Mehryar Mohri
Foundations of Machine Learning 2024
Courant Institute of Mathematical Sciences
Homework assignment 1
October 08, 2024
Due: October 22, 2024

A VC dimension

1. VC-dimension of axis-aligned hyper-rectangles. What is the VC-dimension of axis-aligned hyperrectangles in $\mathbb{R}^3$?
2. VC-dimension of union of two intervals. What is the VC-dimension of subsets of the real line formed by the union of two intervals?
3. VC-dimension of spheres. What is the VC-dimension of the set of all spheres centered at zero in $\mathbb{R}^n$?

B Generalization bound based on covering numbers

1. Let $\mathcal{H}$ be a family of functions mapping $\mathcal{X}$ to a subset of real numbers $\mathcal{Y} \subseteq \mathbb{R}$. For any $\epsilon > 0$, the *covering number* $\mathcal{N}(\mathcal{H}, \epsilon)$ of $\mathcal{H}$ for the L_∞ norm is the minimal $k \in \mathbb{N}$ such that $\mathcal{H}$ can be covered with k open balls of radius ϵ, that is, there exists $\{h_1, \ldots, h_k\} \subseteq \mathcal{H}$ such that, for all $h \in \mathcal{H}$, there exists $i \leq k$ with $\|h - h_i\|_\infty = \max_{x \in \mathcal{X}} |h(x) - h_i(x)| \leq \epsilon$. In particular, when $\mathcal{H}$ is a compact set, a finite covering can be extracted from a covering of $\mathcal{H}$ with balls of radius ϵ and thus $\mathcal{N}(\mathcal{H}, \epsilon)$ is finite.

   Covering numbers provide a measure of the complexity of a class of functions: the larger the covering number, the richer is the family of functions. The objective of this problem is to illustrate this by proving a learning bound in the case of the squared loss. Let $\mathcal{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$ according to which labeled examples are drawn. Then, the generalization error of $h \in \mathcal{H}$ for the squared loss is defined by $R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(h(x) - y)^2]$ and its empirical error for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ by $\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m (h(x_i) - y_i)^2$. We will assume that $\mathcal{H}$ is bounded, that is there exists $M > 0$ such that $|h(x) - y| \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The following is the generalization bound proven in this problem:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{H}} |R(h) - \widehat{R}_S(h)| \geq \epsilon\Big] \leq \mathcal{N}\Big(\mathcal{H}, \frac{\epsilon}{8M}\Big) 2\exp\Big(\frac{-m\epsilon^2}{2M^4}\Big). \tag{1}$$

   The proof is based on the following steps.

   (a) Let $L_S(h) = R(h) - \widehat{R}_S(h)$, then show that for all $h_1, h_2 \in \mathcal{H}$ and any labeled sample S, the following inequality holds:

$$|L_S(h_1) - L_S(h_2)| \leq 4M\|h_1 - h_2\|_\infty .$$

   (b) Assume that $\mathcal{H}$ can be covered by k subsets $\mathcal{B}_1, \ldots, \mathcal{B}_k$, that is $\mathcal{H} = \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_k$. Then, show that, for any $\epsilon > 0$, the following upper bound holds:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{H}} |L_S(h)| \geq \epsilon\Big] \leq \sum_{i=1}^{k} \mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_i} |L_S(h)| \geq \epsilon\Big].$$

   (c) Finally, let $k = \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M})$ and let $\mathcal{B}_1, \ldots, \mathcal{B}_k$ be balls of radius $\epsilon/(8M)$ centered at $h_1, \ldots, h_k$ covering $\mathcal{H}$. Use part (a) to show that for all $i \in [k]$,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\Big[\sup_{h\in\mathcal{B}_i} |L_S(h)| \geq \epsilon\Big] \leq \mathbb{P}_{S\sim\mathcal{D}^m}\Big[|L_S(h_i)| \geq \frac{\epsilon}{2}\Big],$$

   and apply Hoeffding's inequality (theorem D.2) to prove (1).

C Generalization bound for Lipschitz functions

1. Let (E, ρ) be a metric space and consider a set $F \subset E$. From the previous problem, we know that for any $\epsilon > 0$, the *covering number* $\mathcal{N}(F, \epsilon)$ of F is the minimal $k \in \mathbb{N}$ such that F can be covered with k open balls of radius ϵ. We now define the *packing number* $\mathcal{M}(F, \epsilon)$ of F as the maximal $k \in \mathbb{N}$ for which there exists a finite set $P \subset F$ of size k that is ϵ-separated, that is, for all distinct $p, p' \in P$, we have $\rho(p, p') > \epsilon$.

   (a) Show that for any metric, the ϵ-packing number is lower bounded by the ϵ-covering number:

   $$\mathcal{M}(F, \epsilon) \geq \mathcal{N}(F, \epsilon).$$

   (b) Show that for any metric, a d-dimensional ball B of radius r can be covered by $(3r/\epsilon)^d$ balls of radius $\epsilon < r$.

   (c) Let $G = \{g_\theta : \mathcal{Z} \to [0,1] \mid \theta \in \mathbb{R}^d, \|\theta\| \leq 1\}$ be a class of functions parameterized by $\theta \in \mathbb{R}^d$, $\|\theta\| \leq 1$, that is μ-Lipschitz in the following sense: $\|g_\theta - g_{\theta'}\|_\infty \leq \mu\|\theta - \theta'\|$, where the norm $\|\cdot\|$ on $\mathbb{R}^d$ is arbitrary. Show that $\mathcal{N}_\infty(G, \epsilon) \leq (3\mu/\epsilon)^d$. *Hint: show that* $\frac{\epsilon}{\mu}$ *cover of the unit ball in* $\mathbb{R}^d$ *for norm* $\|\cdot\|$ *provides an* ϵ *cover of* G*:* $\mathcal{N}_\infty(G, \epsilon) \leq \mathcal{N}_{\|\cdot\|}(B(0,1), \frac{\epsilon}{\mu})$.

   (d) Using the generalization bound based on covering numbers from the problem above, derive a generalization bound specifically for Lipschitz functions.