^n$ is a vector with all entries equal to 1.*

In the upper bounds (2) and (3) we pay for the Rademacher complexity of the "loss class" $\mathcal{L}$. Whenever the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is L-Lipschitz in its second argument, that is, whenever, for any $y, y_1, y_2 \in \mathcal{Y}$ it holds that

$$|\ell(y, y_1) - \ell(y, y_2)| \le L|y_1 - y_2|,$$

we can pay, up to factor L, for the complexity of the class of predictors $\mathcal{F}$.