Mehryar Mohri
Foundations of Machine Learning 2022
Courant Institute of Mathematical Sciences
Homework assignment 1
February 8, 2022
Due: February 22, 2022

A Concentration bound

1. We denote by $\mathcal{X}$ the input space and S an i.i.d sample of size m.

   (a) Show that there does not exist any hypothesis $h: \mathcal{X} \to \{0, 1\}$ such that the following inequality holds with probability at least $e^{-m/3}$:
$$R(h) - \widehat{R}_S(h) \geq \frac{1}{2}.$$

   (b) Suppose that the target concept to learn is $c \equiv 1$ and the target distribution D is the uniform distribution over the interval $[0, 1]$. Design an algorithm such that for any sample S, the returned hypothesis $h_S: \mathcal{X} \to \{0, 1\}$ satisfies the following equality:
$$R(h_S) - \widehat{R}_S(h_S) = 1.$$

   (c) Why does part (b) not contradict part (a)?

B PAC-Bayesian bound

1. Let $\mathcal{H}$ be a hypothesis set of functions mapping $\mathcal{X}$ to $\mathbb{R}$ and let ℓ be a loss function mapping $\mathbb{R} \times \mathcal{Y}$ to $[0, 1]$. Denote the loss of a hypothesis h at point $z = (x, y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ by $L(h, z) = \ell(h(x), y)$. Let P and Q be probability measures over $\mathcal{H}$. In the PAC-Bayes framework, P represents the *prior probability* over the hypothesis class, i.e., the probability that a particular hypothesis is selected by the learning algorithm. Q represents the *posterior probability* selected after observing the training sample. In this exercise, we will derive learning bounds for randomized algorithms, in terms of the relative entropy of Q and P, denoted by $\mathsf{D}(Q \,\|\, P)$ (See E.2 of the textbook for the definition).

   (a) Define $\mathcal{G}_\mu$ via $\mathcal{G}_\mu = \{Q \in \Delta(\mathcal{H}) : \mathsf{D}(Q \,\|\, P) \leq \mu\}$, where we denote by $\Delta(\mathcal{H})$ the family of distributions over $\mathcal{H}$. Use the Rademacher complexity bound to show that for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $Q \in \mathcal{G}_\mu$:
$$\mathop{\mathbb{E}}_{\substack{h \sim Q \\ z \sim \mathcal{D}}}[L(h, z)] \leq \mathop{\mathbb{E}}_{h \sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h, z_i)\right] + 2\mathfrak{R}_m(\mathcal{G}_\mu) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

   (b) It can be shown that the following inequality holds:
$$\mathfrak{R}_m(\mathcal{G}_\mu) \leq \sqrt{\frac{2\mu}{m}}.$$
   Use this information to show that for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $Q \in \Delta(\mathcal{H})$:
$$\mathop{\mathbb{E}}_{\substack{h \sim Q \\ z \sim \mathcal{D}}}[L(h, z)] \leq \mathop{\mathbb{E}}_{h \sim Q}\left[\frac{1}{m}\sum_{i=1}^{m} L(h, z_i)\right] + \left(4 + \frac{1}{\sqrt{e}}\right)\sqrt{\frac{\max\{\mathsf{D}(Q \,\|\, P), 1\}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$
   (*Hint*: use the doubling trick, i.e., for some $a > 0$, $\Delta(\mathcal{H})$ can be written as the union of $\{Q \in \Delta(\mathcal{H}) : \mathsf{D}(Q \,\|\, P) \leq a\}$ and $\bigcup_{j=1}^{\infty} \left\{Q \in \Delta(\mathcal{H}) : a2^{j-1} < \mathsf{D}(Q \,\|\, P) \leq a2^j\right\}$. Then, use the union bound to extend the result in part (a). Note that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\frac{\log(2t)}{2} \leq \frac{t}{e}$ for $t > 0$.)

C Rademacher complexity

1. Let $\mathcal{X} \subset \mathbb{R}^N$ and let $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ be a sample of size m. In this problem, we consider the following linear hypothesis set

$$\mathcal{H} = \{x \mapsto w \cdot x : \|w\|_1 \leq \Lambda\}.$$

We denote by X the matrix $X = [x_1, \ldots, x_m]$ whose columns are the sample points. The (p, q)-group norm of a matrix M is defined as the q norm of the p norm of the columns of M, that is $\|M\|_{p,q} = \left\|\left(\|M_1\|_p, \ldots, \|M_N\|_p\right)\right\|_q$, where M_is are the columns of M. We denote by $\{\sigma_i\}_{i=1}^m$ the Rademacher variables, that is independent uniform random variables taking values in $\{-1, +1\}$.

(a) Show that the empirical Rademacher complexity of $\mathcal{H}$ admits the following upper bound:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda}{m}\sqrt{2\log(2N)}\left\|X^\top\right\|_{2,\infty}.$$

(*Hint*: use Massart's lemma.)

(b) Show that for any $0 < p < \infty$, there exists a positive constant C_p such that the following inequality holds for all $m \geq 1$ and real numbers $a_1, \ldots, a_m$.

$$\mathop{\mathbb{E}}_{\sigma}\left[\left|\sum_{i=1}^m \sigma_i a_i\right|^p\right] \leq C_p\left(\sum_{i=1}^m a_i^2\right)^{\frac{p}{2}}$$

(*Hint*: For $p \leq 2$, you can use Jensen's inequality. For $p > 2$, w.l.o.g., rescale such that $\sum_{i=1}^m a_i^2 = 1$, use the identity $\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}[X > t]dt$ for $X \geq 0$.)

(c) Show that for any $0 < p < \infty$, there exists a positive constant c_p such that the following inequality holds for all $m \geq 1$ and real numbers $a_1, \ldots, a_m$.

$$c_p\left(\sum_{i=1}^m a_i^2\right)^{\frac{p}{2}} \leq \mathop{\mathbb{E}}_{\sigma}\left[\left|\sum_{i=1}^m \sigma_i a_i\right|^p\right]$$

(*Hint*: For $p \geq 2$, you can use Jensen's inequality. For $p < 2$, use Hölder's inequality and part (b).)

(d) Use the inequality shown in part (c), show that the empirical Rademacher complexity of $\mathcal{H}$ admits the following lower bound:

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \geq c_1\frac{\Lambda}{m}\left\|X^\top\right\|_{2,\infty},$$

where c_1 is some positive constant in part (c) for $p = 1$.

(e) By providing an example, show that the dimension dependence of $\sqrt{\log N}$ in the upper bound in part (a) is tight (*Hint*: consider a data set with $N = 2^m$).